

Rendezetlen kötőmotívumok kinyerése és elemzése a PDB fehérjeszerkezet adatbázisból

XXXII. OTDK Biológia Szekció

Szepes Dóra

Molekuláris bionika mérnöki BSc, 4. évfolyam



Pázmány Péter Katolikus Egyetem Információs Technológiai Kar



Magyar Tudományos Akadémia Természettudományi Kutatóközpont

Enzimológiai Intézet

Témavezető: Kalmár Lajos

2013

Tartalomjegyzék

1. Bevezetés	3
1.1. Rendezetlen fehérjék	3
1.1.1. Fehérjék szerkezeti előfordulásai	4
1.1.2. Rendezetlen fehérjék funkciói.....	5
1.1.3. Rendezetlen fehérjék vizsgálati módszerei	6
1.1.4. Rendezetlen fehérjék tulajdonságai	8
1.2. Lineáris motívumok	10
2. Célkitűzés.....	11
3. Módszerek.....	12
3.1. Felhasznált adatbázisok	12
3.1.1. PDB adatbázis	12
3.1.2. UniProt adatbázis.....	12
3.1.3. ELM adatbázis.....	12
3.2. Felhasznált programok	13
3.2.1. IUPred.....	13
3.2.2. ANCHOR.....	13
3.2.3. USEARCH	14
3.2.4. MUSCLE	15
3.3. Adatok kinyerése a PDB adatbázisból	15
3.4. Adatok kinyerése az ELM adatbázisból.....	20
3.5. Rendezetlen kötőszakaszok jóslása SVM segítségével.....	20
3.6. Markov lánc használata	23
4. Eredmények.....	24
4.1. A PDB és ELM adatbázisból történő szűrés eredményei.....	24
4.2. Az SVM eredményei	27
5. Összefoglalás.....	30
6. Köszönetnyilvánítás.....	31
7. Irodalomjegyzék	32
8. Függelék.....	34

1. Bevezetés

1.1. Rendezetlen fehérjék

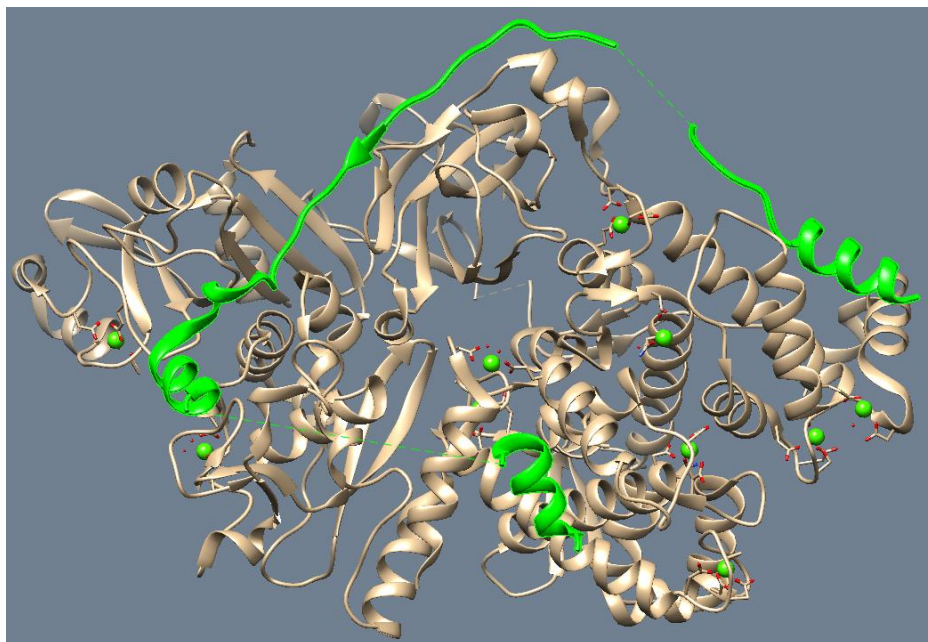
Az ezredfordulón felfedezett rendezetlen fehérjék megváltoztatták azt a nézetet, miszerint a normális funkció ellátásához a proteineknek egy jól definiált harmadlagos szerkezettel kell rendelkezniük. A rendezetlen fehérjéknek funkcionális formájukban sincs stabil harmadlagos szerkezetük, hanem folyamatosan egymásba alakuló konformációkkal rendelkeznek. Funkcióik ellátásához gyakran a partnereik felismerése, illetve ezen molekulák jelenlétében történő feltekeredés szükséges, de vannak olyan esetek is, amikor a flexibilitás ilyenkor is megmarad [1].

Számos rendezetlen fehérje hibás működése vagy hiánya köthető betegségekhez, vagyis a terület orvosbiológiai szempontból is igen jelentős. Gyakran találnak rendezetlen fehérjéket neurodegeneratív betegségek hátterében (Alzheimer kór, Parkinson kór, Prion betegség), illetve a szabályozó rendszerekben betöltött kulcsszerepük miatt daganatos körképek hátterében is állhatnak (pl. p53 fehérje) [9].

Általánosságban elmondható, hogy a komplexebb, a törzsfajlás magasabb szintjén álló organizmusokban nagyobb számban fordulnak elő rendezetlen régiókat tartalmazó fehérjék. Eukariótákban nagy arányban találhatók meg, a proteom mintegy 5-15%-a teljesen rendezetlen fehérje, valamint 40-60%-a legalább egy hosszabb rendezetlen régiót tartalmazó fehérjéből áll [1]. Ennek oka egyes elméletek szerint az, hogy a fejlettebb organizmusokban nagyobb szükség van a gyakran rendezetlen fehérjék által mediált szabályozó, kommunikációs folyamatokra (pl. jelátvitel).

Bár a fehérjék rendezetlensége első ránézésre feleslegesnek tűnhet a hosszú időn át evolválódott funkcionális globuláris doménekkel szemben, valójában több előnye is van [10]. Az egyik, hogy magas specificitású, de tág határok között finomhangolható affinitású kötéseket hozhat létre, ez pedig hasznos az olyan folyamatok során, ahol csak átmeneti kapcsolatra van szükség a partnerek között. Másik előnye, hogy a flexibilitásnak köszönhetően egy rendezetlen fehérje különböző partner molekulákhoz kapcsolódhat, nagymértékű rugalmasságának köszönhetően idomulni tud több eltérő alakú partnerhez is. Ez természetesen a globuláris fehérje oldaláról nézve is igaz, azaz egy kötőfelszínhez akár több szekvenciában igen különböző partner is kötődhet. A nagy kötőfelülettel rendelkező rendezetlen fehérjék körül tudják tekerni, vagy körbe tudják venni partnereiket (1. ábra). Nagy

előny, hogy mivel a térbeli orientációtól kevésbé függ a kapcsolat létrejötte, az gyorsabban tud létrejönni, valamint a disszociáció is gyorsabban történik.

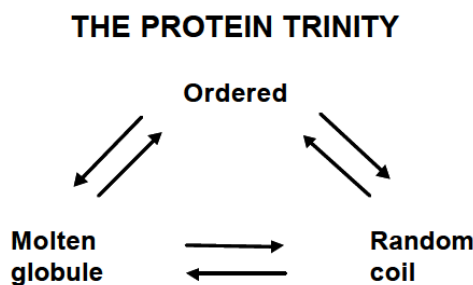


1. ábra: Calpastatin és n-calpain komplex a PDB adatbázisból (ID: 3BOW).

Jól látható, hogy a zölddel kiemelt rendezetlen lánc (calpastatin) körbeöleli a globuláris láncokat, a több kötőhely által biztosított kapcsolat ebben az esetben a nagy mértékű specificitás mellett igen nagy affinitású is.

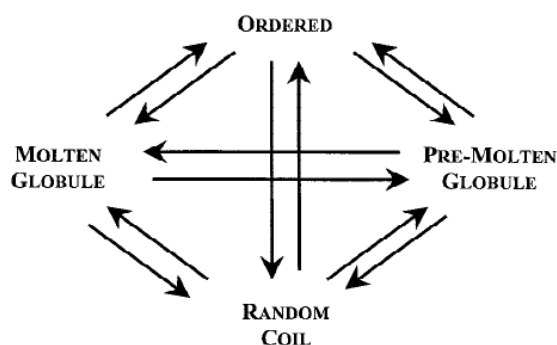
1.1.1. Fehérjék szerkezeti előfordulásai

Egy felvetés szerint a fehérjék háromféle natív állapotban fordulhatnak elő, ezek a rendezett, az úgynevezett molten globula és a random coil állapotok (2. ábra). Tudni kell, hogy tényleges random coil nem jöhet létre a természetben, ezért ez alatt az elnevezés alatt a fehérjék random coil-szerű szerkezetét értjük. Ezt az elképzelést fehérje háromságnak (protein trinity) szokás hívni, melynek lényege, hogy a fehérjék szabadon alakulhatnak át egyik állapotból a másikba, és mindegyik képes lehet funkció ellátására [10]. Egy adott funkció alapulhat az egyik állapoton, vagy akár kettő közti átmeneten is. A rendezett fehérjék jellemzően keletkezésükkor, a feltekeredés folyamatában mennek át ezeken a lépéseken, később a stabil harmadlagos szerkezetüket felvéve már abból az állapotból normál körülmények között nem kerülnek ki. A rendezetlen fehérjék igazán rendezett szerkezetet magukban nagyon ritkán vesznek fel (esetleg kóros amiloid szerkezetet), stabil, merev állapotban csak komplexben kötött formában vannak.



2. ábra: A fehérje háromság szerkezeti állapotai.

Egy másik elképzelésben egy újabb szerkezeti állapot is megjelenik, mégpedig a premolten globula. Emiatt ezt fehérje kvartettnek (protein quartet) nevezzük (3. ábra) [12]. A premolten globula leginkább denaturált állapotra emlékeztet, de képes funkció ellátására, és kompaktabb a random coilnál. Nem rendelkezik stabil harmadlagos szerkezettel, azonban másodlagos szerkezeti elemek találhatók benne, bár kisebb mennyiségben, mint a rendezett vagy molten globula állapotokban.



3. ábra: A fehérje kvartett szerkezeti állapotai.

1.1.2. Rendezetlen fehérjék funkciói

A rendezetlen fehérjék fő funkciója a fehérje-fehérje kölcsönhatások kialakítása, ilyenkor általában egy rövid peptidszakasz kapcsolódik a rendezett partnerhez. A rendezetlen fehérjék gyakran látnak el szabályzó szerepet, például a transzkripció, transzláció során, valamint önfelépítő komplexek összeállása során segédkeznek (állványfehérjék pl. a riboszóma kialakulása során) [9]. A rendezetlen fehérjéknek a sejten belüli jelátvitelben is fontos szerepük van, egyrészt kiválóan alkalmasak átmeneti kötődések kialakítására, így a jel átvételére és továbbadására, valamint annak köszönhetően, hogy a bennük található aminosavak jól hozzáférhetők enzimek számára, gyakran történik rajtuk poszttranszlációs módosítás. Egy új ága a rendezetlen fehérjék kutatásának a rendezetlen chaperonok témaköre,

több teljes mértékben rendezetlen fehérjéről kimutatták már, hogy ATP felhasználása nélkül képesek globuláris fehérjék térszerkezetének helyreállításában közreműködni [8].

