

Peptid-domén kölsönhatásokat stabilizáló tényezők feltérképezése

XXXII. OTDK Biológia Szekció

Gémes Dániel

Pázmány Péter Katolikus Egyetem
Információs Technológiai és Bionikai Kar

Molekuláris bionika mérnöki Bsc

3. évfolyam

Budapest, 2013. november 29.



Témavezető: Dr. Dosztányi Zsuzsanna

MTA TTK Enzimológiai Intézet

Konzulens: Dr. Gáspári Zoltán

Tartalomjegyzék

1. Bevezetés	3
2. Irodalmi áttekintés	4
2.1. A globuláris fehérjék szerkezeti szintjei	4
2.2. A fehérje folding	4
2.3. A fehérjék szerkezeti felderítésében használt kísérleti módszerek ...	5
2.4. A szerkezet-funkció paradigma	6
2.5. A rendezetlen fehérjék	6
2.6. Fehérje-fehérje kapcsolatok	8
2.7. Domén-domén kölcsönhatások	8
2.8. Peptid-domén kölcsönhatások	9
2.9. Lineáris motívumok	11
2.10. Peptid-domén kölcsönhatások predikciója	11
2.11. Statisztikus potenciálok	13
3. Célkitűzés	16
4. Adatbázisok és módszerek	17
4.1. Adatbázisok	17
4.2. Módszerek	18
5. Eredmények	21
5.1. Globuláris fehérjék adatbázisa	22
5.2. Nagyobb peptid adatbázis	24
5.3. Kisebb peptid adatbázis	26
5.4. Véletlenszerűen generált alternatív szekvenciák	27
5.5. Kiindulási mátrix hatása az eredményekre	27
5.6. Az eredmények értelmezése	27
6. Összefoglalás	29
7. Köszönetnyilvánítás	30
8. irodalomjegyzék	31

1. Bevezetés

A fehérjék az élőlények szervezetének legalapvetőbb építőkövei, részt vesznek a sejtek és szövetek szerkezetének felépítésében, működésében és anyagcseréjében. Biológiai jelentőségüket az mutatja, hogy szerepük van az összes sejtekben lejátszódó folyamatban. Többek közt olyan komplex biológiai folyamatokban vesznek részt, mint a sejtciklus szabályozás, a jelátvitel, a transzkripció vagy az enzimatikus reakciók. Ahhoz, hogy az élőlények molekuláris működését megértsük, ezáltal képesek legyünk a működésben beállt zavarokat, a különböző betegségeket megfelelően kezelni, szükségünk van a fehérjék funkcionalitásához szükséges tényezők megismerésére. A fehérjék funkcionalitásának alapja, hogy egymással kapcsolatokat alakítanak ki. Ezek a kapcsolatok a kölcsönhatás időtartalma, a kötődés erőssége és a kapcsolatot kialakító felszínek típusai alapján nagyon változatosak lehetnek.

Manapság egyre nagyobb figyelem irányul a peptid-domén kölcsönhatások felé. Ilyen típusú kölcsönhatásokat nagy számban találtak az élőlények különböző szabályozási folyamataiban. Jelentőségük ellenére ma még csak kevés tudással rendelkezünk ezekről a kapcsolatokról. A szabályozási folyamatokban résztvevő fehérjék sokszor bonyolult hálózatokat alakítanak ki. Az ilyen hálózatokban a kapcsolatok nagy száma illetve a peptid-domén kapcsolatok felderítésére használt kísérleti eljárások nehézségei szükségessé tették a bioinformatikai módszerek alkalmazását.

Az eddigi kísérletesen meghatározott adatokat felhasználva lehetőség nyílik a peptid-domén kapcsolatokat leíró modellek megalkotására. Ezek a modellek lehetnek részletes, a molekuláris kölcsönhatásokat is figyelembevevő modellek, vagy lehetnek kevésbé részletes, statisztikai adatokon alapuló modellek. Ezekből a modellekből aztán különböző szerkezet becslésre alkalmas algoritmusokat lehet alkotni, amelyek lehetővé tehetik kölcsönható partnerek keresését az adatbázisokban, illetve segíthetnek jobban megérteni a kapcsolatok kialakulását és stabilitását alakító tényezőket.

2. Irodalmi áttekintés

2.1. A globuláris fehérjék szerkezeti szintjei

A fehérjéknek négy különböző szerkezeti szintjét különböztethetjük meg. Az elsődleges szerkezeten a polipeptid láncot alkotó, kovalens kötésekkel kapcsolódó aminosavak kapcsolódási sorrendjét értjük. Az aminosavak sorrendje a génekben lévő nukleotidokban van kódolva. A másodlagos szerkezeten a polipeptid lánc lokális rendezettségét értjük. Az α -hélix és a β -redő gyakran előforduló, jól strukturált másodlagos szerkezeti elemek. Ezt a lokális rendezettséget másodrendű kémiai kötések teremtik meg ilyenek a hidrogén hidak, diszulfid hidak vagy van der Waals erők és a hidrofób kölcsönhatások. A harmadlagos szerkezet a polipeptid lánc teljes térbeli konformációját adja meg. Ezt a szerkezeti szintet is másodlagos kötési erők stabilizálják, de legfőbb képen a hidrofób kölcsönhatások. Ha egy fehérje több alegységből áll, akkor az alegységek elrendeződését a negyedleges szerkezettel írhatjuk le.

A domén a fehérje harmadlagos szerkezeti egysége, amely főként α -hélixbe vagy β -redőbe rendeződött másodlagos szerkezeti elemekből áll. A doménekre többféle képen tekinthetünk, attól függően, hogy milyen szempontból közelítjük meg szerepüket. Tekinthejük rájuk úgy, mint önálló feltekeredési (folding) egység, önálló funkcióval bíró egység vagy fehérje evolúciós egység. Ezek a meghatározások természetesen nem zárják ki egymást, sőt egy domén általában rendelkezik az első hárommal. Egy fehérjében több fajta domén is lehet, illetve egy adott típusú domén megjelenhet különböző fehérjékben is, ilyen például az ATP-kötő domén [1].

2.2. Fehérje folding

A folding az a folyamat, amely során a fehérjék felgomolyodnak, azaz a 3D-s szerkezetük kialakul. Az élőlényekben a riboszómákon a fehérjék az mRNS transzlációja után egy lineáris láncként születnek meg. Az aminosav sorrend által meghatározott elsődleges szerkezetből ezt követően a folding során a fehérje felveszi a rá jellemző térszerkezetet. A folyamat nem véletlenszerű, különböző útvonalakon és köztes szerkezeteken át a fehérje konformációja a natív szerkezet felé van irányítva. A natív szerkezetnek azt az állapotot hívjuk, amiben a fehérje funkcióképes. A natív szerkezet kialakításában különböző stabilizáló tényezők vesznek részt, többek között másodrendű kémiai kötések pl. hidrogén hidak vagy van der Waals erők, elektrosztatikus erők, de nagyon fontos a hidrofób effektus is. A fehérjék

foldingja során bekövetkezett hibák súlyos következményekkel járhatnak, többek között ennek tulajdonítják néhány neurodegeneratív betegség kialakulását [2].

2.3. A fehérjék szerkezeti felderítésében használt kísérleti módszerek

Ahhoz, hogy a fehérjék gazdag világát megértsük, szükségünk van a térszerkezetük meghatározására. A PDB fehérje adatbázisa szerint ma már több mint 90.000 fehérje pontos térszerkezetét ismerjük, amikhez bárki hozzáférhet a www.pdb.org webszerveren található adatbázisban. Ebbe az adatbázisba egy egységes, úgy nevezett pdb formátumban kerülnek be az adatok a fehérjékről, amelyek szerkezetét meghatározták [3]. A fehérjék szerkezetének meghatározására leggyakrabban használt és legelterjedtebb módszerek a röntgen diffrakció, a mágneses rezonancia (NMR) spektroszkópia és az elektronmikroszkópia.

A röntgen diffrakció az anyagtudomány egyik legalapvetőbb vizsgálati módszere. A 19. század elején Max von Laue volt az, aki először azt feltételezte, hogy a kristályos anyagok rácsszerkezettel rendelkeznek. Friedlich és Knipping (1912) bebizonyították a röntgen sugárzás hullámtermészetét és a kristályos anyagok rácsszerkezetét. A módszer alkalmazása során a fehérjéket kikristályosítják, majd röntgensugárzást bocsájtanak rá. A röntgensugarak elhajlásával (diffrakciójával) kapott intenzitásmaximumok térbeli helyzetéből és relatív erősségéből meghatározható a vizsgált anyag 3D-s szerkezete [4]. Az eljárás hátránya, hogy alkalmazása közben azzal a feltételezéssel élnek, hogy a fehérjék szerkezete kikristályosított állapotban ugyan olyan, mint amilyen fiziológias környezetben, oldott állapotban. Ez a feltételezés azonban nem mindig állja meg a helyét.

Az NMR spektroszkópia, vagy más néven mágneses magrezonancia spektroszkópia napjaink talán legelterjedtebb és leginkább széles körben alkalmazott kémiai szerkezetkezelési vizsgálati módszere. Az eljárást a fehérjék térszerkezetének meghatározása mellett gyógyszeripari és anyagtudományi kutatásokban is alkalmazzák. Az NMR spektroszkópia használata során nincs szükség a minta kikristályosítására, a módszer egyaránt képes molekuláris szintű szerkezeti információt nyújtani szilárd és oldatfázisú mintákról. Ez lehetőséget ad arra, hogy a fehérjéket roncsolás nélkül, eredeti közegükben lehessen vizsgálni. A módszer a vizsgálandó molekulát alkotó atommagok kvantummechanikai mágneses tulajdonságának mérésén alapszik. Erős mágneses térben az atommagok, amelyekben a protonok vagy neutronok száma páratlan, vizsgálhatóak és felvilágosítást tudnak nyújtani az őket körülvevő elektronok

sűrűségéről, ami viszont függ a szomszédos atomokhoz fűződő viszonytól, a kémiai kötéstől. Az atommagok ezen tulajdonságait kihasználva magspektroszkópiai módszerrel kapott spektrumból lehet következtetni az atomok elhelyezkedésére, aminek segítségével pedig fel lehet térképezni a molekula teljes szerkezetét. Az NMR spektroszkópia hátránya, hogy csak viszonylag kisméretű, nagyjából 200 aminosav rövidebb fehérjéket lehet vele vizsgálni.

