SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 227

AUTOMATSKA IZRADA TESTNOG SKUPA ZA PREDVIĐANJE PROTEINSKIH INTERAKCIJA

Juraj Petrović

Zagreb, lipanj 2008.

Zahvaljujem mr. sc. Mili Šikiću na vodstvu i pomoći u izradi ovog rada.

Također zahvaljujem kolegi Ivoru Prebegu na pomoći pri uspostavljanju baza podataka.

Sadržaj:

1. Uvod

Izrada neredundantnoga testnog skupa proteina postupak je koji prethodi brojnim istraživanjima i radovima s područja bioinformatike. Opisa izrade takvih skupova različitih svojstava već postoji nekoliko, a glavnu primjenu u praksi ti skupovi nalaze u predviđanju proteinskih interakcija, aktivnih mjesta na površini proteina i njihovoj strukturi. Budući da je riječ o eksperimentalnim podacima manje ili veće točnosti, u ovom su završnom zadatku postavljeni uvjeti ne samo glede raznolikosti i svojstava samih proteina u skupu, nego i kvalitete i preciznosti podataka.

Ovim radom htio sam objasniti postupak dobivanja takvoga neredundantnog skupa proteina za predviđanje proteinskih interakcija, ali ne samo to nego i pojasniti pojmove, materijale i metode vezane uz isti proces.

U drugom sam poglavlju sam objasnio što su proteini, kako i gdje nastaju, od čega su građeni, čemu služe i, budući da je riječ o makromolekulama čija se veličina obično izražava u nanometrima, kako se do tih podataka o njima uopće dolazi. Izrada zadanoga neredundantnoga testnog skupa proteina bez poznavanja ove materije nije moguća.

Treće poglavlje analizira tipove i problematiku proteinskih reakcija i uvjete koji na njih utječu. Ovo je područje vrlo kompleksno, a zbog mogućnosti da se njime objasni kako nastaju neke bolesti interes je za njega u porastu.

U četvrtom poglavlju nabrojao sam i objasnio podatke dostupne za izradu samog neredundantnog skupa. Najbitnija je dakako PDB baza podataka koja sadrži oko 50 000 eksperimentalno dobivenih opisa proteina, ali tu su i još neki podaci čijim sam kombiniranjem došao do konačnog rješenja.

Peto poglavlje opisuje metode koje sam koristio nad dostupnim podacima. To se prvenstveno odnosi na parsiranje u programskom jeziku Python i SQL upite. Svi postupci su, radi lakšeg razumijevanja, dokumentirani i grafičkim prikazima.

Rezultati rada izloženi su u šestom, a zaključak u sedmom poglavlju ovog rada.

2. Proteini

2.1 Aminokiseline

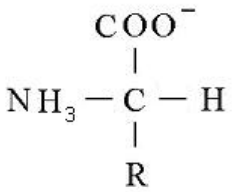
2.1.1 Uvod

Aminokiseline su molekule čijim spajanjem nastaju kratki polimerski lanci peptidi ili dugi lanci polipeptidi odnosno proteini. Dogovorena razlika je da se spojevi koji broje manje od pedeset aminokiselina u nizu nazivaju polipeptidi, a da se makromolekule sa više od pedeset aminokiselina nazivaju proteini. U prirodi postoji dvadeset različitih aminokiselina poznatih kao standardne, proteinogenske ili alfa-aminokiseline i to su: alanin, arginin, asparagin, aspartinska kiselina, cistein, glutaminska kiselina, glutamin, glicin, histidin, izoleucin, leucin, lizin, metionin, fenilanin, prolin, serin, treonin, triptofan, tirozin i valin. Aminokiseline, dakle, možemo promatrati kao građevne jedinice proteina i u prirodi ih vrlo rijetko nalazimo u slobodnom stanju [1].

Osim standardnih postoje i nestandardne aminokiseline. Neke od njih stvaraju biljke i mikroorganizmi i to najčešće promjenom pojedinog aminokiselinskog ostatka u proteinu (eng. *residue*) nakon što je on sintetiziran [8].