A funkcióik szerint alapvetően négy kategóriába sorolhatjuk a rendezetlen fehérjéket [10]. Az első az úgynevezett „entropic chain”, azaz entropikus lánc. Ezeket a funkciókat kizárólag rendezetlen fehérjék tudják elvégezni. Ide tartoznak például a linker fehérjék, melyeknek szerepe a multidomén fehérjékben az egyes domének összekötése úgy, hogy azok egymáshoz képest elmozdulhassanak (emiat van szükség a flexibilis rendezetlen fehérjékre). A spacer fehérjék (pl. mikrotubulus asszociált fehérje) fehérje doméneket, vagy nagyobb fehérje komplexeket kötnek össze, szigorúan szabályozva az azok közti minimális távolságot, meggátolva összetapadásukat. Ezek a rendezetlen fehérjék tehát a funkciójuk ellátása közben sem vesznek fel egy rendezett szerkezetet, végig random coil jellegű konformációban maradnak.

A második funkció, melyet igen gyakran rendezetlen szakaszok látnak el, a poszttranszlációs módosítások célszakaszai. Ezek a módosítások nagy szerepet játszanak a sejtek szabályozó folyamataiban, és mivel a nem feltekeredett rendezetlen szakaszok igen jól hozzáférhetők más enzimek számára, így mind a kémiai addíció (foszforiláció, glikoziláció, stb.), mind a proteáz hasítás jobban jellemző ezekre a szakaszokra.

A harmadik kategória a molekuláris összeszerelés. Megfigyelték, hogy főként a nagy méretű komplexek összeszerelésében a részlegesen rendezetlen alegységek előnyt élveznek a teljesen rendezett alegységekkel szemben. Ennek oka, hogy sokkal könnyebben véghezvihető a komplex létrehozása úgy, rugalmas alegységek mintegy ragasztóként behálózva a komplexet több ponton biztosítják annak integritását.

Az utolsó funkció a molekuláris felismerés, melyben legtöbbször a rendezetlen fehérjén lévő kötőmotívum a rendezett fehérje specifikus kötőfelszínéhez kapcsolódik. Ez a funkció kulcsfontosságú a legtöbb sejten belüli szabályozó folyamatban, a kötődés finomhangolásával (pl. poszttranszlációs módosítással) jól szabályozható a kötődés erőssége.

1.1.3. Rendezetlen fehérjék vizsgálati módszerei

A rendezetlen fehérjék kísérleti úton történő vizsgálata korlátozott, mivel a globuláris fehérjékre kidolgozott módszerek közül több nem alkalmas a gyorsan változó szerkezettel bíró proteinek hatékony elemzésére. Bár rendelkezésre állnak olyan módszerek, melyekkel ez

mégis lehetséges (főleg több kombinálásával) [11], azonban ezek magas költségigénye miatt a bioinformatika a rendezetlen fehérjék vizsgálatában komoly szerepet kap.

A röntgen krisztallográfia, mellyel elsősorban rendezett, kristályos szerkezetű anyagok szerkezete vizsgálható, korlátozott hatékonysággal rendezetlen fehérjék esetében is használható. A rendezetlen régiók flexibilis atomjairól a röntgensugarak inkoherens módon szóródnak szét, ezért ezek helyén hiányzó elektronsűrűséget lehet tapasztalni. Ezt a jelenséget azonban más is okozhatja, így érdemes más módszerekkel is bizonyítani egy ilyen szakasz rendezetlenségét.

Gyakran használják az NMR (nuclear magnetic resonance, mágneses magrezonancia) spektroszkópiát is. Ennek segítségével szilárd vagy oldatban előforduló minták 3D-s szerkezetét lehet megállapítani. Rendezetlen fehérjék vizsgálatánál sokszor a ^{15}N - ^1H heteronukleáris NOE (nuclear overhauser effect) mérést alkalmazzák. Ennél a módszernél az eredményben a rendezett régiók pozitív, a rendezetlenek negatív jelet adnak, így könnyen elkülöníthetők. A fehérjék aminosavainak asszignálására, a különböző hatások (pl. fehérje-fehérje interakció) követésére használható HSQC spektrumban is nagy különbség tapasztalható rendezett és rendezetlen fehérje között, a rendezetlen fehérjék jóval szűkebb jeldiszperziót mutatnak. Azonban az NMR spektroszkópiának is van hátránya, mégpedig az, hogy a molten globula állapotú proteineket nehéz vele vizsgálni. Ennek egyik fő oka, hogy a méréshez szükséges koncentrációban már aggregálódnak ezek a fehérjék.

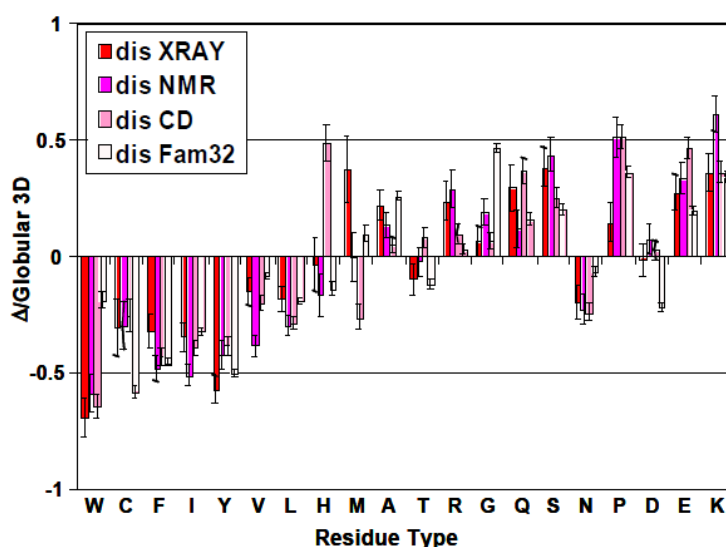
A CD (circular dichroism, cirkuláris dikroizmus) spektroszkópia alapja az a fizikai tulajdonság, hogy a jobbra vagy balra cirkulárisan polarizált fény elnyelésében különbségek mutatkoznak másodlagos szerkezeti elemek között. Segítségével oldatban levő minták szerkezeti információit vizsgálhatjuk. A módszernek több fajtája is van, az egyik ilyen a távoli UV CD, mellyel a másodlagos szerkezetről nyerhető információ, egy másik pedig a közeli UV CD, ami aromás csoportokat vizsgál. A kettő kombinálásával a fehérjék különböző állapotai elválaszthatók egymástól, mivel a távoli a random coil fehérjéket különbözteti meg a többitől, a közeli pedig a rendezetteket. Azonban pont ez a hiányossága is, hiszen többféle elemet tartalmazó proteineket nehézkesen vizsgálni vele, mivel ilyenkor sokszor egy kevert spektrumot kapunk.

Egy aránylag egyszerűen kivitelezhető módszer a limitált proteáz emésztésen alapuló vizsgálat. Amikor egy rendezett, globuláris fehérje proteáz hasítása történik, annak előbb le kell tekerednie, hogy az enzim a megfelelő helyen hozzá tudjon férni. Azonban a rendezetlen

fehérjék esetében erre nincs szükség, mivel alaphól nincsenek feltekeredve, emiatt az emésztés gyorsasága nagyságrendekkel nő. Ennek segítségével akár pozíció specifikus információt is kaphatunk a rendezetlen régiókról. Természetesen a módszerrel csak olyan fehérjék vizsgálhatók, melyekben a proteáz enzimek által felismert szekvencia található.

1.1.4. Rendezetlen fehérjék tulajdonságai

A rendezetlen fehérjék esetében általánosságban elmondható, hogy a rendezett társaiktól eltérő, jellemző aminosav eloszlás figyelhető meg bennük (5. ábra).

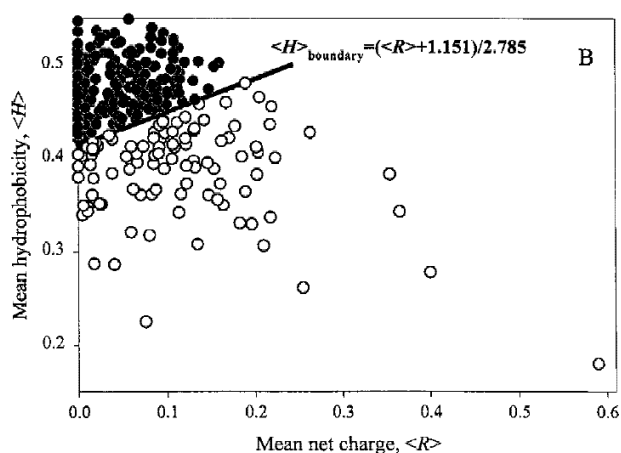


5. ábra: A rendezetlen fehérjék aminosav eloszlása a globuláris fehérjékhez képest különböző vizsgálati módszerek alapján [11].

A grafikonon jól megfigyelhető, hogy a triptofán, cisztein, fenil-alanin, izoleucin, tirozin, valin, leucin, vagis általánosságban a hidrofób aminosavak kisebb mennyiségben fordulnak elő, ezzel szemben az hidrofil, poláros, töltött, vagy kicsi és mozgékony aminosavak jóval gyakoribbak. Emiatt az előbbi csoportba tartozókat rendezettséget elősegítő, míg az utóbbiakat rendezetlenséget elősegítő aminosavaknak nevezzük. A hisztidin, metionin, treonin és aszparaginsav mennyiségében nincs jelentős eltérés a kétféle fehérje csoport között [11]. Megfigyelték azt is, hogy a rendezetlen fehérjék gyakran kisebb komplexitásúak a rendezett társaiknál.

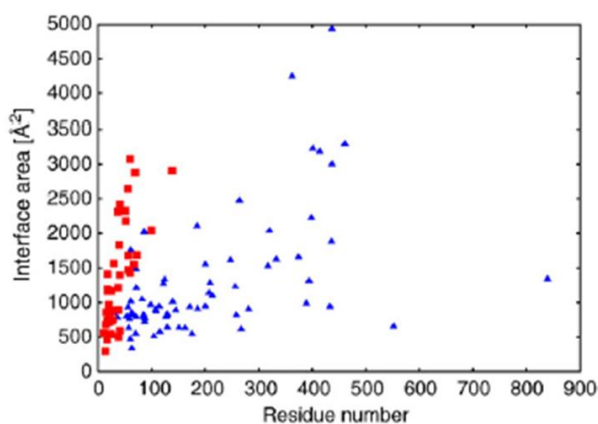
Általánosságban megállapítható tehát, hogy a rendezetlen régiókra a kevés hidrofób, több hidrofil, valamint gyakori töltött aminosavak jellemzőek [8]. Ha a hidrofobicitás és a

töltés alapján rendezetlen és rendezett fehérjéket egy grafikonon ábrázolunk, jól látható lesz, hogy könnyen elkülöníthetőek ezen tulajdonságok alapján (6. ábra).