2.4. Szerkezet-funkció paradigma

A klasszikus szerkezet-funkció paradigma szerint a fehérjék funkciójához jól definiált térszerkezetre van szükség. Bár tagadhatatlan, hogy a fehérjék által ellátott feladatok egy jó része szorosan kapcsolódik a térszerkezethez, sőt bizonyos funkciók valóban csak így láthatóak el, mint pl. az enzimátikus folyamatok, azonban az elmúlt egy-két évtizedben egyre több kísérleti bizonyíték mutatott arra, hogy ez az elképzelés túlságosan leegyszerűsített. Egyrészt már régóta ismert tény, hogy bizonyos fehérjeszakaszok, illetve kisebb hormon peptidek nem képesek önmagukban jól strukturált, rendezett szerkezetet felvenni, hanem térszerkezetük dinamikusan változik. Ezek a peptidek röntgen diffrakciós mérés során nem adnak koherens szórási képet. Másrészt az irodalomban sorra bukkantak fel azok az esetek, ahol eredménytelenül próbálkoztak egy adott fehérje kristályosításával, illetve szerkezetének meghatározásával. Azonban általános nézet volt, hogy ez csupán a kísérleti körülmények következménye. Egészen 1999-ig váratott magára az a felismerés, hogy a szerkezet-funkció összefüggést újra kell fogalmazni, mivel fiziológias körülmények között is léteznek olyan másod és harmadlagos szerkezet nélküli fehérjék, amelyek képesek egy adott funkciót ellátni [5]. Ezek a fehérjék többek közt olyan kritikus folyamatokban is részt vehetnek, mint a sejtciklus-szabályzás, transzkripció és programozott sejthalál.

2.5. Rendezetlen fehérjék

Rendezetlennek azokat a fehérjéket vagy fehérje szakaszokat nevezzük, amelyek natív, funkcionális állapotukban sem rendelkeznek jól definiált 3D-s szerkezettel tehát nem képesek stabil másodlagos és harmadlagos szerkezetet felvenni. Térszerkezetük dinamikusan változik, leginkább a globuláris fehérjék denaturált állapotára emlékeztetnek. Mivel ez a rendezetlenség nem külső hatásokra jön létre, hanem a fehérje saját tulajdonságaiból adódik ezért „*intrinsically unstructured protein*” IUP-nak vagy „*intrinsically disordered protein*” IDP-nek nevezzük őket. Funkciójuk ellátása közben kapcsolatot alakíthatnak ki más fehérjékkel, ekkor

a szerkezetük rendeződik és olyan strukturált másodlagos szerkezetet is felvehet, mint az α -hélix [6].

A természetüknél fogva rendezetlen fehérjék jelentőségét jól mutatja, hogy becslések szerint az eukarióta genom 30-60%-a rendezetlen régióval rendelkező, míg 5-15%-a teljes hosszúságukban rendezetlen fehérjéket kódol [7]. Evolúciós vizsgálatok bizonyítják, hogy egy élőlény komplexitásának növekedésével nő a genetikai állományában kódolt rendezetlen fehérjék aránya is. Természetüknél fogva rendezetlen fehérjékkel sokszor sejtciklus szabályozó, jelátviteli és transzkripciós folyamatokban találkozhatunk, ezért mutációjuk gyakran okozhat zavart a sejtosztódás szabályozásában, ami rákos megbetegedésekhez vezethet. Számos esszenciális fehérje, mint a p53, BRCA1, CREB, RNS polimeráz II és a prion fehérje esetében írtak le szerkezeti rendezetlenséget.

A természetüknél fogva rendezetlen fehérjék felfedezése sokáig váratott magára, hiszen a sokáig leginkább alkalmazott szerkezet vizsgálati módszer, a röntgen diffrakció alkalmatlan volt vizsgálatukra. Ahhoz, hogy ezzel az eljárással használható eredményt kapjunk, azt kell feltételeznünk, hogy a fehérje 3D-s szerkezete kikristályosított állapotban megegyezik az oldott állapotban felvett [4]. Ez a feltétel a rendezetlen fehérjék esetén nem teljesül, ezért a módszer alkalmatlan rendezetlen fehérjék, illetve domének valamint flexibilis hurkok és régiók vizsgálatára. A rendezetlen fehérjék vizsgálatára használt kísérleti módszerek általában más jellegű információt szolgáltatnak, amelyek értelmezése nem annyira egyértelmű, mint a röntgen diffrakciós kép. A ma leggyakrabban használt módszerek az NMR spektroszkópia, a távoli és közeli UV CD spektroszkópia, a fluoreszcens spektroszkópia és a Raman spektroszkópia.

A rendezetlenség a speciális, a fehérje-feltekeredéssel össze nem egyeztethető aminosav összetétel következménye, ezekben a fehérjékben igen gyakoriak a rendezetlenséget elősegítő aminosavak, mint például az Ala, Arg, Gly, Gln, Ser, Pro, Glu és Lys, míg csaknem teljesen hiányoznak belőlük a rendezettséget elősegítő aminosavak: a Trp, Tyr, Cys, Phe, Ile, Val, Leu és Asn [7]. Az aminosav szekvenciájuk alacsony komplexitású és gyakran tartalmaz ismétlődéseket. A rendezett szerkezet hiányának legfőbb fizikai okai a magas nettó töltés és az alacsony hidrofóbicitás. Az ellentétesen töltött aminosavak taszítják egymást, a hidrofób aminosavak alacsony aránya pedig megakadályozza a hidrofób mag kialakulását [8]. Ezek alapján a rendezetlen fehérjék a globuláris fehérjékhez képest eltérő szekvenciális

tulajdonságokkal rendelkeznek, amely lehetővé teszi az aminosav sorrendből történő predikciójukat. Jelenleg több mint ötven módszer létezik, ezek közül az egyik az Enzimológiai Intézetben, a csoportban kifejlesztett IUPred módszer [9].

2.6. Fehérje-fehérje kapcsolatok

A fehérjék az élőlények komplex biológiai szabályozási folyamataiban vesznek részt. Ahhoz hogy ezek a szabályozási folyamatok egymással párhuzamosan, nagy hatékonysággal tudjanak működni összetett hálózatok alakultak ki [6]. Az információ közvetítéséhez a fehérjék kapcsolatokat alakítanak ki egymás között. Ezek a kapcsolatok tulajdonságaikban nagyon eltérőek lehetnek egymástól, bizonyos esetben rövid idegi fennmaradó, kis stabilitású, úgy nevezett átmeneti komplex kialakítása, máskor hosszú idegi, akár napokig, évekig megmaradó, stabil, úgy nevezett állandósult komplex kialakítása a cél. Bár a kapcsolatok kialakulását több dolog is befolyásolhatja, mint a kapcsolódási partnerek koncentrációja, a kémiai környezet, bizonyos molekulák jelenléte, a legfontosabb tényező mégis a kapcsolódási felület, más néven interfész. Alapvetően az interfész mérete, alakja és összetétele határozza meg a kapcsolatok időtartalmát, erősségét és specificitását [10]. A kölcsönható felületeket szempontjából két nagyobb csoportot tudunk elkülöníteni, a domén-domén kölcsönhatásokat és a peptid-domén kölcsönhatásokat.

2.6.1. Domén-domén kölcsönhatások

Domén- domén kölcsönhatásokban két fehérje funkcionális egysége, doménje között alakul ki kapcsolat. Ezt a kölcsönhatás típust már régóta tanulmányozzák, számos komplex röntgen diffrakcióval meghatározott térszerkezetét meg lehet találni a PDB adatbázisban. A kapcsolódási felületet, más néven interfészt, úgy definiálhatjuk, mint azok a felületen található hozzáférhető aminosavakat, amik a komplex kialakulása után fedésbe kerülnek. A kötő felület a fehérjék több mint 75%-ban 500-1500 Å² közé esik, általában nagyobb interfész erősebb kapcsolatok kialakulását teszi lehetővé, tehát az interfész mérete összefüggésben van a kapcsolat erősségével. Nagyobb interfésszel rendelkeznek például az állandósult komplexek kialakításában résztvevő fehérjék. Érdekes módon a fehérje hossza és a kapcsolódási felület mérete között nincs lineáris összefüggés.

A fehérjéknek a transzkripció után a folding során alakul ki a térszerkezetük. Ennek következtében a 3D-s térben a polipeptid láncnak a szekvenciában egymástól távoli részei

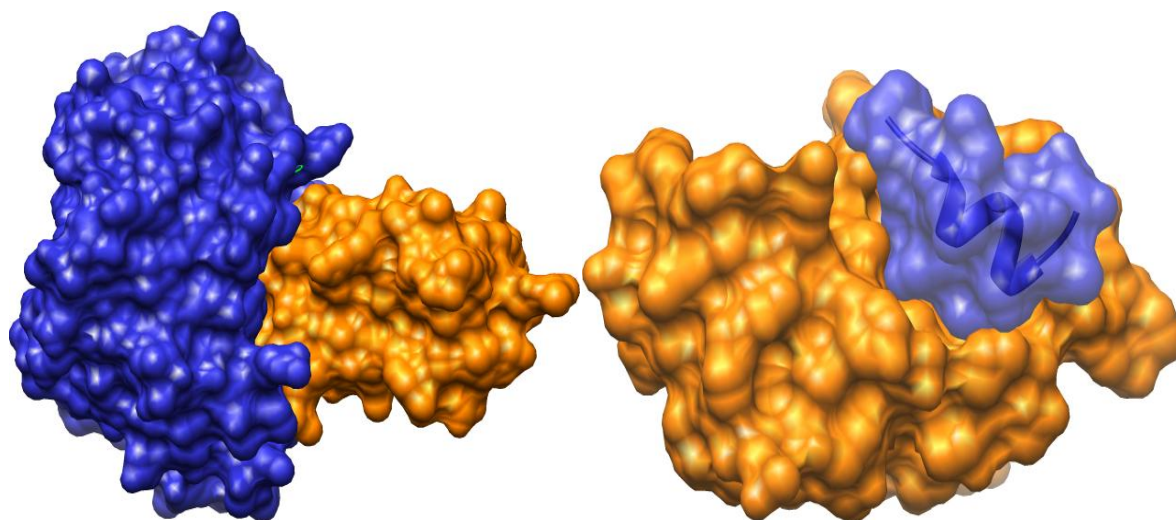
kerülhetnek egymás közvetlen közelébe. Ez az oka annak, hogy a domén-domén kölcsönhatásoknál a kapcsolódási felület kialakításában résztvevő aminosavak a szekvenciában nem folytonosan, hanem általában 5-7 nem folytonos szegmensben találhatóak meg. Ezek az aminosavak hidrogénkötéseket, hidrofób kölcsönhatásokat, van der Waals kölcsönhatásokat és só hidakat képeznek a partner molekula interfészén lévő aminosavakkal, ezzel összekapcsolva a két fehérjét. A fehérjéken egymással kölcsönható aminosavak között általában 4-8 atomi kontaktus alakul ki. A kapcsolatok specifikitásában közrejátszhat az interfészek geometriai formája is. Az interfészek funkcionális fontosságát legjobban az evolúciós konzerváltságuk mutatja. Az evolúciós vizsgálatok azt mutatták, hogy a kötőfelületeknek az evolúciós konzerváltsága jóval nagyobb, mint a fehérje többi részének [5].