6. ábra: Az x tengelyen az átlagos nettó töltés, az y tengelyen az átlagos hidrofobicitás szerepel. A fekete körök a rendezett, a fehér körök a rendezetlen fehérjéket jelölik. Ezt a grafikai megjelenítést kitalálója után „Uversky plot”-nak nevezzük [12].

A rendezetlen fehérjék szerkezetéből adódik az is, hogy a kevés aminosavból álló rendezetlen fehérjék, fehérje szakaszok is rendelkezhetnek nagy interakciós felülettel (7. ábra). Ez egyben azt is jelenti, hogy ugyanakkora interakciós felület kialakításához egy rendezetlen fehérjének sokkal kevesebb aminosavra van szüksége, mint egy globuláris fehérjének, vagyis jóval hatékonyabb (evolúciósan költséghatékonyabb) rendezetlen régióból kialakítani kölcsönható szakaszt.



7. ábra: A kötőfelszín az aminosavak számának függvényében ábrázolva. A piros négyzetek a rendezetlen, míg a kék háromszögek a globuláris fehérjéket jelölik [14].

1.2. Lineáris motívumok

A lineáris motívumok (linear motif, LM) olyan rövid fehérje szakaszok, melyek a leggyakrabban rendezetlen régiókban találhatók, és funkcióikat harmadlagos szerkezet nélkül látják el. Másik elnevezésük a rövid lineáris motívum (short linear motif, SLiM).

Ezek a rövid motívumok tehát a rendezetlen kötőhelyek egyik alkategóriáját képviselik, és fontos feladatokat látnak el többek között a jelátvitelben, fehérje-fehérje interakcióknál, és szabályzó szerepeik is vannak [6].

Jellemző rájuk, hogy a körülöttük található régióknál, evolúciósan jobban konzerváltak. Sokszor másodlagos szerkezeti elemeket alakítanak ki a partnerhez való kapcsolódáskor, ezt a folyamatot a határoló régiók segíthetik. Két fontos tulajdonságuk is következik abból, hogy ezekben a motívumok általában rövidek, kevés konzervált aminosavból állnak. Egyrészt hajlamosak a konvergens evolúcióra, vagyis közös őstől könnyedén kialakulhat több fehérjében ugyanazt a funkciót ellátó motívum. Másrészt aránylag kis affinitással bírnak a partner molekulákhoz való kötődéskor, emiatt könnyen hoznak létre tranziens, átmeneti kapcsolatokat [7].

Megfigyelték, hogy az átlagos rendezetlen aminosav kompozícióhoz képest kötőmotívumokban gyakrabban fordulnak elő hidrofób aminosavak, a töltöttek pedig kevésbé gyakoriak. Ezzel szemben a motívumot határoló régiókban a hidrofil aminosavak a jellemzőek, aminek az evolúciós magyarázata az, hogy míg a hidrofób motívum hajlamos lenne az aggregációra, amiloid képzésre, addig a flankingben található hidrofil aminosavak ezt meggátolják [13].

2. Célkitűzés

A továbbiakban ismertetett munkám során az elsődleges feladatom a rendezetlen kötő régiók PDB adatbázisból való kiszűrése volt. A feladat elvégzésekor figyelnem kellett arra, hogy a PDB adatbázisban található számos hibás adatot kiszűrjem. Ezen kívül a programok írása közben arra is törekedtem, hogy az így elkészített adatbázis a lehető legkevesebb emberi beavatkozással tudjon frissülni, amit az indokol, hogy a PDB hetente frissül. Bár munkám jelentős részét a PDB adatbázis feldolgozása tette ki, és ez elsősorban a munkám eredménye is, ennek tárgyalását a módszerek fejezetbe illesztettem a munka jellege miatt.

Az adatok kiszűrése után azok feldolgozása is munkám része volt. Egyrészt a kötő motívumok, valamint az azokat határoló régiók fizikai és kémiai tulajdonságait vizsgáltam bioinformatikai módszerekkel, másrészt egy olyan algoritmus elkészítése is a feladatom volt, mely segítségével rendezetlen kötőrégiók ismerhetők fel.

A munkám eredménye a csoportban folyó átfogó kutatásának szerves részét képezi majd. Ennek során elsősorban a rendezetlen motívumok kialakulását, általános felépítését, fizikai és kémiai tulajdonságaikat, valamint a működésünkhöz szükséges kulcs faktorokat vizsgáljuk. A bioinformatikai módszerekkel kapott eredményeket később kísérletes úton is szeretnénk bizonyítani.

3. Módszerek

3.1. Felhasznált adatbázisok

3.1.1. PDB adatbázis

A PDB, azaz a Protein Data Bank egy olyan hetente frissülő adatbázis, mely fehérjék, nukleinsavak, valamint ezek komplexeinek szerkezeti információit tárolja. Ezeket az adatokat elsősorban röntgenkristallográfia, valamint NMR spektroszkópia segítségével nyerik ki, és 3D atomkoordináták formájában tárolják. Mára már több mint 95 000 makromolekula struktúráját tartalmazza.

3.1.2. UniProt adatbázis

A UniProt, vagyis Universal Protein Resource adatbázisban fehérje szekvenciákat és ezek annotációit tartalmazza. Nagy mennyiségű információ található benne a proteinek biológiai funkciójáról is. Az adatbázis négy különböző adatbázisból tevődik össze, ezek a következők: UniProtKB, UniParc, UniRef, és UniMES.

A UniProtKB (UniProt Knowledgebase) a központi része az adatbázisnak, nagyon jól annotált szekvenciákat tartalmaz. Az adatok egyik részének annotációja manuálisan történik (SwissProt), míg a másiké számítógépes analízis segítségével (TrEMBL). A UniParc (UniProt Archive) egy nem-redundáns adatbázis, mely a legtöbb publikus fehérje szekvenciát tartalmazza. Ez csakis a szekvenciákat tartalmazza, további információk az adatbázis többi részéből nyerhetők ki.

A UniRef (UniProt Reference Clusters) a UniProtKB-ből valamint a UniParcból származó fehérjék clusterezett csoportjait tartalmazza. A UniMES (UniProt Metagenomic and Environmental Sequences) adatbázis specifikusan metagenomikus és környezeti adatok tárolására szolgál.

3.1.3. ELM adatbázis

Az ELM, azaz Eukaryotic Linear Motif resource egy olyan adatbázis, amely eukariótákban található rövid lineáris motívumokat tartalmaz, több mint 2000 bejegyzés található benne [3].

Az adatbázisban 6 csoportra vannak osztva a lineáris motívumok. Az egyik a LIG (ligand site), ebbe az osztályba ligandumként szolgáló motívumok tartoznak. Egy másik csoport a MOD (modification site), mely a poszt-transzlációs modifikáció helyeit tartalmazza. A CLV (cleavage site) csoportban a proteolitikus hasítási és feldolgozási folyamatokban résztvevő motívumok szerepelnek. A TRG (targeting site) osztályban a sejten belüli célzással kapcsolatos szakaszok találhatók. A DEG (degradation site) és a DOC (docking site) csoport a degradációért és dokkolásért felelős motívumokat tartalmazza.

Az adatbázisból való lekérdezéskor bemenetként egy vagy több szekvenciát, illetve UniProt ID-t adhatunk meg. Ezen kívül lehetőség van egyéb adatok megadására is, ilyen a sejten belüli elhelyezkedés, valamint a faj. Kimenetként a szekvenciában talált motívumokat és azok helyét kapjuk meg különböző információkkal kiegészítve [2].

3.2. Felhasznált programok

3.2.1. IUPred

Az IUPred segítségével egy adott aminosav szekvencia rendezetlensége jósolható. A módszer a polipeptidek stabilizáló kapcsolat kialakítására való hajlamát becsli meg [4]. Az alapelv az, hogy a globuláris fehérjéknél sok ilyen összeköttetés keletkezik, azonban a rendezetlen fehérjék nem képesek hatékonyan ilyen kapcsolatok létrehozására. Az IUPred algoritmus az interakciós energiák kiszámításával jósolja az egyes szakaszok rendezetlenségét.

A szerzőtől elkért programot csoportunkban kis mértékben átalakítva használjuk (kizárólag a kimeneti formátum lett megváltoztatva, a predikciós algoritmus nem), mely bemenetként mindig FASTA formátumban adtam meg a szekvenciákat, a kimenet pedig egy olyan fájl, melyben az eredeti adatok kiegészülnek szekvenciánként egy sorral, ami minden aminosav rendezetlenségi értékét tartalmazza. Ez az érték annak a valószínűségét mutatja, hogy az adott aminosav rendezetlen-e vagy sem, értéke 0 és 1 között lehet, a program által felajánlott alapértelmezett határérték 0,5, az efeletti régiókat, aminosavakat vesszük rendezetlennek.

3.2.2. ANCHOR

Az ANCHOR segítségével rendezetlen kötőrégiókat jósolhatunk. Ez az algoritmus olyan szakaszokat keres, melyek rendezetlen régiók részei, tehát nem tudnak önmagukban feltekeredni, azonban nagy lokális interakciós potenciáljuknak köszönhetően nagy

valószínűséggel kötődhetnek egy globuláris partnerhez [5]. A módszer az IUPrednél is használt energiabecslésen alapul, biofizikai tulajdonságokra koncentrál. Az alapelv az, hogy a rendezetlen fehérjék kevésbé kedvező energiaállapotokban vannak, de kedvezőbb állapotba kerülnek, ha egy partnerrel kapcsolódnak.

A jóslás három fő kritériumon alapszik. Az első az, hogy a vizsgált szakasz egy hosszabb rendezetlen régió része legyen, a második biztosítja, hogy az adott aminosav ne tudjon elegendő kapcsolatot létesíteni a környezetével, azaz ne legyen képes önmagában feltekeredni, ne rendelkezzen jól definiálható szerkezettel. Az utolsó kitétel azt teszteli, hogy egy aminosavnak mennyi esélye van elegendő kötés kialakítására egy globuláris partnerrel való kapcsolódáskor.

Bemenetként megadhatunk egyetlen vagy több szekvenciát (akár FASTA formátumban), de UniProt ID alapján is végezhetjük a jóslást. Fontos megemlíteni, hogy a partner molekuláról nem szükséges információt megadni, a program sem jósol konkrét kötőpartnert, csupán az interakciós készséget. A kimenetben minden sorban egy aminosav szerepel, a pozíciójával, a rendezetlenségi értékkel, és azzal az információval kiegészítve, hogy kötő aminosavnak lett-e predikálva. Ezen kívül grafikus ábrázolást is kérhetünk, ekkor egy görbe mutatja az IUPred által jósolt rendezetlenségi valószínűséget, valamint egy másik annak a valószínűségét, hogy az adott aminosav egy rendezetlen kötőrégió része.