2.6.2. Peptid-domén kölcsönhatások

Habár a ma ismert fehérje-fehérje kölcsönhatások közül a legtöbb klasszikus globuláris domén-domén kölcsönhatás, egyre nagyobb számban találkozunk olyan esetekkel, ahol egy fehérje rövid lineáris része, ez az úgy nevezett peptid rész, köt egy globuláris fehérje doménjához. Peptid-domén kölcsönhatásokkal olyan funkcionálisan jelentős biológiai folyamatokban találkozhatunk, mint a jelátvitel, sejtosztódás és egyéb szabályozási folyamatok. A domén-domén kölcsönhatások és a peptid-domén kölcsönhatások közt folytonos átmenet van, nehéz meghatározni a pontos határokat. Azonban a két kapcsolati típus tulajdonságai közt olyan jelentős különbségek vannak, amik indokoltá teszik, hogy külön vizsgálják őket.

Definíciótól függően, a 15-30 aminosavnál rövidebb kötőrégiókat nevezzük peptidnek. Ezek a kötőrégiók nagyrészt rendezetlen régiókba esnek, bár előfordulhatnak rendezett részeken is [11]. A peptidek fontos tulajdonsága, hogy az őket alkotó aminosavak folytonosan, vagy maximum egy-két szegmensben helyezkednek el a szekvenciában, ellentétben a globuláris domén kötőhelyekkel, ahol a kötést kialakító aminosavak egymástól távol, 5-7 szegmensben találhatóak a szekvenciában és csak a 3D-s szerkezetben kerülnek egymás közelébe. A kötést kialakító aminosavak nem egyenlő arányban járulnak hozzá a kapcsolódáshoz szükséges szabad energiához. Csupán néhány kiemelt jelentőségű aminosav adja a kapcsolódáshoz szükséges energia közel 70%-át [12]. A peptidek poláros, nem poláros aminosav összetétele nem különbözik jelentősen a domén-domén kölcsönhatásokban közrejátszó aminosavak összetételétől.

A legtöbbször rendezetlen szerkezetű peptidek, a kötődés során hosszan elnyújtott, rendezett szerkezetet képesek felvenni, néhány esetben akár olyan komplex másodlagos szerkezetet is, mint az α -hélix vagy a β -redő. Ha a kötést kialakító peptid egy nagyobb rendezetlen régió része, akkor a kapcsolat kialakulása során csak a peptid régiójában található, kapcsolatban közvetlenül résztvevő szakasz rendeződik. Sok esetben a peptidek könnyen hozzáférhetőek a partner molekula számára és a kötődés során képesek mélyen eltemetődni a doménon található árkokba, mélyedésekbe. Az interfész, a felület, amely a kapcsolódás során eltemetődik, körül belül $500 \text{ \AA}^2$ nagyságú [12]. Ez nagyjából a fele a domén-domén kölcsönhatások kapcsolódási felületének. A peptidek evolúciós konzerváltságának vizsgálata alapján a kötőfelület jóval erősebben konzervált, mint a fehérje többi része, azonban ez a konzerváltság alacsonyabb, mint amit a domén-domén kölcsönhatásoknál tapasztaltak.



1. ábra. A bal oldali ábrán egy domén-domén kölcsönhatás jellemző példáját láthatjuk, a tripszin-BPTI komplexet, míg a jobb oldali ábra egy peptid-domén kölcsönhatás által kialakított MDM2-p53 komplexet mutat, ahol a rendezetlen kötő peptid rendeződik és α -hélix struktúrát vesz fel kötődés során.

2.7. Lineáris motívumok

A lineáris motívumok rövid, szekvenciájukban sokszor erősen konzervált peptid szegmenssek. Funkciójuk során azonos doménnel vagy domén családdal alakítanak ki kapcsolatot [11]. Lehetnek kötőhelyek, amik kis affinitású fehérje-fehérje kapcsolat kialakulását teszik lehetővé, foszforilációs helyek vagy hasítási helyek a proteázok számára. Általában 3-11 aminosavból állnak és lineárisan találhatóak meg a szekvenciában [12]. A lineáris motívumokat tartalmazó szekvenciákban bizonyos pozíciókba csak meghatározott aminosavak kerülhetnek, míg más pozíciókban nem találunk szabályszerűséget, ezekbe bármilyen aminosav kerülhet. A lineáris motívumokat legtöbbször reguláris kifejezésekkel szokták megadni. Az ELM (Eukaryotic Linear Motif) adatbázis a legnagyobb, eukarióták lineáris motívumait tartalmazó adatbázis, amely jelenleg körülbelül 200 kísérletesen igazolt motívumot tartalmaz [13, 14].

2.8. Peptid-domén kölcsönhatások predikciója

A peptid-domén kölcsönhatások jelentősége az utóbbi időben vált egyre nyilvánvalóbbá. Az ilyen típusú kölcsönhatások a jelátviteli és egyéb szabályozási funkciókat ellátó fehérjék dinamikus kölcsönhatásainak kialakításában játszanak fontos szerepet. Ugyanakkor a domén-domén kölcsönhatásokhoz képest, jóval kevesebb adattal és tudással rendelkezünk róluk. Ennek számos oka lehet, mint a rendezetlen kötőhelyek kísérletes vizsgálatának nehézségei vagy, hogy ezek a kapcsolatok sokszor csak tranziens, rövid életű vagy olyan speciális környezetet igénylő komplexekben jelennek meg, amelyek vizsgálata nehéz feladat és különleges technikákat igényel [5, 6].

Bioinformatikai módszereket felhasználva egy átfogóbb képet kaphatunk ezekről a kapcsolatokról, ami segíthet jobban megérteni a kialakulásukat befolyásoló tényezőket. Ezen felül, ismert adatokból kiindulva, olyan számítógépes algoritmusokat lehet készíteni, amelyek használhatóak például potenciális kötőpartnerek megtalálására. A peptid-domén kölcsönhatások speciális tulajdonságai miatt, az ilyen típusú kölcsönhatások jóslására egészen más megközelítéseket kell alkalmazni, mint a domén-domén kölcsönhatások predikciójánál. Egyszerű csoportosítással élve, a peptid-domén kapcsolatok predikciójára használt algoritmusok két főbb csoportját lehet elkülöníteni, a szekvencia alapú és a szerkezet alapú predikciós módszereket, attól függően, hogy a szekvencia vagy a szerkezeti adatok ismeretében képes a potenciális kapcsolatokat megjósolni.

A szekvencia alapú predikció egyik megközelítése, hogy az ismert lineáris motívumokat használja fel a kölcsönható partnerek azonosításához. Azonban ezek a rövid szekvencia elemek akár véletlenül is előfordulhatnak a szekvenciában és gyakran ezek a rossz találatok dominálnak [14]. A valódi találatok aránya jelentősen növelhető, ha egyéb információkat is figyelembe veszünk. A lineáris motívumok gyakran fednek át rendezetlen kötőhelyekkel [15]. Az ANCHOR elnevezésű predikciós eljárás ezeken a rendezetlen kötőhelyeken található lineáris motívumok felismerésére specializálódott [11]. Ez a módszer ezeknek a rendezetlen kötőrégióknak két alapvető fizikai tulajdonságát próbálja megragadni, azt, hogy ezek nem képesek szekvenciális környezetükkel kedvező kölcsönhatásokat kialakítani, másrészt hogy egy globuláris fehérjével kedvező kölcsönhatást tudnak képezni. Az eljárás egy egyszerűsített energiaszámoláson alapuló fizikai modellre épül. A megemlített példák mellett léteznek még egyéb, különböző machine-learning eljárásokon alapuló módszerek is.

A peptid-domén kölcsönhatások szerkezet alapú predikciója esetén a kiinduló pontot, a kölcsönhatás pontos atomi részleteit leíró szerkezetek adják. Szemben azonban a globuláris fehérjék ismert szerkezetű komplexeinek a számával, eddig jóval kevesebb peptid-domén komplex szerkezetet határoztak meg. Ezek száma a PDB térszerkezeti adatbázisban csupán néhány százra tehető. Az eddig vizsgálatok arra utaltak, hogy a kötő domén szerkezete nem változik jelentősen a kötődés során, illetve az egy adott doménhez kötő peptid konformációja sok esetben hasonló [16]. Ezért az ismert peptid-domén szerkezetek jó kiinduló pontjai lehetnek egy adott doménhez kötő peptid kölcsönhatásainak modellezésére. Több olyan eljárást is kifejlesztettek, amely a peptid-domén kölcsönhatások szerkezeti tulajdonságait veszik figyelembe [17,18]. Ezeket a komplexek térszerkezeti adataiból kiindulva, lényegében szerkezetfinomításon alapuló módszereket már sok esetben sikeresen használták többek között foszforiláció hatásának modellezésére, új kötőmotívumok azonosítására, kötési energia jóslására, illetve szintetikus gátló peptid kifejlesztésére. Azonban ezek a módszerek általában nagyon időigényesek, és gyakran csak bizonyos típusú peptid-domén kölcsönhatásokra működnek jól, ezért kevésbé alkalmasak nagyszálás vizsgálatokra.

A különböző térszerkezet jósló eljárások legfontosabb eleme az energiafüggvény, amely az adott modell jóságát jellemzi. Ezeknek az energiafüggvényeknek három nagy csoportja van: a molekula dinamikai számításoknál gyakran alkalmazott fizikai energiafüggvények, az empirikus energiafüggvények, melyek különböző, elsősorban mutációs kísérletes adatokon alapulnak, illetve a statisztikai potenciálok, melyek atom vagy aminosav kontaktusok

gyakoriságán alapulnak [18]. A tapasztalat szerint az empirikus vagy statisztikus potenciálokra alapuló energiafüggvények sokkal jobban működnek a térszerkezet becselő eljárásokban. Egyrészt, a fizikai potenciálokkal szemben gyorsan számolhatóak és kevésbé érzékenyek a kisebb változásokra a fehérje struktúrájában. Ennek az alapja egyfajta durva szemcsés leírás, ahol az egyes atomokat együttesen, egyetlen kölcsönhatási centrumként kezelnek. Másrészt, az ilyen energiafüggvények jól működnek olyan komplex effektusok esetén, mint például a hidrofóbicitás, ahol az egyes tényezők leírása önmagában nagyon nehéz. Azonban egy jól használható és megbízható energiafüggvény megalkotása nehéz feladat. Meg kell találni az egyensúlyt a különböző energia tagok között, amelyek az adott megközelítésben fontosak a stabilitás kialakításában. Általában ezek az energiafüggvények probléma-függők. Az eddigi fehérje-fehérje kapcsolatok szerkezeti predikciójánál használt energiafüggvények a globuláris fehérjékre lettek kifejlesztve. A peptid-domén kölcsönhatások eltérő jellege miatt szükség van az ilyen jellegű kölcsönhatásokat leíró energiafüggvények kifejlesztésére is.