3.2.3. USEARCH

A USEARCH algoritmus segítségével nagyon hatékonyan rendezhetünk csoportokba fehérje szekvenciákat. Az algoritmus a magas pontszámmal rendelkező globális illesztéseken alapszik [16]. Egy adott lekérdezés esetében az adatbázisban található szekvenciák csökkenő szószám alapján vannak rendezve, mivel a hasonló szekvenciák közös részei általában rövid szavak. Emiatt hogyha van megfelelő találat a lekérdezett szekvenciához, akkor ez az első néhány jelölt között lesz, és az esély arra, hogy van megfelelő találat, nagyon gyorsan csökken a sikertelen próbálkozások számának növekedésével. Tehát az algoritmust néhány jelölt vizsgálata után le lehet állítani, így nagyon gyors.

A felhasználó határozza meg, hogy mekkora mértékű hasonlóság esetén tekintsen az algoritmus két szekvenciát hasonlóknak. Ezen kívül több paramétert is változtathat, például azt, hogy hány sikertelen próbálkozás után álljon le az algoritmus.

3.2.4. MUSCLE

A MUSCLE (multiple sequence comparison by log-expectation) egy olyan program, mellyel fehérje szekvenciák többszörös illesztését készíthetjük el. A sebessége és a gyorsasága is jobb, mint más hasonló funkcióval rendelkező programoknak (pl. CLUSTALW), ezért is ezt választottam [15]. Az algoritmusnak három szakasza van, mindegyik végén elérhető egy többszörös alignment, és az algoritmus leállítható.

Az első szakasz egy progresszív illesztést készít. Először kiszámítja a szekvenciák páronkénti hasonlóságát, majd egy távolsági mátrixot hoz létre ezek alapján. Ezután egy fát készít, melynek meghatározza a gyökerét. A progresszív illesztést ezen fa alapján készíti el a szakasz végén.

A második lépésben megpróbálja ezt a fát finomítani, és új, javított progresszív alignmenteket készíteni. Ez a szakasz többször ismételhető. A szekvenciák páronkénti hasonlóságát a már meglevő illesztés alapján számolja ki. Ezután egy új fát állít össze egy új távolsági mátrix használatával. A régi és az új fákat összehasonlítja, és megállapítja, hogy hol történt változás. Ha többször ismételtük ezt a szakaszt, és úgy ítéli a program, hogy a fák konvergálnak, akkor befejeződik az iteráció. A végén egy új progresszív alignment készül.

Az utolsó szakasz során még tovább próbálja finomítani az algoritmus az eredményt. Először két csoportra osztja a szekvenciákat, majd az aktuális többszörös illesztésből kinyeri a program az egyes alcsoportokhoz tartozó illesztéseket. Ezt a két alignmentet aztán egymáshoz igazítja. Hogyha így jobb lett az illesztés, akkor megtartja, majd másféle felosztással megismétli a lépéseket, amíg az összesen végig nem ér, vagy a maximális iteráció számot el nem éri.

3.3. Az adatok kinyerése a PDB adatbázisból

A munkám során a saját készítésű programokhoz végig a perl programozási nyelvet használtam. A teljes folyamatot a függelékben található ábra tartalmazza. Az első lépés elején rendelkezésemre állt letöltve az aktuális PDB adatbázis összes állománya. Először a komplexeket szerettem volna kiszűrni, tehát az első lépésben a PDB fájlok „DBREF” sorait vizsgáltam meg. Ezekben a sorokban egyéb információk mellett az szerepel, hogy a fehérje egy adott láncához milyen UniProt ID tartozik. Itt tehát olyan eseteket kerestem, melyeknél ezekben a sorokban legalább két különböző UniProt ID szerepel, azaz több fehérje,

feltehetően komplexben szerepel a szerkezetben. Ha ilyet találtam, a PDB és a UniProt ID-t is kiírtam egy fájlba.

A következő lépéshez rendelkezésemre álltak a UniProt adatbázisból letöltött szekvenciák (a hozzájuk tartozó információkkal). Ebből az ID-k alapján csak azokat a szekvenciákat kerestem és tartottam meg, melyek szerepeltek az előzőleg elkészült (PDB adatbázisból kiszűrt) fájlban. Erre a lépésre egyrészt azért volt szükség, mert míg a UniProt adatbázisban a teljes szekvencia szerepel, addig a PDB-ben sokszor csak egy szakasz, másrészt, tapasztalataink alapján a PDB adatbázis gyakran tartalmaz hibás adatokat (például az adott helyen más aminosav található, vagy a helyzetük el van csúszva), a UniProt adatbázis alapvetően megbízhatóbb forrás, tehát érdemes összevetni a kettőt.

Ezután a kiszűrt szekvenciákon az IUPred program segítségével elvégeztem a rendezetlenség jóslását. Így egy olyan fájlt kaptam, melyben szerepelt a szekvencia, alatta pedig az adott aminosavhoz tartozó rendezetlenséget jelző, 0 és 1 közötti valószínűség. Ez alapján kiszűrtem a rendezetlennek tekinthető szekvenciákat. Ehhez egy 25 aminosav hosszú ablakkal végigmentem az adott láncon, és akkor tekintettem rendezetlennek, ha minden ablakban legalább 23 aminosav IUPred értéke 0,6 felett volt (tehát egy valamivel szigorúbb határt használtam a hagyományos 0,5-ös határértékhez képest).

Miután ez megtörtént, már csak azokkal a komplexekkel dolgoztam tovább, melyekben legalább egy (az előzőekben kiszűrt) rendezetlen lánc szerepelt. Ezekben a PDB szerkezetekben kiszámoltam a rendezetlen és a többi lánc nehézatomjainak (nem hidrogén) távolságát, a PDB fájl „ATOM” kezdetű sorainak segítségével. Az „ATOM” sorokban többek között az atom, az azt tartalmazó aminosav, valamint az atom térbeli elhelyezkedése szerepel. Ha két atom közti távolság kisebb, mint 6Å, akkor képesek lehetnek kötést kialakítani egymás között. Ez alapján kigyűjtöttem azokat az aminosavakat, melyek kapcsolódhatnak egymáshoz (a feltétel az volt, hogy mindkettőben legalább 1-1 nehézatom elég közel legyen egymáshoz). Az ebben a lépésben elkészített kimeneti fájlomban minden sorban szerepel az adott PDB ID, az aminosav neve, a lánc betűjele, az aminosav helyzete a láncban, valamint az ehhez kötődő aminosav neve, pozíciója, és az azt tartalmazó lánc jele.

A következőkben az így elkészült fájl, és a UniProt adatbázisból kiszűrt rendezetlen szekvenciákat tartalmazó fájl alapján ellenőrizem, hogy az aminosavak pozíciója egyezik-e a

két adatbázis szerint. Amennyiben olyan esetet találtam, melyben nem volt egyezés, az adott UniProt és PDB ID-kat elmentettem egy fájlba.

Ezek után a célom az volt, hogy egy olyan fájlt készítsek, mely formailag megegyezik a kötődő aminosavakat tartalmazó fájlommal, de ahol a pozíciók eltérnek a két adatbázisban, ott a PDB helyett a UniProtból származó információ szerepeljen. Ehhez először a már említett „ATOM” sorok alapján kivettem a szükséges szekvenciákat a PDB fájlokból, majd ellenőriztem, hogy ez megtalálható-e a megfelelő UniProt szekvenciában. Amennyiben igen, az azt jelenti, hogy az aminosavak maguk jók, csak a helyzetük el van csúszva a PDB-ben, tehát ezt a későbbiekben egy MUSCLE alignment elkészítésével kijavítottam. Azonban ha nem, akkor az aminosav sorrendben is van eltérés a két adatbázisban.

Az utóbbi esetben első lépésként szintén egy MUSCLE alignmentet végeztem el. Észrevettem, hogy néhány esetben a PDB fájlban hibásan, felcserélve szerepelnek a láncok, tehát a „DBREF” sorokban egy adott lánchoz egy adott UniProt ID tartozik, ám az „ATOM” sorokban már egy másik láncnál szerepel a UniProt ID-hez tartozó szekvencia. Ahhoz, hogy ezeket az eseteket kiszűrjem, a MUSCLE eredményéből csak azokat az eseteket tartottam meg, melyekben az aminosavak legalább 80% egyezik az eredeti és az illesztett szekvenciákban.

Az előző lépésekben kapott adatok alapján végül el tudtam készíteni egy olyan fájlt, mely a már korábban elkészített kötő aminosavakat tartalmazó fájlra hasonlít, annyi különbséggel, hogy ha a PDB és a UniProt adatbázis adatai nem egyeztek meg, akkor a UniProt adatai szerepelnek. Ezt a fájlt a további feladatokhoz ki kellett egészítenem egy olyan oszloppal, melyben az adott aminosav, és a szomszédságában levő 2-2 aminosav rendezetlenségi értékének átlaga is szerepel (8. ábra).

pdb3bum.ent	TYR	B	274	0.06358	ASN	A	53	0.7684
pdb3bum.ent	TYR	B	274	0.06358	GLU	A	54	0.79124
pdb3bum.ent	GLN	B	316	0.13712	GLU	A	54	0.79124
pdb3bum.ent	ILE	B	318	0.1623	THR	A	56	0.82692
pdb3bum.ent	GLN	B	316	0.13712	THR	A	56	0.82692
pdb3bum.ent	TYR	B	307	0.0809	PRO	A	59	0.8742

8. ábra: Néhány sor a fentiek alapján elkészült fájlból.

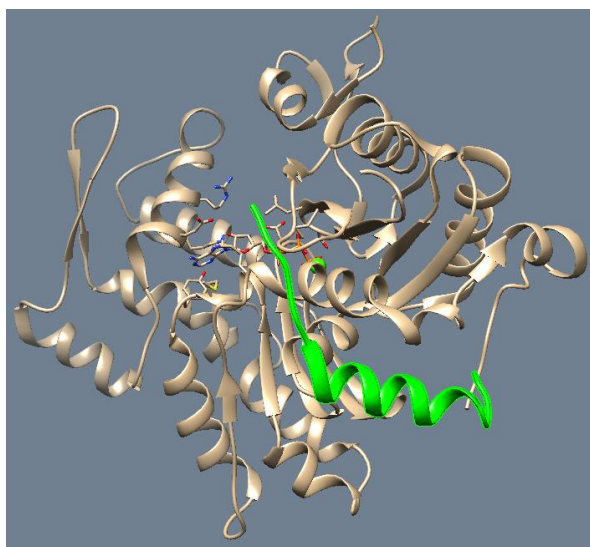
Az így elkészült fájl, és néhány korábbi segítségével egy új adatállományt készítettem, melyben a PDB ID, a lánc jele, és a UniProt ID szerepel, alatta pedig egy (a szekvenciának

megfelelő hosszúságú) bináris string, ahol 1 áll az adott pozícióban, ha az abban a helyzetben levő aminosavhoz legalább 0,6-os (már átlagolt) IUPred érték tartozik, és 0, ha ennél kisebb.

A következőkben egy olyan fájlt készítettem el, melyben a legszélső kötőhelyek közti aminosav szekvencia, valamint a bináris string szerepel, amennyiben abban van 1-es. A két legszélső kötőhelyet a kötő aminosavakat tartalmazó fájl alapján kerestem meg. A továbbiakban kiszűrtem azokat a szekvenciákat, amelyeknek legalább a fele rendezetlen (azaz a string legalább fele 1-esekből áll).

Mivel a kutatás során a kötő motívumokat határoló, úgynevezett flanking régióknak nagy szerepük lesz, ezért a UniProt adatbázis alapján az előzőleg kiszűrt szekvenciákat kiegészítettem (az elején és a végén) 20-20 aminosav hosszú szakaszokkal. Miután ez megtörtént, egy új fájlt készítettem, melyben már az is szerepelt, hogy melyik aminosav kötő aminosav. Ezt szintén egy nullákból és egyesekből álló sorral oldottam meg, ahol az 1 a kötő, a 0 pedig a nemkötő aminosavat jelenti (9. ábra).

```
>pdb3m1f.ent      V   Q87GE5  110..171
PAPPLNLSKGAPKSSSNSLD  SVNEDHSKLMEQIRQGVKLKSA  TKSLSADKSSADAHSKLMEE
00000000000000000000  1110110111111111111111  00000000000000000000
11111111111111111111  11111111111111111000111  11111111111111111110
```



9. ábra: Egy példa a leírtak alapján elkészült fájlból. Az ID sor alatt a szekvencia található, a motívum elejét és végét szókőz jelzi. Az első bináris sor a megfelelő pozícióban lévő aminosav kötésben való részvételét jelzi, míg a második a rendezetlenségre ad információt. Alatta a 3D-s szerkezete a PDB-ből, mely igazolja, hogy valóban egy kötő régióval van dolgunk.

Többször is előfordult, hogy egy láncon belül több különálló motívum is szerepelt (ilyen például a bevezetőben említett kalpain-kapasztatin komplex), ám mivel az előzőekben a két legszélső aminosav közti szakaszt írtam ki, ezek nem különültek el. Azonban a kötő aminosavakat jelző string alapján szét tudtam darabolni a motívumokat. Ez úgy történt, hogy ha legalább 4 nemkötő aminosavat találtam egymás mellett, akkor ott kettévágtam a láncot, és két külön motívumként írtam ki.

A további munkám során elsősorban a redundancia csökkentésére, és a nem kívánatos adatok törlésére törekedtem. Sokszor egy komplexen belül többször is szerepel ugyanaz a szekvencia más láncként, először ezeket az ismétlődéseket távolítottam el. Azonban az is gyakran előfordult, hogy különböző PDB ID-val jelölt komplexekben teljesen egyező motívumok szerepeltek, így ezekből is csak egyet hagytam meg.

Az előző lépések elvégzése után továbbra is maradtak redundáns adatok. Ezek főleg abból adódtak, hogy sok esetben ugyanolyan UniProt ID-vel rendelkező, de nem teljesen egyező motívumok találhatók az adatsztben. Ennek oka, hogy ugyanannak a fehérjének különböző, de átfedő szakaszai szerepelnek motívumként. Az ilyen esetekben kézzel végeztem el a szűrést, és mindig a leghosszabb szekvenciát tartottam meg.

A következőkben a riboszomális fehérjék is szűrésre kerültek, a PDB fájl HEADER sorában szereplő „ribosome” szó alapján. Erre elsősorban azért volt szükség, mert jelenlegi vizsgálatunkban a molekuláris felismerő elemekre vagyunk kíváncsiak, melyek elsősorban tranziens kötődéssel kapcsolódnak a partnerükhöz, a riboszomában található rendezetlen fehérjék azonban a nagy komplex állandó részei, abban állványfehérje szerepét töltik be. Ezután az úgynevezett coiled coil fehérjéket is eltávolítottam az adatsztből. Ezekben a fehérje komplexekben kettő vagy több alfa-hélix tekeredik egymás köré, és gyakran rendezetlennek jóslódik, bár valójában egy igen merev és stabil szerkezeti elem, ezért volt szükség erre a lépésre. Az ilyen fehérjék felismerése a COILS program segítségével történt, mely az adott szekvenciát egy olyan adatbázissal veti össze, melyben ismert coiled-coil fehérjék találhatók, majd kiszámolja annak a valószínűségét, hogy ilyen konformációt vegyen fel a megadott fehérjén.

Így elkészült a végleges adatszt, mely a PDB adatbázisban található rendezetlen kötőszakaszokat tartalmazza.

3.4. Az adatok kinyerése az ELM adatbázisból

Mivel az ELM adatbázisban szereplő lineáris motívumok közül számomra csak a ligandumként szereplő fehérje szakaszok voltak fontosak, ezért az első lépésben a „LIG” kezdetű bejegyzéseket szűrtem ki. Míg az ELM adatbázis alkalmas ismeretlen szekvenciából történő motívum jóslásra is (bár igen magas fals pozitív aránnyal), én csupán a benne szereplő ismert, kísérletes úton bizonyított motívumokat használtam fel PDB-s adataim kiegészítésére. Ezután az adatokat az eddig is használt formátumban tároltam, azaz egy ID sor után a kötőmotívum szerepelt a 20-20 aminosav hosszú határoló régiókkal kiegészítve.

A kinyert adatokon elvégeztem az IUPred rendezetlenség jóslást, majd csak a rendezetlen szekvenciákat tartottam meg, azaz azokat, melyeknél a motívum több mint fele legalább 0,6-os értékel rendelkezett.

A következő lépésben a USEARCH program segítségével clusterezést végeztem 80%-os határral. Ez alapján kiszűrtem a legalább 80%-ban hasonlító szekvenciákat, és mindegyik hasonló fehérje szakaszból csak egyet hagytam meg.

Miután ezzel elkészültem, összefésültem ezeket az adatokat a korábban elkészített PDB-ből származó adatszettel, majd ezen is végeztem egy clusterezést, és eltávolítottam a nagyon hasonló szekvenciákat.

3.5. Rendezetlen kötőszakasz jóslás SVM segítségével

A következőkben az volt a célom, hogy egy olyan módszert találjak, mely segítségével egy véletlenszerűen megadott fehérjeszakasról eldönthető, hogy rendezetlen kötőmotívum-e. Első lépésként megismertem a support vector machine és a Kohonen map módszereket, hogy ezek közül az egyiket fel tudjam használni a feladat megoldása során.

A support vector machine (SVM) egy olyan tanuló algoritmus, melyet adatok klasszifikációjára használhatunk. Olyan hipersíkot hoz létre, amellyel a lehető legjobban el lehet szeparálni a különböző osztályokba tartozó adatokat. Bináris esetben, azaz két csoport elkülönítésekor a tanítás során szükség van egy pozitív és egy negatív adatszetre.

Ezzel szemben a Kohonen map, más nevén self-organizing map (SOM) egy olyan neurális hálózat, mely nem igényel tanító halmazt. A beadott többdimenziós adatokat kétdimenzióssá redukálja úgy, hogy az egyes elemek között távolságot számol, és ez alapján

helyezi el őket a síkban. Ezzel a módszerrel tehát a hasonlóság alapján csoportosíthatunk különböző adatokat.

Úgy döntöttem, hogy bár a Kohonen map a későbbi kutatás során hasznos lesz, de a mostani feladatom elvégzésre a support vector machine az alkalmasabb.

A tanításhoz a pozitív szett a korábban kiszűrt, PDB és ELM adatbázisban található motívumokból állt. Ezen kívül egy negatív szettre is szükségem volt, azaz olyan szekvenciákra, melyek rendezetlennek minősülnek, ám nem kötőmotívumok. Ezeket a UniProt adatbázisból nyertem ki.

Először csak a humán fehérjéket szűrtem ki, majd az ANCHOR program segítségével megjósoltam azt, hogy egy adott aminosavnak mekkora a rendezetlensége, valamint azt, hogy kötő aminosav-e. Miután ez megtörtént, olyan 50 aminosav hosszú szekvenciákat választottam ki, melyekben csak nemkötő, 0,5-nél nagyobb rendezetlenségű aminosavak szerepeltek.

Az így elkészült negatív szettben sok alacsony komplexitású szekvencia volt, ezeket szerettem volna eltávolítani. Ezért komplexitást rendeltem az összes fehérje szakaszhoz a Shannon entrópia képletét használva.

$$K_2 = - \sum_{i=1}^N \frac{n_i}{L} \left(\log_2 \frac{n_i}{L} \right)$$

A képletben szereplő N ebben az esetben egyenlő az aminosavak számával, azaz 20, L az ablak hossza, n_i pedig azt mutatja meg, hogy az i. aminosav hányszor szerepel. Ennek kiszámítása után csak a legalább 3-as komplexitási értékkel rendelkező szekvenciákat tartottam meg.

Ezek után a pozitív és negatív szettben található szekvenciákat egy számsorozattá kellett alakítanom, ennek oka, hogy az SVM csak numerikus információ feldolgozására képes. Azonban nem lett volna elegendő az egyes aminosavaknak egy-egy számot megfeleltetni, mivel minket elsősorban nem a konkrét aminosav sorrend érdekel, hanem a motívumban (és annak környékén) található aminosavak különböző jellemzői, tulajdonságai.

Emiatt a legjobb megoldásnak az aminosav indexek használata bizonyult. Az aminosav indexek (AAindex) az egyes aminosavak fizikai és kémiai tulajdonságait adják meg

numerikus formában. Mivel ezekből az indexekből több mint 500 van, ezért természetesen nem használhattam mindet, hanem ki kellett válogatnom a legfontosabbakat. Ehhez alapul szolgált egy olyan fájl, melyet csoporttársam, Hatos András biztosított számomra, és azt mutatta, hogy mely AAindexek azok, melyek értékei a legjobban különböznek az egyes régiók esetén (távoli régió, flanking régió, motívum).

Az így kiválasztott 6 db aminosav index a következő volt: GOLD730101, SIMZ760101, KARP850101, ZIMJ680105, GRAR740102, és QIAN880114. Ezek sorban a következőkről adnak információt: hidrofobicitás, egy energetikai tulajdonság, flexibilitás, egy másik energetikai tulajdonság, polaritás, béta-redőre való hajlam.

A pozitív és negatív szettbe tartozó szekvenciáimban található aminosavakhoz tartozó konkrét AAindex értékeket szintén a csoporttársam biztosította. Az aminosav indexek értéktartománya az 1-től 5-ig terjedő egész számokból állt. Ez alapján készítettem el a végleges tanító halmazokat. Mivel számunkra elsősorban a motívum, valamint az azt határoló régiók az érdekesek, ezért csak a motívum előtt található határoló régió utolsó 3-3, a motívum első és utolsó 3, valamint a motívum utáni flanking első 3-3 aminosavához tartozó értékeket használtam. Ezeket 3 aminosavanként átlagoltam, majd egymás mögé írtam ki az összes index alapján. Természetesen az így kapott vektorok alatt azt is jeleztem, hogy az adott szekvencia a pozitív vagy a negatív szett része (1 vagy 0).