2.9. Statisztikus potenciálok

A különböző térszerkezet becselő eljárásoknál gyakran alkalmazott energiafüggvények egyik fő fajtáját a statisztikus potenciálok alkotják. Általánosságban, a statisztikus potenciálok olyan energiaszerű mennyiségek, amelyek ismert térszerkezetű fehérjékből származtatott statisztikai adatokon alapulnak. A legegyszerűbb statisztikus potenciálok az egyes aminosavak közötti kontaktusokat veszik figyelembe. Ehhez a megfigyelt aminosav párkölcsönhatási mintázatokat és az ebből származó párkölcsönhatási gyakoriságokat használják fel. A megfigyelt párkölcsönhatási gyakoriságokat az aminosav kapcsolatokhoz tartozó párkölcsönhatási energiákká lehet alakítani a Boltzmann hipotézis alapján. Ezeket az energiákat egy 20x20-as szimmetrikus mátrixként lehet reprezentálni. A párkölcsönhatási energia meghatározásához először is ki kell számolni az i -dik aminosav j -dik aminosavval vett párkölcsönhatási gyakoriságát az adatbázisban és a referencia állapotban. A kapott értékek hányadosának a természetes alapú logaritmusát véve kapunk egy dimenzió nélküli energiamennyiséget.

$$e_{ij} = -\ln \left(\frac{p_{ij}}{p *_{ij}} \right)$$

Ahol a p_{ij} az i -dik aminosav j -dik aminosavval megfigyelt párkölcsönhatási gyakorisága és a p^*_{ij} az i -dik és a j -dik aminosav közötti párkölcsönhatási gyakoriság a referencia állapotban [19]. Nehézséget okozhat azonban a referencia állapot megfelelő leírása. Az általában alkalmazott modellben azt feltételezik, hogy az aminosavak egymástól függetlenül lépnek kapcsolatba. Ez a feltételezés azonban nem állja meg a helyét, a várható kapcsolatok száma függ a fehérje hosszától, aminosav összetételétől, illetve kölcsönható aminosavak elhelyezkedésétől is.

Ezeknek a problémáknak a kiküszöbölésére javasolt egy újfajta algoritmust Paul D. Thomas és Ken A. Dill [20]. A megközelítés legfontosabb eleme, hogy a referencia állapot a szekvencia által meghatározott teljes lánc, amely különböző konformációkat vesz fel. Így nem abból indul ki, hogy az aminosavak véletlenszerűen, egymástól függetlenül képesek kölcsönhatni. A Paul D. Thomas és Ken A. Dill által leírt ENERGI algoritmus azon alapszik, hogy a natív szerkezet energetikailag lényegesen kedvezőbb az alternatív szerkezetekhez képest. Ismert térszerkezetű fehérjék adatbázisait felhasználva, a megfigyelt párkölcsönhatási gyakoriságokból számolt energiákból készít egy energiafüggvényt, amit úgy optimalizál, hogy az energetikailag kedvező natív szerkezet és az energetikailag kedvezőtlen alternatív szerkezetek minél jobban elkülöníthetőek legyenek. Mindezt egy iteratív eljárás során valósítja meg, ahol minden iterációban úgy igazítja a kiindulási statisztikus potenciál mátrixában található energia értékeket, hogy a natív szerkezetet minél jobban el tudja különíteni az alternatívoktól az energiájuk alapján. Az ENERGI hatékonyan alkalmazható a natív szerkezet megtalálásában a globuláris fehérjék esetén. Érdekes módon azonban, az eljárást fel lehetett használni a rendezetlenség és a rendezetlen kötőhelyek jóslására is. A már említett ANCHOR és IUPred módszerek az ENERGI algoritmus által számolt energiaértékek szekvenciából becslésén alapul [9, 11].

A statisztikus potenciálokat sokkal általánosabb fehérje stabilitási problémákban is alkalmazzák, mint a fehérjemodellek minőségének ellenőrzése, ismeretlen struktúrájú fehérje lehetséges szerkezeteinek megtalálására az aminosav szekvencia ismeretében vagy ismert struktúrájú fehérjéhez tartozó aminosav szekvenciák megtalálására. Az előbbi threading problémának, míg az utóbbit inverz folding problémának vagy fehérje tervezésnek hívják.

A threading probléma során a fehérje szekvenciáját ismerjük és azt a szerkezetet akarjuk meghatározni, ami ehhez az ismert szekvenciához a legjobban illeszkedik. A szekvenciához legjobban illeszkedő szerkezetnek azt tekintjük, ahol a legmélyebb az energiája az összes

többi alternatívához képest. Az inverz folding probléma a threading problémának pont a megfordítottja. A fehérje térszerkezetét ismerjük és azt a szekvenciát akarjuk meghatározni, ami az adott térszerkezetre a legjobban illeszkedik. Az illeszkedés ebben az esetben is azt jelenti, hogy a natív szekvencia által meghatározott szerkezet energiája jóval kedvezőbb, mint az alternatív szekvenciák által meghatározott szerkezetek energiája.

3. Célkitűzés

Az utóbbi években egyre világosabbá vált azoknak a fehérje-fehérje kapcsolatoknak a szerepe, amelyekben az egyik partner rövid peptid része alakít ki kapcsolatot egy másik fehérje globuláris doménjával. Ezek az úgy nevezetett peptid-domén kölcsönhatások gyakran előfordulnak különböző szabályozási hálózatokban, ezért megértésük és felismerésük nélkülözhetetlen lehet egyes betegségek kezeléséhez. Jelentőségük ellenére ma még kevés adattal és tudással rendelkezünk róluk.

A munkám során egy olyan módszer kifejlesztését kíséreltem meg, aminek segítségével meg lehet majd jósolni, hogy az ismert térszerkezetű domének milyen peptidekkel képesek kölcsönhatást kialakítani. Az eredményeim ezen kívül segíthetnek majd abban, hogy jobban megérthessük a peptid-domén kölcsönhatásokat hajtó és stabilizáló erőket.

A peptid-domén kölcsönhatások vizsgálatához egy olyan programot készítettem, amely a különböző térszerkezet becslő eljárásoknál már jól bevált statisztikus potenciálok módszerén alapszik. A programomban a térszerkezeti adatokon megfigyelt párkölcsönhatási gyakoriságokat felhasználva konstruálok egy energiafüggvényt, a Paul D. Thomas és Ken A. Dill által javasolt algoritmus alapján.

A munkámhoz ismert, kísérleti úton meghatározott, peptid-domén kölcsönhatásokból kialakult komplexeket használok fel. Az eredeti algoritmus alternatív szerkezetek halmazára optimalizált, ez azonban nehezen alkalmazható komplexekre. Ezért az eredeti algoritmust adaptálom szekvencia optimalizációra, majd alkalmazom az általam összeállított peptid-domén adatbázisra.

4. Adatbázisok, módszerek

4.1. Adatbázisok

Munkám során három adatbázist használtam, ebből az egyik globuláris fehérjékből áll, míg a másik kettő peptid-domén kölcsönhatásokat tartalmazott. Mindhárom a PDB fehérje térszerkezet adatbázisból összeállított nem redundáns fehérje lista [3]. A globuláris fehérjék adatbázisához az IUPred megalkotása során használt fehérjéket használtam fel [9]. Ez egy olyan fehérje lista, amely nem tartalmaz több fehérje láncból álló szerkezeteket, illetve olyan szerkezeteket, amelyeket nukleinsavval, fémionokkal, vagy *hem*-mel való kölcsönhatások stabilizálnak. Ez az adatbázis 553 fehérjéből állt.

A két peptid-domén adatbázis közül az egyik specifikusan lineáris motívumon alapuló komplexet tartalmazott. Ez a csoport korábbi munkája során használt adatbázis [11] további szűrése révén jött léte. Ennek kiindulópontja az ELM adatbázis volt [13], amely jelenleg elérhető legteljesebb lineáris motívum adatbázis, amely tartalmaz információt az adott motívumhoz tartozó PDB szerkezetekről is. Míg a korábbi munkákban csak a peptidek szekvenciájára végeztek hasonlósági szűrést, most a domén partner szekvenciájára is szűrtünk a PISCES szerver segítségével [21]. Tovább csak azokat a komplexet tartottuk meg, ahol a lineáris motívumot tartalmazó lánc rövidebb volt 30 aminosavnál. Az így kapott adatbázis 84 elemet tartalmazott.

Egy általánosabb peptid-domén adatbázist is létrehoztunk. Ehhez a teljes PDB adatbázisból kiszűrtük azokat az eseteket, ahol a PDB szerkezetben pontosan kétfajta fehérjelánc szerepelt, és ebből az egyik rövidebb volt, mint 50 aminosav. Az így kapott komplexeket szűrtük a hosszabbik lánc aminosav szekvenciája alapján, csak egy reprezentánst megtartva 25% szekvencia hasonlóság alapján a PISCES szervert használva [21]. Az adatbázis 384 elemet tartalmazott. Az így létrejött adatbázis jóval több peptid-domén kölcsönhatást tartalmaz, mint amennyi jelenleg az ELM adatbázisban található.

4.2. Módszerek

A projektben szereplő algoritmus elkészítéséhez LINUX operációs rendszert és az Eclipse fejlesztői környezetét használtam. A számítási feladatok komplexitása miatt a C++ programozási nyelvre esett a választásom. A program nagy memóriaigénye miatt többször is újra kellett terveznem a programot, hogy optimálisabban számoljon. Végül a dinamikus memória használat mellett döntöttem. A program futtatásához igénybe vettem az MTA Enzimológiai Intézetén található szerver gépeket is.

A legelső feladatomból az volt, hogy az összeállított adatbázisban szereplő PDB formátumú fájlokból a számomra szükséges információkat kinyerjem. Ehhez készítettem egy PDB típusú fájlokat adatai beolvasni, és feldolgozni képes programot. Az általam összeállított adatbázisban szereplő fájlokból generáltam egy olyan kimenetet, ami csak a vizsgált peptid-domén kölcsönhatásban résztvevő peptid szekvenciáját, a peptiden belüli kapcsolatokat és azokat a peptid-domén közötti aminosav kölcsönhatásokat tartalmazza, amelyek részt vesznek a kötés kialakításában. Ezt a fajta kimenetet a globuláris fehérje adatbázisomra is legeneráltam. Itt a kimenet a fehérje szekvenciáját és a fehérjén belüli aminosav kapcsolatokat tartalmazta.