Az így kapott fájlt 10 csoportra osztottam, ügyelve arra, hogy mindegyik nagyjából ugyanannyi szekvenciát tartalmazzon, és vegyesen legyenek egy csoportban pozitív és negatív szettből származó elemek. Az SVM tanítása során mindig 9 csoportot használtam, 1 csoporttal pedig a validálást végeztem. Ezt addig ismételtem, míg az összes halmaz szerepelt a validálás során. A folyamat során nem csak azt figyeltem, hogy hány szekvenciát sorol helyes osztályba az SVM, hanem azt is, hogy ezekből mennyi a fals pozitív és a fals negatív eredmény.

Természetesen ebben az esetben mindig tudtam, hogy hol van a motívum, de ha egy véletlenszerű szekvenciáról szeretnénk megállapítani, hogy tartalmaz-e rendezetlen kötőszakaszt, és ha igen, az hol található, akkor egy csúszóablakos módszert kell alkalmazni.

A következőkben tehát nem csak egy, hanem több SVM-et tanítottam. Az egyiket a motívum előtti flanking utolsó 3-3 aminosavára vonatkozó AAindexek értékének átlaga alapján, egy másikat pedig ugyanígy, csak a motívum utáni határoló régió szerint tanítottam.

Ezen kívül a motívumokhoz 6 külön SVM-et készítettem, mégpedig a különböző hosszúságúakhoz, 4-től 9 aminosavig. Ezeknél az egész motívumra egy átlagot számítottam az egyes indexek szerint, tehát egy 6 hosszú vektor tartozott az egyes kötőszakaszokhoz.

A folyamat az előzőhöz hasonlóan zajlott, tehát 10 csoportra osztottam az adatszettet, és mindig 9-cel tanítottam, 1-gyel pedig validáltam. Miután a határoló régiókat és a különböző hosszúságú fehérjeszakaszokat felismerő SVM-eket betanítottam az aktuális tanító halmazokkal, a maradék csoporton úgy végeztem el a validálást, hogy egy csúszóablakkal végigmentem az adott szekvencián, és azt figyeltem, hogy az adott részt határoló régiónak jósolja-e. Mind a motívum előtti, mind az azutáni régióknak jósolt szakaszok pozícióját elmentettem. Miután ez megtörtént, végigmentem az összes lehetséges motívumon (az 1. flanking és a 2. flanking közti összes lehetséges szakaszon), és a megfelelő SVM-be beadva figyeltem, hogy melyiket jósolja valóban motívumnak.

3.6. Markov lánc

A következő feladatom a Markov lánc, illetve a rejtett Markov modell megismerése, és a számomra megfelelőbb alkalmazása volt. A Markov lánc egy olyan diszkrét sztochasztikus folyamat, melynek alapelve, hogy egy adott jelenbeli állapot mellett a jövőbeli állapot nem függ a múltbéliektől. Megmutatja, hogy egy adott állapotból mekkora valószínűséggel léphetünk más állapotokba.

A rejtett Markov modell (hidden Markov model, HMM) hasonló elveken alapszik, ám míg a Markov lánc esetében az egyes állapotok jól megfigyelhetők, a HMM-nél maga az állapot nem látható, ám az attól függő kimenet igen. Tehát itt minden állapothoz tartozik egy eloszlás, mely megmutatja, hogy melyik kimenetet milyen valószínűséggel következik belőle.

Mivel én arra voltam kíváncsi, hogy az egyes pozíciókban egy adott aminosav index érték esetén a következő pozícióban milyen értékek milyen valószínűséggel fordulnak elő, ezért elegendő volt a Markov láncot alkalmazni. Ehhez egy olyan programot készítettem, mely a PDB-ből és az ELM-ből szűrt adatszettben szereplő szekvenciák minden pozíciójánál kiszámolja, hogy egy adott aminosav index egyes értékeiből mekkora valószínűséggel lesz egy másik érték a következő pozícióban. A kapott eredmények értelmezése későbbi munkám része lesz.

4. Eredmények

4.1. A PDB és ELM adatbázisból történő szűrés eredményei

A szűrés első lépésében több mint 9000 fehérje komplexet nyertem ki a PDB-ből. Az ezekben a komplexekben előforduló UniProtból származó szekvenciák kigyűjtése után majdnem 7000 fehérjét kaptam eredményül. Miután ezekre a szekvenciákra elvégeztem a rendezetlenség jóslást, és csak a rendezetlennek ítélteteket tartottam meg, körülbelül 2600 fehérjére csökkent az adatszettem, melyek kicsivel több, mint 5000 PDB komplexben szerepeltek.

Amikor a pozíciókat ellenőriztem, kiderült, hogy ebből az 5000-ból majdnem 3000 PDB tartalmaz valamilyen hibás adatot. A MUSCLE illesztés elvégzése után kiesett kb. 100 PDB komplex amiatt, hogy az illesztett és az eredeti szekvencia nem egyezett meg legalább 80%-ban. A hibás adatok kijavítása után nagyjából 6500 olyan láncot kaptam, melyben rendezetlen aminosavak találhatóak. Miután ezekből kiszűrtem a rendezetlennek ítélt szekvenciákat (legalább a fele rendezetlen aminosav), a számuk jelentősen lecsökkent, mindössze valamivel több, mint 2000 lánc maradt.

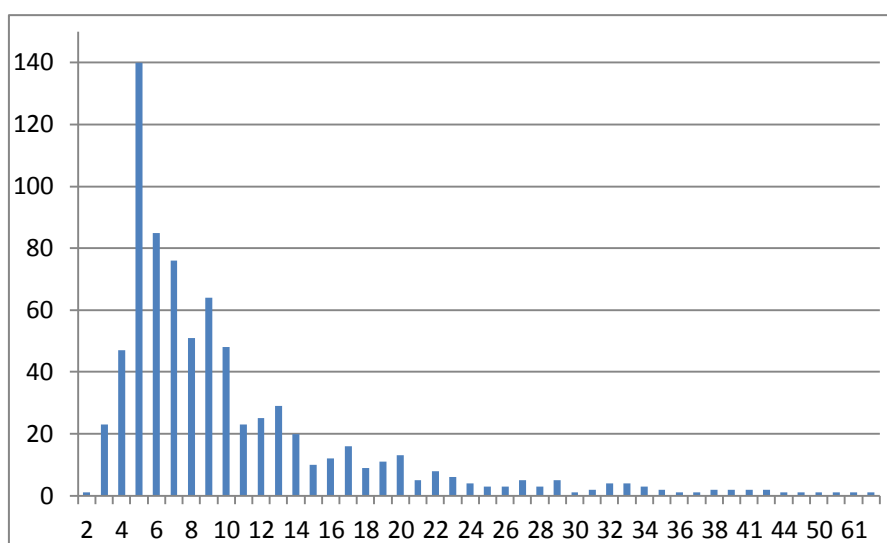
Mivel ez az állomány még jelentős mennyiségű redundáns adatot tartalmazott, kiszűrtem az azonos komplexben többször szereplő láncokat, valamint a teljesen azonos motívumokat tartalmazó különböző PDB-eket, így körülbelül 800 fehérje szakasz maradt. Az eltávolított riboszomális fehérjék száma 50, a coiled coil képzőké pedig 20 körül volt.

Végül a különálló motívumok különválasztása, és az átfedő szekvenciák esetén a leghosszabbak meghagyása után megkaptam a végső fájlt, melyben pontosan 424 rendezetlen kötőmotívum szerepel. Természetesen ez a PDB adatbázis egy korábbi állapotán alapszik, de később feladatom lesz a folyamat lehetőség szerinti teljes automatizálása, így frissülni fog.

A PDB adatbázisból kinyert adataim kiegészítésére az ELM adatbázisból először a „LIG” jelű bejegyzéseket szűrtem ki, ezekből közel 1500 rekord volt. Miután az így kapott motívumokból csak a rendezetleneket tartottam meg, ez a szám 700 alá csökkent. A legalább 80%-ban egyező motívumok szűrése után már csak kb. 500 motívum maradt. Ezeket a PDB-ből származó adatokkal összefésültem, majd az újra elvégzett hasonlósági szűrés után a végleges adatszett 777 rendezetlen kötőmotívumot tartalmaz.

A feladatomban az is része volt, hogy minél kevesebb emberi beavatkozás legyen szükséges az adatok kinyerésénél, hogy egyszerűen lehessen követni a hetente frissülő PDB adatbázisban bekövetkezett változásokat. Ezt részben sikerült csak megvalósítanom, mivel a motívumok szétválasztását, valamint az átfedő szekvenciák szűrését nem tudtam automatizálni. Ez azonban nem okoz különösebb problémát, mivel a heti frissülés során nagyon kevés olyan fehérje kerül a PDB adatbázisba, amely eljut a szűréseken keresztül ezekhez a lépésekhez, tehát kis mértékű manuális feldolgozás lehetségessé válik.

Az így kapott fájl alapján elkészítettem a motívumok méretének eloszlását, és grafikonon ábrázoltam (10. ábra).

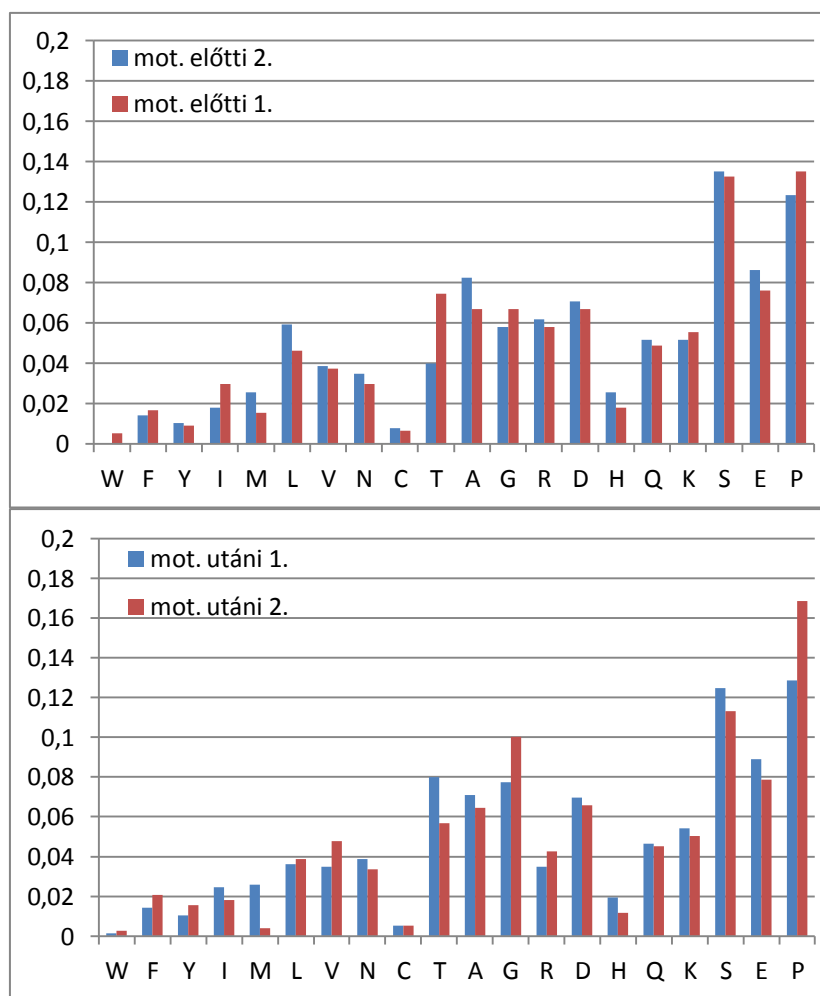


10. ábra: A PDB-ből és ELM-ből származó rendezetlen motívumok méretének eloszlása.

Jól látható, hogy a legtöbb motívum 5 aminosav hosszúságú, és általában a 4-10 aminosav hosszú régiók számítanak gyakorinak, a néhány 10 aminosavval rendelkezők pedig már elenyésző mértékben fordulnak elő.

Ezen kívül az aminosav eloszlást is elkészítettem, mégpedig a motívum előtti flanking utolsó 2 pozíciójára, a motívum első és utolsó 2-2 pozíciójára, valamint a motívum utáni határoló régió első 2 pozíciójára. Bár még pontosan nem tudjuk, hogy a motívumot határoló régió milyen mértékben befolyásolja a motívum szerkezetét, kötődését, egyelőre 2-2 aminosavas régiót elemeztem. Mind a flanking régiókra (11. ábra), mind a kötőrégióra (12. ábra) vonatkozó információkat grafikonon ábrázoltam.

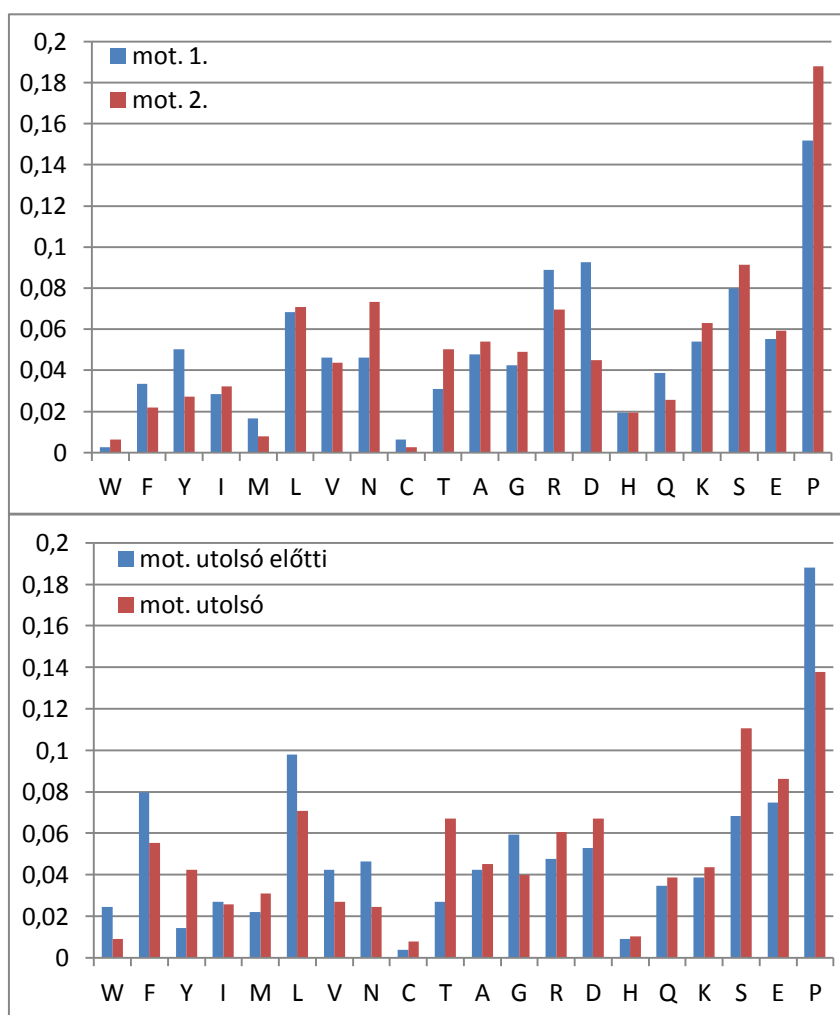
Ha ezeket megfigyeljük, általánosságban elmondható, hogy a rendezetlen motívumokban és közvetlen környezetükben a legjobban jellemző aminosav a prolin (P). Ezt követi a szerin (S), de jelentősebb mennyiségben fordul elő még glutamát (E), arginin (R), aszparaginsav (D), és treonin (T) is. Ezek közül az R, G, S, P és E valóban rendezetlenséget elősegítő aminosavak, a D és T pedig semlegesek ilyen szempontból. Nagyon kis mennyiségben fordul elő triptofán (W), cisztein (C), metionin (M), izoleucin (I) és hisztidin (H). A W, C, I rendezettséget elősegítő aminosavak, a T és H pedig a semleges aminosavak közül került ki.



11. ábra: A motívum előtti flanking utolsó 2 pozíciójának, valamint a motívum utáni flanking első 2 pozíciójának aminosav eloszlása

Elmondható még a grafikonok alapján, hogy az adott helyen megfigyelt két pozíció aminosav eloszlása a flanking régiókban nagyjából megegyezik, nagyjából megfelel a rendezetlen fehérjék átlagos aminosav kompozíciójának, bár a prolin és a poláros aminosavak kis mértékben túlreprezentáltak.

Ha a motívum környezetében levő aminosavak eloszlását összehasonlítjuk a motívumban találhatókéval, akkor is eltéréseket figyelhetünk meg. Szembetűnő, hogy a motívumban megnövekszik a prolin mennyisége, főleg a belsőbb részekben. Ezzel szemben a szerin mennyisége a motívum belsejében lecsökken. Láthatjuk azt is, hogy a glutamin (Q) és az alanin (A) előfordulása kisebb a kötőszakaszban, az aszparaginé (N) pedig a motívum belsőbb részén csökken. A fenilalanin (F), leucin (L) nagyobb mennyiségben van jelen a motívumban, ez kisebb mértékben a tirozinra (Y) is elmondható, az aszparaginsav és arginin pedig csak a motívum elején fordul elő gyakrabban. A hidrofób aminosavak megemelkedett aránya egyértelműen arra utal, hogy a rendezetlen régió ezen szakasza hajlamosabb lehet fehérje-fehérje kapcsolat kialakítására.



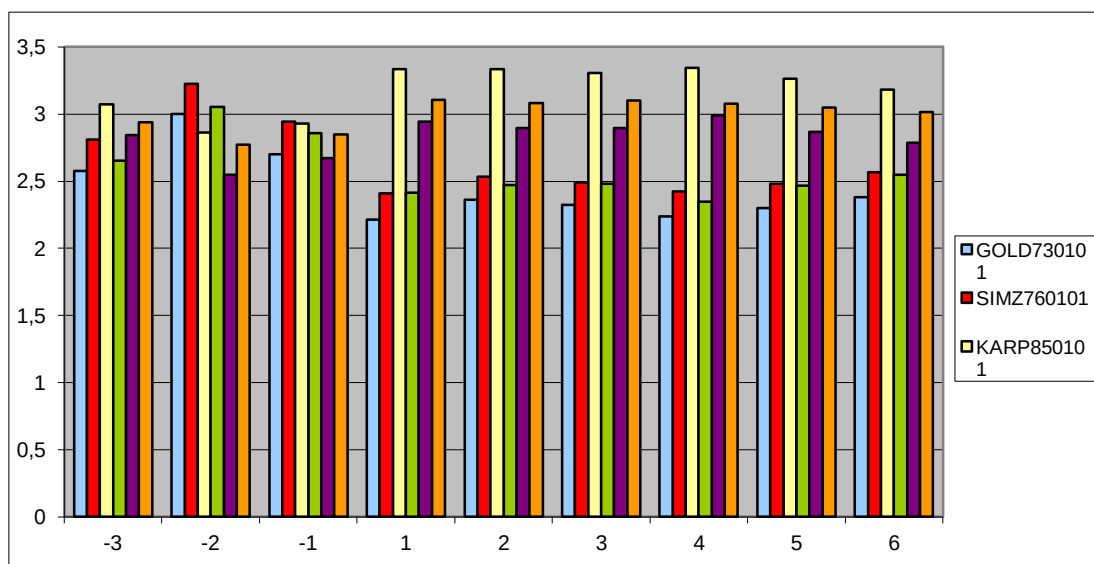
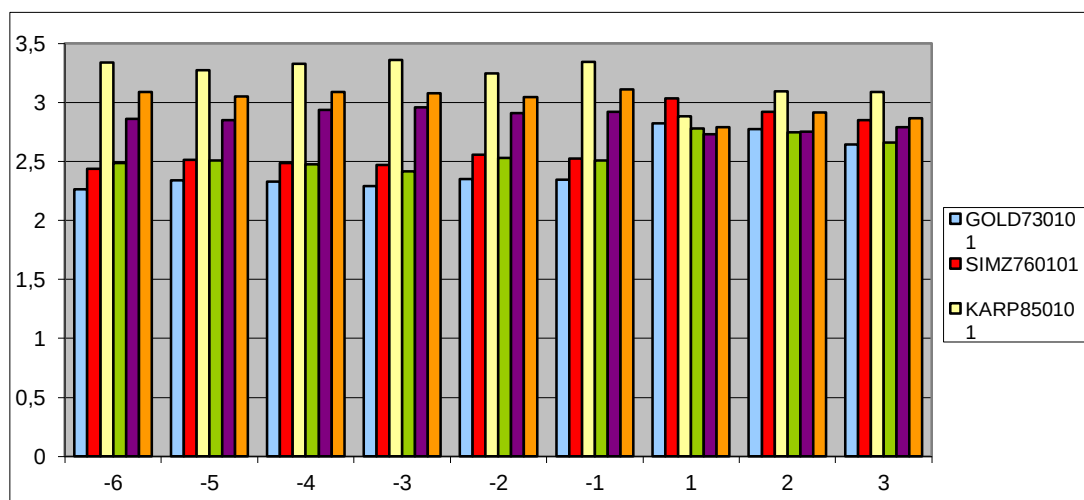
12. ábra: A motívum első és utolsó két pozíciójának aminosav eloszlása

4.2. Az SVM eredményei

Először úgy alkalmaztam az SVM-et, hogy tudtam, hol a motívum helye, és ez alapján végezhettem a validációt. Többféle paramétert is változtattam a feladat során, hogy megtaláljam a legmegfelelőbbeket. Végül a C-SVC használta radiális függvénnyel bizonyult a legjobbnak. Ezzel a módszerrel nagyon jó eredményeket sikerült elérni, átlagosan 80%-os találati aránnyal, valamint alacsony fals pozitív és fals negatív rátával.