Ezt követően megkíséreltem egy olyan algoritmus elkészítését, amely képes egy adott konformáció esetén a natív szekvencia felismerésére. Ezt alkalmaztam globuláris fehérjékre és peptid-domén komplexek esetén a natív peptidek felismerésére is. Az algoritmus elkészítéséhez a Paul D. Thomas és Ken A. Dill által publikált ENERGI módszert vettem alapul [20]. Az ENERGI célja egy olyan eljárás kifejlesztése volt, ami a folding probléma megoldásával megtalálja a natív szerkezetet egy ismert aminosav szekvenciához. Ezt úgy valósította meg, hogy ismert térszerkezetű fehérjék adatbázisát felhasználva készített egy energiafüggvényt, ami speciálisan a natív szerkezet és az szerkezetek elkülönítésére volt optimalizálva. Ezzel szemben én, az ENERGI-hez hasonló iteratív módszert felhasználva, egy olyan eljárást próbáltam kifejleszteni ami, az inverz folding probléma megoldását keresi, azaz egy adott szerkezethez a natív szekvenciát. Az optimalizáció során a cél az volt, hogy a natív szekvencia energiája minél alacsonyabb legyen az alternatív szekvenciák energiájához képest.

Az alternatív szekvenciákat kétféle módon állítottam elő. Először az alternatív szekvenciába véletlenszerűen sorsoltam az aminosavakat. A második módszer során a natív szekvenciákat összefűztem egy hosszú szekvenciába majd ebből vágtam ki a megfelelő hosszúságú

szakaszokat. Ezzel a megoldással valóság hűbb alternatív szekvenciákat állíthattam elő, hiszen az aminosavak nem véletlenszerűen követik egymást a szekvenciában, bizonyos mintázatok sokkal gyakrabban fordulnak elő, míg mások pedig csak nagyon ritkán vagy soha. Mindkét módszernél ügyeltem arra, hogy az alternatív szekvenciák közé véletlenül ne kerüljön be a natív szekvencia. Fontos megjegyezni, hogy a generálásnál olyan szekvencia is bekerülhetett az alternatív szekvenciák közé, ami valójában kötni képes a peptidhez. Az alternatív szekvenciák előállításánál a kapcsolatok kialakításáért felelős aminosavak pozícióit a peptid szekvenciájában nem változtattam, csak az aminosavak típusait.

Az alternatív szekvenciák elkülönítését a natív szekvenciától az ENERGI által prezentált iteratív módszer szerint végeztem [20]. Az algoritmus a 0. iterációban egy kiindulási aminosav párkölcsönhatási energiákat tartalmazó statisztikus potenciál mátrixot használ fel. Az energiamátrix felhasználásával kiszámolom a natív és az általam generált alternatív szekvenciákhoz tartozó energiákat. Ezeket felhasználva a Boltzmann statisztika segítségével megbecsülöm az egész adatbázisra vonatkozó aminosav párkölcsönhatási gyakoriságokat. A 0. lépésben még az alternatív szekvenciák sokaságából becsült párkölcsönhatási gyakoriságok nem fognak megegyezni a natív szekvenciában megfigyelttel. A hiba becslésére az alábbi formula használható:

$$\Delta e_{ij}^{(1)} = -\ln\left(\frac{p_{ij}(\text{observed})}{p_{ij}^{(0)}(\text{predicted})}\right)$$

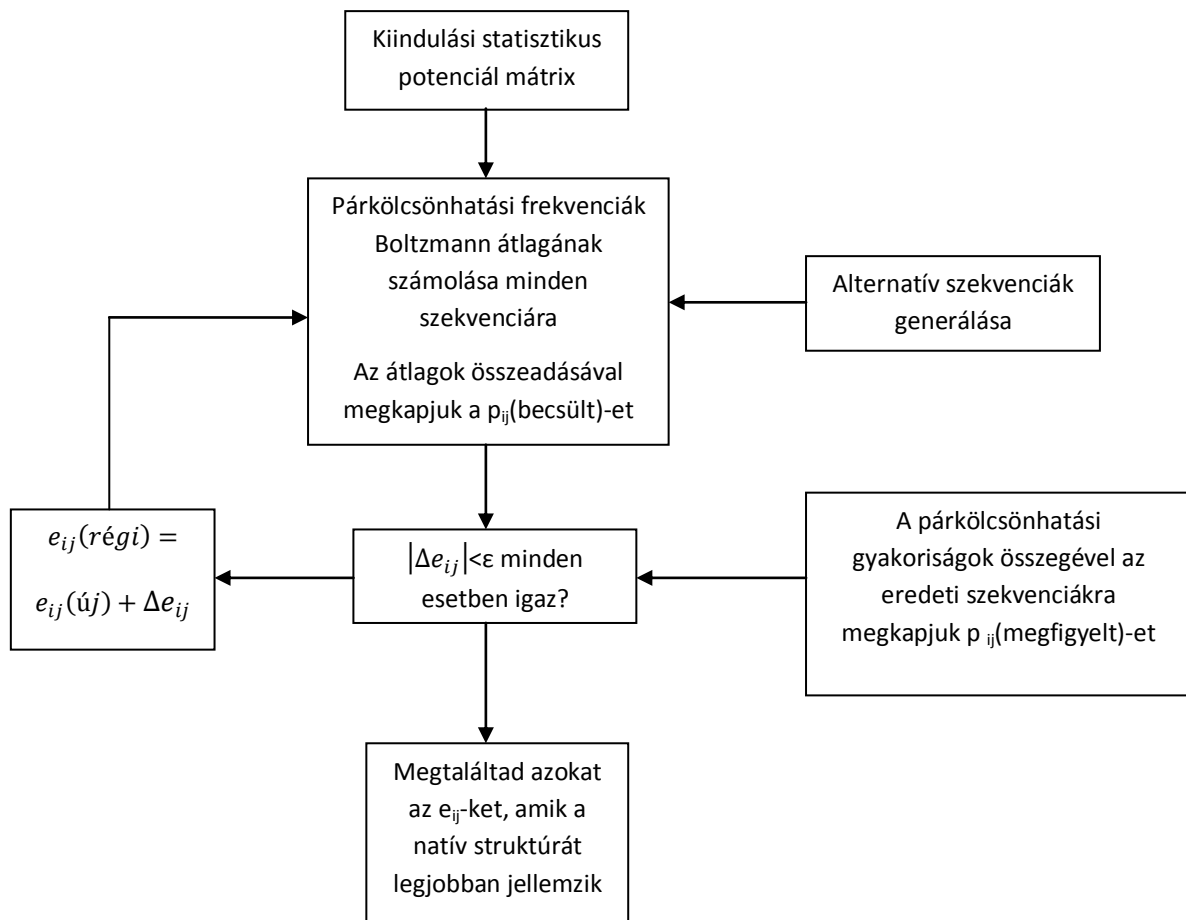
Ahol a p_{ij} az i-dik aminosav j-dik aminosavval megfigyelt párkölcsönhatási gyakorisága a natív szekvenciában, a $p_{ij}^{(0)}$ az i-dik aminosav és a j-dik aminosav között becsült párkölcsönhatási gyakoriság. A következő iterációban használandó párkölcsönhatási energiákat, azaz az új statisztikus potenciál mátrixunkat a most használt energiák hibával való korrigálásával kapjuk.

$$e_{ij}^{(1)} = e_{ij}^{(0)} + \Delta e_{ij}^{(0)}$$

Általánosan az n+1-dik iterációban használt párkölcsönhatási energiákat az n-dik lépésben használt párkölcsönhatási energiák n-dik lépésben számolt hibával való korrigálásával kapjuk [20]. A statisztikus potenciál mátrixunkat, tehát úgy korrigáljuk, hogy a belőle számolt energiákkal minél jobban képesek legyünk elkülöníteni a natív szekvenciákat az alternatívoktól. Azt, hogy egy szekvencia mennyire kedvező energetikailag, a Boltzmann eloszlás segítségével az adott szekvencia előfordulásának valószínűségére fordíthatjuk le. Ha

az energiafüggvény optimalizálása során a natív szekvencia előfordulásának az esélyét növeljük, az alternatívokkal szemben, akkor az energiamátrixunkban található értékek konvergálni fognak és becsült párkölcsönhatási gyakoriságok egyre közelebb kerülnek a megfigyelt párkölcsönhatási gyakoriságokhoz.

A kiindulási statisztikus potenciál mátrixom egy -1 és 0 közötti, véletlenszerűen generált értékeket tartalmazó 20x20-as szimmetrikus energiamátrix.



2. ábra. Az alábbi folyamat ábra bemutatja az algoritmusom lépéseit. A $|\Delta e_{ij}| < \epsilon$ hiba küszöb figyelése helyett én 100 iterációig futattam minden esetben a programomat.

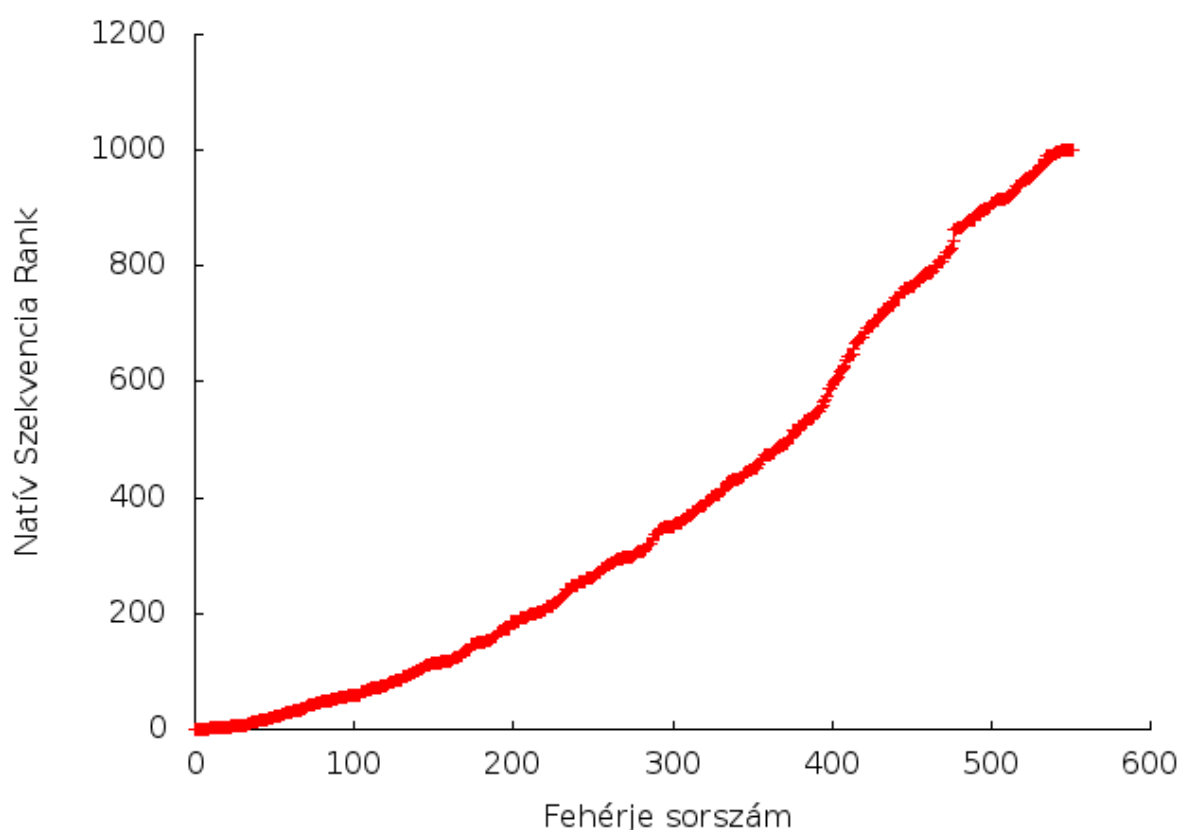
5. Eredmények és összefoglalás

Az algoritmusom által felhasznált statisztikus potenciál energia mátrixomban a párkölcsönhatási energiák konvergálnak. A becsült aminosav párkölcsönhatási energiák is konvergálnak a megfigyelt aminosav párkölcsönhatási energiákhoz, tehát az algoritmus megfelelően működik.

Az eredmények kiértékelésénél egy grafikonon tüntettem fel, hogy mennyire jól sikerült a komplexben a peptid natív szekvenciáját az alternatívoktól elkülönítenem. Az eredmények szemléltetéséhez készítettem egy rangsort, ami során azt vizsgálom, hogy hány alternatív szekvenciának kedvezőbb, azaz kisebb az energiája a natív szekvenciákhoz képest, majd ezt a rangsort grafikonon ábrázolom. Akkor működik jól az algoritmusom, ha minél több komplex esetén kerül a natív szekvencia a rangsor elejére. A grafikon x tengelyére az adatbázisban szereplő komplexek kerültek, a rangsorban elért helyezésük szerint csökkenő sorrendben. Az y tengely a rangsorolást mutatja.

5.1. Globuláris fehérjék adatbázisa

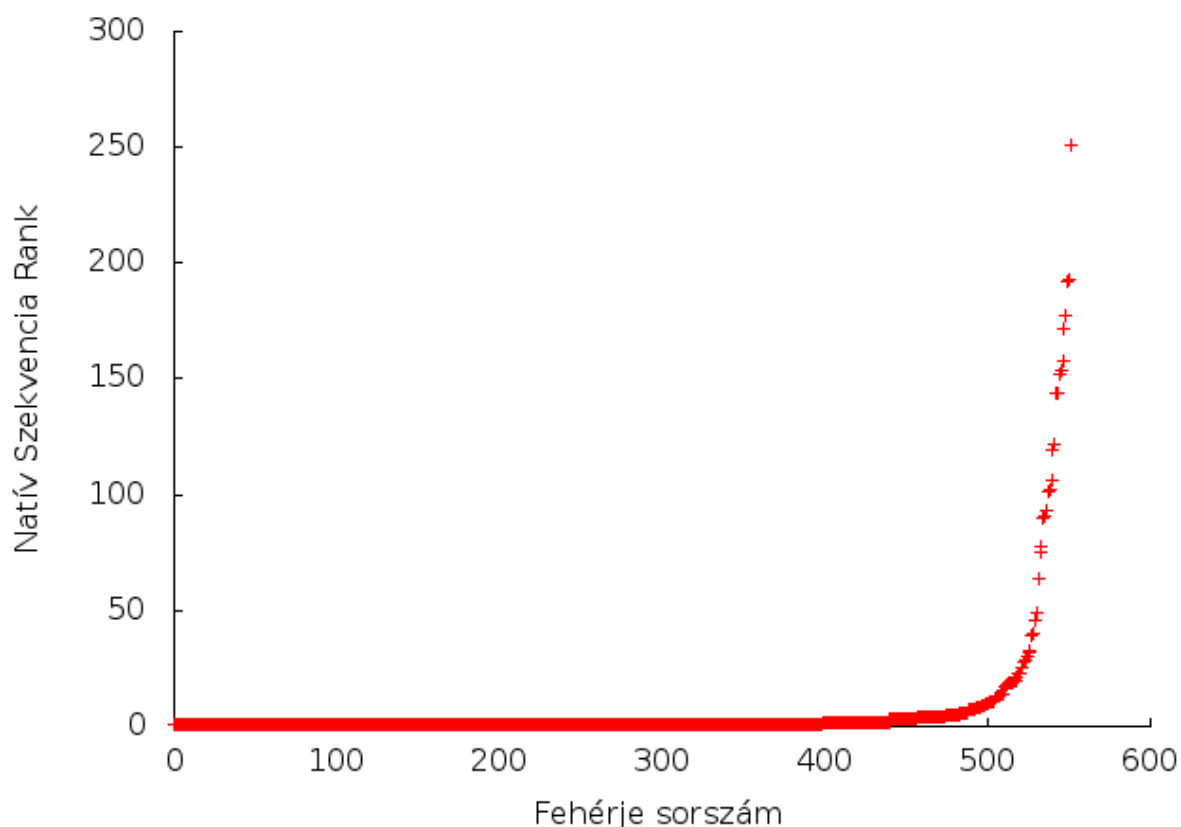
A programomat először az 553 globuláris fehérjéből összeállított adatbázison teszteltem. Az adatbázisban szereplő összes fehérje minden natív szerkezetéhez 1000 alternatív szekvenciát generáltam, a módszereknél bemutatott illesztési módszerrel. Először elvégeztem egy rangsorolást a kiindulási energiamátrixomat felhasználva. A rangsorolás eredményét a 3. ábra mutatja.



3. ábra. A kiindulási mátrixomat felhasználásával készült rangsorolás a globuláris adatbázis esetén.

A véletlenszerű energiaértékekkel történt rangsorolás a várt eredményt hozta, a natív szekvenciák helyezéseit egy egyenessel közelíthetjük.

Ezt követően az általam készített algoritmust felhasználva 100 iteráción keresztül optimalizáltam az energiafüggvényemet. Az ennek eredményeképpen kapott párkölcsönhatási energiákat tartalmazó statisztikus potenciál mátrixomat felhasználva a rangsorolást a 4. ábra szemlélteti.

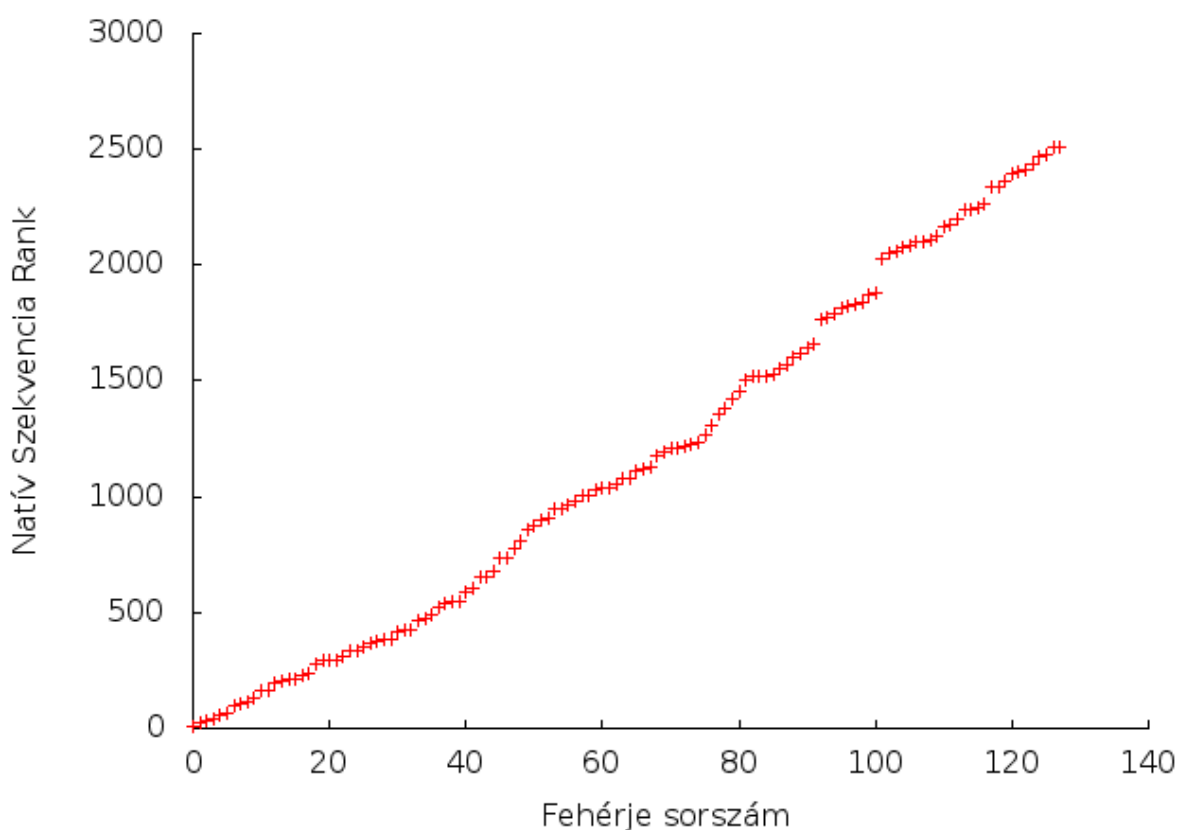


4. ábra. Az optimalizált mátrixom felhasználásával készült rangsorolás a globuláris adatbázis esetén.

A grafikonon jól látszik, hogy a natív szekvencia elkülönítése az alternatívoktól a legtöbb esetben sikeres volt. Az esetek több mint a felénél a natív szekvencia az első helyre került és az esetek 90%-ban a natív szekvenciák az első 10 közé kerültek a rangsorolás után.