Amikor már csúszóablakot használtam, sajnos nem sikerült megfelelő eredményt elérni, mert a motívum határait sokszor rosszul jósolta a program. Emiatt több változtatásra is szükség volt. Egyrészt új aminosav indexeket választottam, másrészt később a már említett Kohonen map alapján csoportosított szekvenciákon külön fogom elvégezni tanítást (a fehérje szakaszok osztályozását Hatos András végezte). Az új módszerrel történő jóslásra a későbbi munkám során kerül majd sor.

A korábban választott aminosav indexekre kiszámoltam, hogy az egyes pozíciókban mennyi az átlagos értékük, majd ezt grafikonon ábrázoltam (13. ábra). Az első ábrán a motívum előtti 6, valamint a motívum első 6 aminosava látható, a második ábrán pedig a motívum utolsó 3, és a motívum utáni 6 aminosav.



13. ábra: Az egyes pozíciókban az aminosav indexek átlagos értéke.

A grafikonokon jól látható, hogy a motívumon belül más érték a jellemző, mint a határoló régióknál. Mivel a narancssárgával jelölt QIAN880114 AAindex esetében nincs túl nagy változás az egyes pozíciókban, ezért úgy döntöttem, hogy egy másik béta-redőre vonatkozó indexet keresek helyette. Ezen kívül hozzávettem még 4 indexet, így a továbbiakban összesen 10 aminosav indexet használlok majd, még hozzá a következőket: GOLD730101, SIMZ760101, KARP850101, ZIMJ680105, GRAR740102, CRAJ730101, FAUJ880111, FAUJ880112, KOEP990102, és ZHOH040103. Az első 5 az eddig használtakkal megegyezik, a többi pedig a következő tulajdonságokkal kapcsolatos: alfa-hélix gyakoriság, pozitív töltés, negatív töltés, béta-redőre való hajlam (más, mint amit eddig használtam), és eltemethetőség.

5. Összefoglalás

Munkám során sikerült elvégezni az elsődleges feladatomat, azaz létrehoztam egy adatszettet, mely a PDB adatbázisban található rendezetlen kötőmotívumokat tartalmazza, az ELM adatbázisban található rendezetlen lineáris motívumokkal kiegészítve. A kötőrégiók és azok környezetének vizsgálatával kiderült, hogy az általam kinyert szekvenciák nagyrészt követik a szakirodalomban korábban leírt tulajdonságokat.

Munkám során végig mindig szem előtt tartottam, hogy bár az adatok kinyerésére, válogatására, összehasonlítására bioinformatikai módszereket használtam, a kiindulási adatok semmiképpen ne számítógépes annotációk eredményéből származzanak, hanem mindig valós kísérletes adatokon alapuljanak.

Sikerült betanítanom egy support vector machine-t úgy, hogy nagyon jó hatékonysággal meg tudja állapítani egy szekvenciáról, hogy rendezetlen motívumot tartalmaz-e. Azonban a motívum konkrét helyének azonosítására már nem működik jól az SVM-eket alkalmazó módszer. A problémát a későbbi munkám során szeretném megoldani, ezt már elő is készítettem az új AAindexek kiválasztásával, valamint a motívumok különböző csoportokba sorolása is segítségemre lesz benne. Ezen kívül a Markov lánc elkészítésével egy másik szemszögből is meg tudom majd közelíteni a feladatot.

Az eredményként létrejött adatszettet nem csak én fogom további munkámban használni, hanem a Rendezetlen Fehérje Kutatócsoport számos más munkatársának is hasznos „alapanyagot” szolgáltatott. Bár az automatizálással kiegészített, rendszeresen frissülő adatbázis tudományos folyóiratban önálló eredményként leközzölhető lenne, mi már az elemzéssel (SVM, Markov lánc, stb.) együtt tervezzük a cikk megírását.

6. Köszönetnyilvánítás

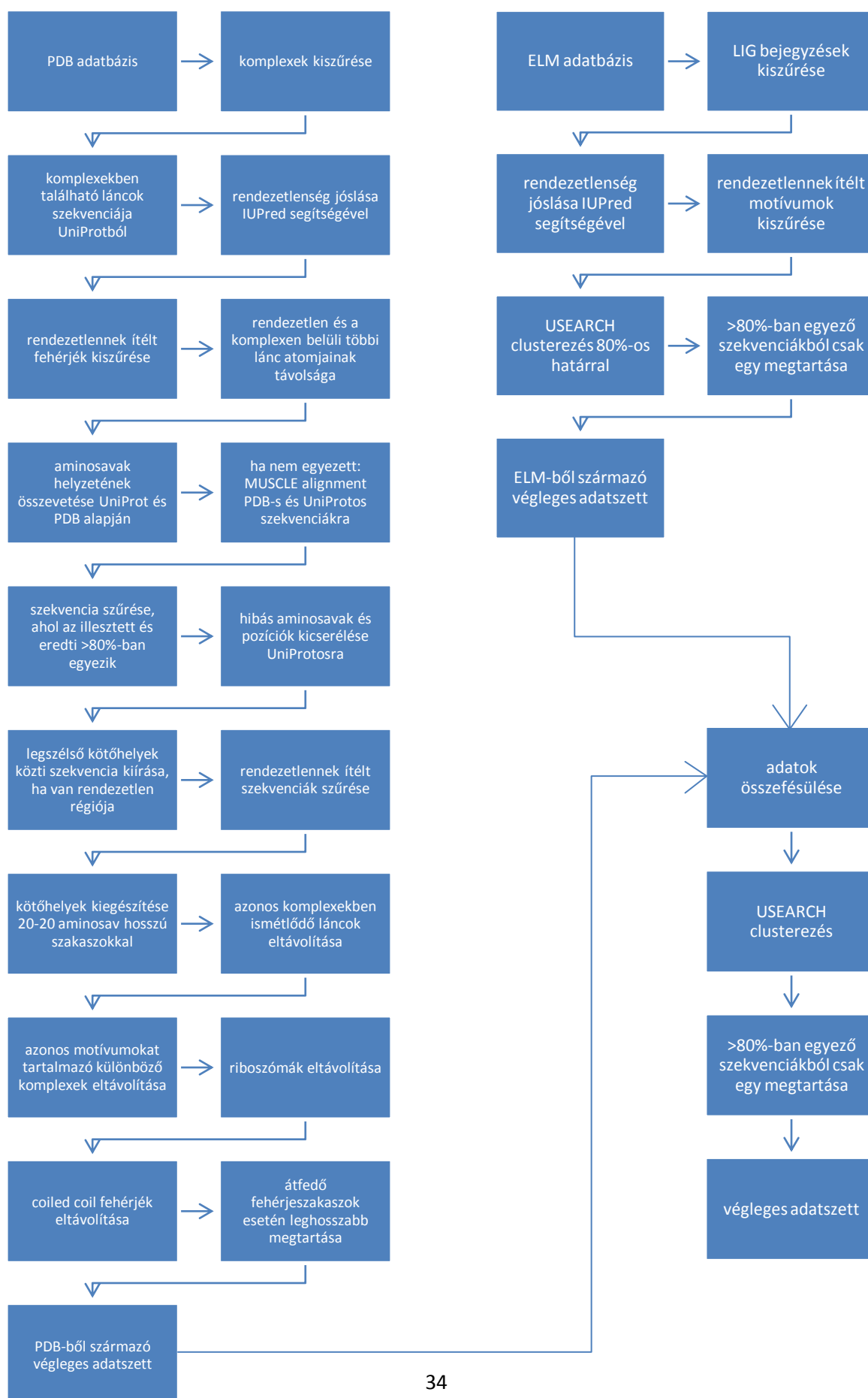
Szeretnék köszönetet mondani témavezetőmnek, Kalmár Lajosnak, aki folyamatos segítséget nyújtott a munkám során. Köszönöm csoporttársamnak, Hatos Andrásnak, hogy biztosította számomra az aminosav indexekkel kapcsolatos adatokat, és segített a megfelelő indexek kiválasztásában. Végül köszönettel tartozom Horváth Tamás PhD hallgatónak a módszerekkel kapcsolatos meglátásaiért.

7. Irodalomjegyzék

- [1] Dosztányi Z, Tompa P: Prediction of Protein Disorder, *Structural Proteomics*, 426:103-115, 2008
- [2] Puntervoll P, Linding R, Gemünd C et al.: ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins, *Nucleic Acids Research*, 31: 3625-3630, 2003
- [3] Gould CM, Diella F et al.: ELM: the status of the 2010 eukaryotic linear motif resource, *Nucleic Acids Research*, 38: D167-D180, 2010
- [4] Tompa P, Simon I, Csizmók V, Dosztányi Z: IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, *Bioinformatics*, 21:3433-3434, 2005
- [5] Dosztányi Z, Mészáros B, Simon I: ANCHOR: web server for predicting protein binding regions in disordered proteins, *Bioinformatics*, 25(20): 2745-2746, 2009
- [6] Fuxreiter M, Tompa P, Simon I: Local structural disorder imparts plasticity on linear motifs, *Bioinformatics*, 23:950-956, 2007
- [7] Davey NE et al.: Attributes of short linear motifs, *Molecular BioSystems*, 8:268–281, 2012
- [8] Dyson HJ, Wright PE: Intrinsically unstructured proteins and their functions, *Nature reviews*, 6:197-208, 2005
- [9] Dosztányi Z, Mészáros B, Simon I: Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins, *Briefings in bioinformatics*, 2:225-243, 2009
- [10] Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z: Intrinsic Disorder and Protein Function, *Biochemistry*, 41:6573-6582, 2002
- [11] Dunker AK et al.: Intrinsically Disordered Protein, *J Mol Graph Model.*, 19:26-59, 2001
- [12] Uversky VN: Natively unfolded proteins: A point where biology waits for physics, *Protein Science*, 11:739-756, 2002

- [13] Abeln S, Frenkel D: Disordered Flanks Prevent Peptide Aggregation, PLoS Computational Biology, 4: e1000241, 2008
- [14] Gunasekaran K, Tsai CJ, Nussinov R: Analysis of Ordered and Disordered Protein Complexes Reveals Structural Features Discriminating Between Stable and Unstable Monomers, J Mol Biol, 341:1327-1341, 2004
- [15] Edgar RC: MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics, 5:113, 2004
- [16] Edgar RC: Search and clustering orders of magnitude faster than BLAST, Bioinformatics, 26:2460-2461, 2010

8. Függelék



Tételes szerzői hozzájárulás

Alulírott Szepes Dóra a dolgozatban leírt munkafolyamatok közül a:

- a rendezetlen kötőmotívumok PDB adatbázisból történő szűrésének minden részfolyamatát,
 - a rendezetlen kötőmotívumok ELM adatbázisból történő szűrésének minden részfolyamatát,
 - a végleges adatszett létrehozását,
 - az SVM tanítását és validálását
- végeztem el.

Budapest, 2014.05.14.

.....

aláírás