5.2. Nagyobb peptid adatbázis

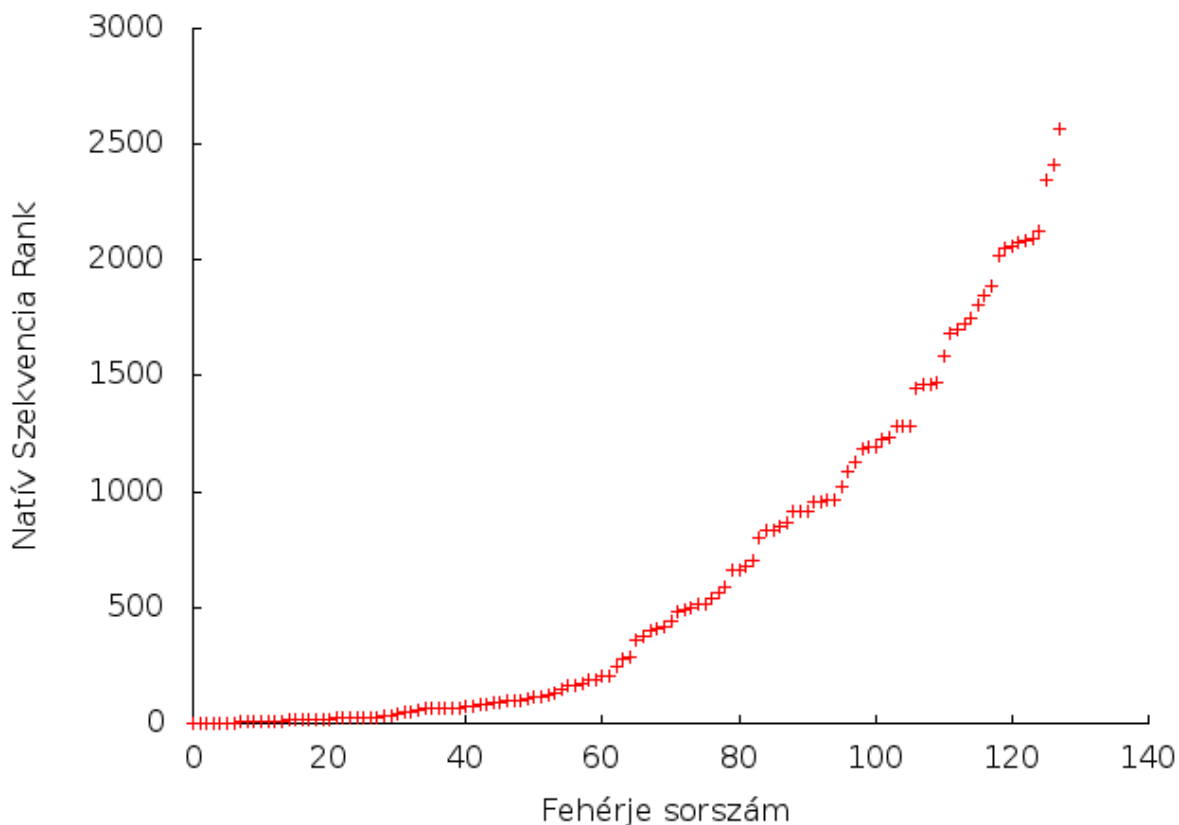
A programomat ez után a peptid-domén kölcsönhatások által kialakított komplexeket tartalmazó, nagyobb méretű adatbázison teszteltem. Ez az adatbázis 384 darab komplexet tartalmaz. Az adatbázist 3 részre osztottam, a 2/3-át felhasználva futattam az algoritmusomat. Az optimalizálás után eredményül kapott statisztikus potenciál mátrixot felhasználva az adatbázis maradék 1/3-ra pedig elkészítettem a rangsorolást. Ezt elvégeztem mind a három lehetséges esetben. Az adatbázisban lévő összes peptid natív szekvenciájához 10.000 alternatív szekvenciát generáltam. Az ábrákon csak az egyik felosztásnál kapott eredményeket szemléltetem. Az alábbi ábra (5. ábra) a kiindulási energiamátrix által elvégzett rangsorolást mutatja.



5. ábra. A kiindulási mátrixomat felhasználásával készült rangsorolás a nagyobb peptid adatbázis esetén.

A véletlenszerűen generált energiaértékekkel történt rangsorolás a várt eredményt hozta, a natív szekvenciák helyezéseit ismét egy egyenessel közelíthetjük.

Az általam fejlesztett programmal ezt követően 100 iteráción keresztül optimalizáltam az energiafüggvényemet. Az eredményül kapott energiamátrixommal újra elvégeztem a rangsorolást, ennek eredményét a 6. ábra szemlélteti.



6. ábra. Az optimalizált mátrixom felhasználásával készült rangsorolás a nagyobb peptid adatbázis esetén

A grafikonon látható, hogy javult a natív szekvencia rangsorolása. Az energiafüggvény optimalizálása után kapott energiamátrixot felhasználva a rangsoroláshoz a natív szekvenciák 10%-a az első 10 közé került, míg több mint egyharmaduk az első 100 közé.

Az adatbázisom további felosztásainál is hasonló eredményeket kaptam. Az optimalizáció során kapott energiamátrixok közötti korrelációt az alábbi táblázat mutatja.

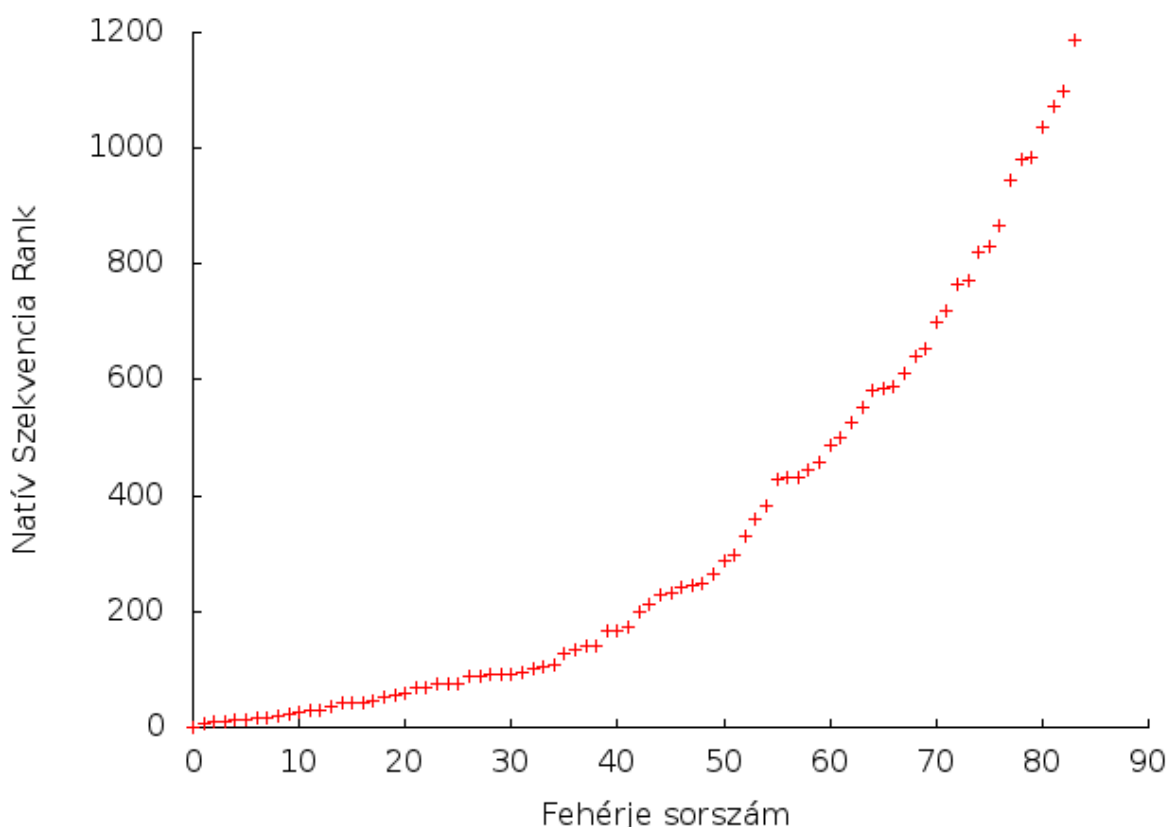
Listák:	lista1-lista2	lista1-lista3	lista2-lista3
Korreláció:	0.816804	0.738807	0.762425

7. ábra. Korrelációs táblázat az optimalizált energiamátrixok közt a nagyobb peptid adatbázisban .

A magas korrelációs adatokból arra lehet következtetni, hogy az adatbázis különböző felosztásaiból kapott statisztikus potenciál energiamátrixok nagyfokú azonosságot mutatnak.

5.3. Kisebb peptid adatbázis

Ez adatbázis specifikusan lineáris motívumon alapuló komplexet tartalmazott. A nagypeptid adatbázishoz hasonlóan itt is minden natív szekvenciához 1000 alternatív szekvenciát generáltam. A kiindulási mátrixommal történő rangsorolás során az eddigiekhez hasonló képet kaptam (3., 5. ábra). Az adatbázis kisebb mérete miatt itt nem volt lehetőség az adatbázis felosztására, ezért a nagyobb peptid adatbázisból az optimalizáció után kapott statisztikus potenciál energiamátrixszal végeztem el a rangsorolást. A kapott eredmények a 8. ábrán láthatóak.



8. ábra. A nagyobb peptid adatbázisban kifejlesztett energiamátrix felhasználásával készült rangsorolás a kisebb peptid adatbázis esetén

A grafikonon látható, hogy a natív szekvenciák rangsorolása az energiafüggvény optimalizálása után kapott statisztikus potenciál mátrixot felhasználva javult. A natív szekvenciák az esetek több mint egyharmadában az első 100 közé kerültek.

5.4. Véletlenszerűen generált alternatív szekvenciák

Az eddigi eredményeim során az alternatív szekvenciákat úgy állítottam elő, hogy a natív szekvenciákat összefűztem egy hosszú szekvenciába, majd ebből vágtam ki megfelelő méretű szakaszokat, ezzel megőrizve az aminosavak egymás melletti elhelyezkedésének törvényszerűségeit. A nagyobb, peptid-domén komplexeket tartalmazó adatbázisomat felhasználva megvizsgáltam, hogy milyen eredményeket kapok, ha az alternatív szekvenciákba véletlenszerűen rakom bele az aminosavakat tehát úgy tekintem, mintha minden aminosav megegyező valószínűséggel kerülne egymás mellé. A véletlenszerűen generált -1 és 0 energiaértékeket tartalmazó mátrixomból kiindulva az eddigiekhez hasonló eredményeket kaptam.

5.5. A kiindulási mátrix hatása az eredményekre

Az eddigi összes eredményemnél ugyan azt a véletlen generált -1 és 0 közé eső energiaértékeket tartalmazó kiindulási mátrixot használtam fel a számolásomhoz. Megvizsgáltam, hogy milyen eredményt kapok, ha más energiamátrixokból indulok ki. Megegyező adatbázisokat felhasználva különböző energiamátrixokból kiindulva futtattam a programomat. Míg a kiindulási energiamátrixok korrelációja mindössze -0.2 és 0.2 közé esett, 100 iteráció után, az eredményül kapott energiamátrixok korrelációja majdnem minden esetben 1 értékű lett. A kapott eredmény alapján a kiindulási mátrixokban szereplő energiaértékektől nem függ az optimalizáló kimenete, feltéve, ha az energiaértékek ésszerű határokon belül vannak.

5.6. Az eredmények értelmezése

A véletlenszerűen generált értékeket tartalmazó kiindulási energiamátrixot felhasználva a rangsoroláshoz minden esetben ugyan azt az eredményt kaptam. Eszerint, a natív szekvencia helyezése az alternatív szekvenciák között lényegében egyenletesen eloszolva, bárhol lehetett (3., 5. ábra).

Az algoritmussal optimalizált mátrixszal számolt rangsor a globuláris fehérjék esetén jelentősen javult, a natív szekvenciák a rangsorolás után az esetek több mint a felénél az első

helyre kerültek és az esetek 90%-nál az első 10-be kerültek (4. ábra). Néhány kivételes esettől eltekintve az energetikailag kedvező natív szekvenciát sikerült jól elkülöníteni az energetikailag kedvezőtlen alternatív szekvenciáktól. Az általam készített program, tehát jól alkalmazható az inverz folding problémára a globuláris fehérjék esetén.

A továbbiakban megvizsgáltam domén peptid kölcsönhatások esetén, hogy alkalmazható-e az algoritmus a natív peptid megtalálására véletlenszerűen generál alternatív szekvenciák között. A nagyobb peptid adatbázisból való peptid-domén komplexek esetén is láthatóan javul a rangsorolás, amit az algoritmusom futása után kapott energiamátrixszal számoltam, ahhoz képest, amit a véletlenszerűen generált energiaértékeket tartalmazó mátrixszal készítettem. Azonban a natív szekvenciák elkülönítése az alternatív szekvenciáktól nem volt olyan sikeres, mint a globuláris fehérjék esetén. Csupán az esetek 10%-ban került a natív szekvencia az első 10 közé, és az esetek egyharmadában az első 100 közé (6. ábra).

A kisebb peptid adatbázisból való, specifikusan lineáris motívumokon alapuló komplexek esetén is hasonló eredményeket tapasztaltam, mint a nagyobb peptid adatbázisom esetén. A kiindulási mátrixommal elvégzett rangsoroláshoz képest a kiszámolt energiamátrixommal javult a rangsorolás, de nem olyan jelentős mértékben, mint a globuláris fehérjék esetében. Az esetek több mint egyharmadában a natív szekvenciák a rangsorolás során az első 100 közé kerültek (8. ábra).

További vizsgálataim során megállapítottam, hogy az eredmények nem függtek lényegesen sem az alternatív szekvenciák generálásának a módjától, sem a kiindulási energiamátrixtól.

A kapott eredményekből azt a következtetést lehet levonni, hogy a programom a globuláris fehérjék natív szekvenciájának keresésére jól használható. A peptid-domén komplexek esetén az algoritmusom a peptidek natív szekvenciájának a megtalálásában javulást ért el, azonban a predikció nem volt olyan eredményes, hogy azt praktikusán fel lehessen használni. A további munkám során energiatagok bevezetésével és az entrópikus tényezők figyelembevételével igyekszem majd javítani a predikció eredményességén, továbbá a kapott energiamátrixok elemzésével szeretném megvizsgálni, hogy miben térnek el a domén-domén illetve peptid-domén kölcsönhatásokat stabilizáló tényezők.

6. Összefoglalás

A fehérjék az élőlények olyan komplex biológiai folyamataiban játszanak szerepet, mint a jelátvitel vagy az enzimatikus reakciók. Ha ezek nem megfelelően működnek, az komoly problémát okozhat a szervezetben, ezért a modern orvostudomány számára fontos a pontos megértésük. A szabályozási folyamatokban résztvevő fehérjék számára elengedhetetlen hogy megfelelő kapcsolatokat tudjanak kialakítani egymással. A fehérjék közt kialakuló kölcsönhatások egyik gyakori típusa a peptid-domén kölcsönhatás. Jelentőségük ellenére, az ilyen típusú kölcsönhatásokról azonban jóval kevesebb tudással rendelkezünk a globuláris fehérjék kölcsönhatásaihoz képest.

A célom egy olyan szerkezet alapú predikciós eljárás kifejlesztése volt, amely segítségével meg lehet jósolni, hogy ismert térszerkezetű domének milyen peptidekkel képesek kölcsönhatást kialakítani. Ehhez készítettem egy programot, ami a statisztikus potenciálok módszerét alkalmazza. Ismert térszerkezetű peptid-domén kölcsönhatások komplexeinek térszerkezeti adatait felhasználva konstruáltam egy energiafüggvényt, ami a párkölcsönhatási gyakoriságokon alapszik. Ezt követően energiafüggvényt olyan módon optimalizáltam, hogy minél jobban képes legyen specifikusan az adott doménekhez kötő és a véletlenszerűen generált nem kötő peptidek megkülönböztetésére.

A kapott eredményekből kiderült, hogy az általam készített algoritmus jól alkalmazható az globuláris fehérjék esetén a natív szekvencia azonosítására alternatív szekvenciák között. A peptid-domén kölcsönhatások esetén, bár a programom használatával sikerült jól látható javulást elérnem a predikcióban, az algoritmusom nem volt olyan eredményes a natív szekvenciák megtalálásába, mint a globuláris fehérjék esetén. Ennek az oka az lehet, hogy egy nagyon leegyszerűsített modellt használok. Azonban a jövőben további energiatagok bevezetésével és az entrópikus tényezők figyelembevételével tovább lehet javítani a predikció sikerességén.

A rendelkezésemre álló adatokból a peptid-domén kölcsönhatásokat stabilizáló tényezők feltárása még további vizsgálatokat igényel.

7. Köszönetnyilvánítás

Köszönettel tartozom Dr. Dosztányi Zsuzsannának, témavezetőmnek és Mészáros Bálintnak, a kitartó, fáradtságos munkájukért és hogy bevezettek a bioinformatika világába. Köszönöm továbbá Dr. Simon Istvánnak, a biológiai tudományok doktorának, hogy biztosította a lehetőséget az MTA Enzimológiai Intézetében való kutatáshoz.

8. Irodalomjegyzék

1. Ádám, Dux, Faragó, Fésüs, Machovich, Mandl, Sümegi (2006): "Orvosi Biokémia": Medicina Könyvkiadó Zrt
2. Tsytlonok M, Itzhaki LS (2013) The how's and why's of protein folding intermediates. Arch Biochem Biophys. 531:14-23.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank Nucleic. Acids Res. 28:235-42.
4. Dorottya Blaho (2008) X-Ray Crystallography of Biomacromolecules. Yale J Biol Med. 81: 51–52.
5. Mészáros B, Simon I, Dosztányi Z. (2011) The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. Phys Biol. 8:035003.
6. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. (2010) Transient protein-protein interactions: structural, functional, and network properties. Structure. 18:1233-43.
7. Tompa Péter (2009) Rendezetlen fehérjék amelyek új megvilágításba helyezik a szerkezet-funkció összefüggést. Magyar Tudományos akadémia folyóirata.
8. Uversky VN, Gillespie JR, Fink AL. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins. 41:415-27.
9. Dosztányi Z, Csizmók V, Tompa P, Simon I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 347:827-39.
10. Nooren IM, Thornton JM. (2003) Diversity of protein-protein interactions. EMBO J. 22:3486-92.
11. Mészáros B, Dosztányi Z, Simon I. (2012) Disordered binding regions and linear motifs--bridging the gap between two models of molecular recognition. PLoS One. 7:e46829.
12. London N, Movshovitz-Attias D, Schueler-Furman O. (2010) The structural basis of peptide-protein binding strategies. Structure. 18:188-99.
13. Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A, Jödicke L, Dammert MA, Schroeter C, Hammer M, Schmidt T, Jehl P, McGuigan C, Dymecka M, Chica C, Luck K, Via A, Chatr-Aryamontri A, Haslam N, Grebnev G, Edwards RJ, Steinmetz MO, Meiselbach H, Diella F, Gibson TJ. (2012) ELM--the database of eukaryotic linear motifs. Nucleic Acids Res. 40:D242-51.
14. Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, Born D, Speck T, Krüger D, Grebnev G, Kuban M, Strumillo M, Uyar B, Budd A, Altenberg B, Seiler M, Chemes LB, Glavina J, Sánchez IE, Diella F, Gibson TJ. (2013) The eukaryotic linear

motif resource ELM: 10 years and counting. *Nucleic Acids Res.*

15. Mészáros B, Simon I, Dosztányi Z. (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol.* 5:e1000376.
16. London N, Raveh B, Schueler-Furman O. (2013) Peptide docking and structure-based characterization of peptide binding: from knowledge to know-how. *Curr Opin Struct Biol.*
17. London N, Raveh B, Cohen E, Fathi G, Schueler-Furman O. Rosetta (2011) FlexPepDock web server--high resolution modeling of peptide-protein interactions. *Nucleic Acids Res.* 39:W249-53.
18. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.* 33:W382-8.
19. Thomas PD, Dill KA. (1996) Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol.* 257:457-69.
20. Thomas PD, Dill KA. (1996) An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci U S A.* 93:11628-33.
21. Wang G, Dunbrack RL Jr. (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* 33:W94-8.

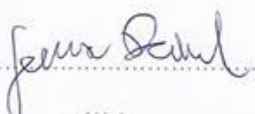
3. számú melléklet

Tételes szerzői hozzájárulás

Alulírott GÉMES DANIEL..... a dolgozatban leírt munkafolyamatok közül, az alábbiakat végeztem el:

- Programot írtam C++-ban amely kigyűjti a munkámhoz releváns adatokat a PDB formátumú fájlokat tartalmazó adatbázisokból.
- Programot írtam C++-ban, amely a térszerkezeti adatokból kinyert párkölcsönhatási gyakoriságokat felhasználva konstruál egy energiafüggvényt, majd ezt az energiafüggvényt olyan módon optimalizálja, hogy az minél jobban képes legyen egy adott szerkezethez tartozó natív szekvencia megtalálására.
- A rendelkezésemre álló adatbázisokon teszteltem a programot.
- Elvégeztem a kapott eredmények értékelését.

Budapest, 2014. 05.12


.....
aláírás