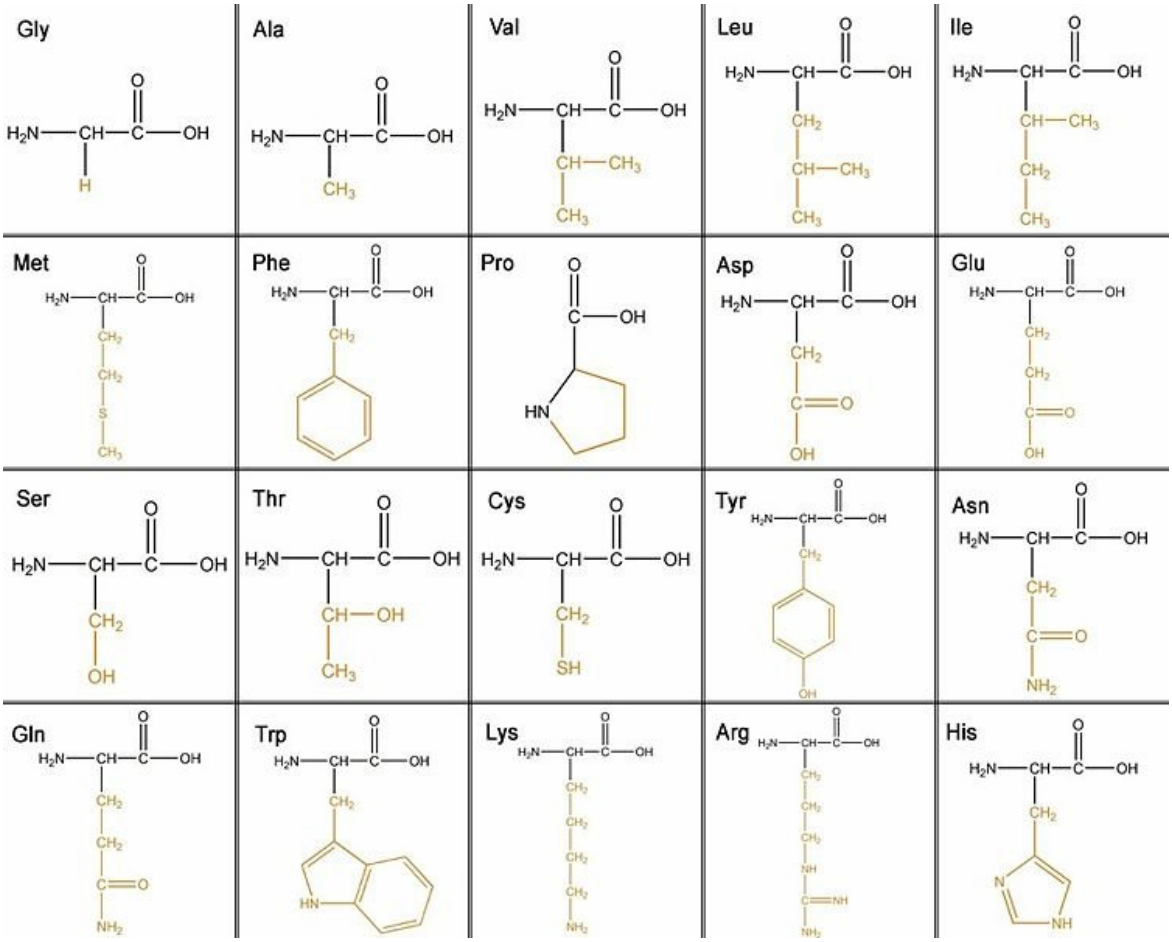
2.1.2 Standardne aminokiseline

Aminokiseline tipično sadrže amino skupinu (- NH_2) ili karboksilnu skupinu (-*COOH*), te bočni lanac koji se najčešće označava s *R*. Općenita struktura aminokiselina prikazana je na slici 2.1, dok su strukture svih dvadeset standardnih aminokiselina prikazane na slici 2.2.



Slika 2.1 Općenita struktura aminokiseline

Svih 20 standardnih aminokiselina ima istu osnovnu strukturu u kojoj se nalaze sva tri navedena elementa (amino skupina, karboksilna skupina i bočni lanac) vezana na isti ugljikov atom, takozvani alfa ugljikov atom. Malu razliku nalazimo samo kod prolina. Kod te aminokiseline dušikov atom iz amino skupine na sebe ne veže tri atoma vodika nego samo dva, a treću kovalentnu vezu uspostavlja s drugim atomom ugljika [1].



Slika 2.2 Strukture 20 standardnih aminokiselina

Bočni ogranak *R* vezan za alfa ugljik odgovoran je za karakteristične osobine pojedinih aminokiselina. On može biti primjerice atom vodika, alkilna skupina, aromatski prsten ili heterociklički prsten, i predstavlja osnovu za podjelu aminokiselina [8]:

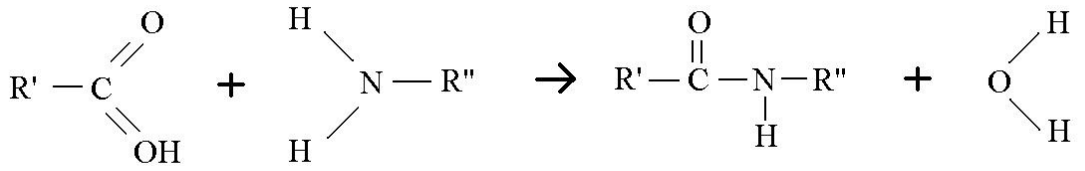
- aminokiseline nepolarnoga bočnog ogranka: alanin, valin, leucin, izoleucin, prolin, fenilalanin, triptofan i metionin
- aminokiseline polarnoga bočnog ogranka: glicin, serin, treonin, cistein, tirozin, asparagin i glutamin

- aminokiseline sa kiselim bočnim ogrankom: asparaginska kiselina i glutaminska kiselina
- aminokiseline sa bazičnim bočnim ogrankom: lizin, arginin, histidin

Aminokiseline se također dijele i na *lijeve* i *desne*, prema položaju amino skupine u odnosu na ostatak lanca. U proteinima isto tako možemo razlikovati esencijalne i neesencijalne aminokiseline, to jest one koje se ne mogu samostalno obnavljati u organizmu pa ih moramo unositi putem hrane, i one koje organizam može sintetizirati. Arginin i histinin spadaju u poluesencijalne aminokiseline koje ne mogu stvarati djeca.

2.1.3 Vezivanje aminokiselina

Zahvaljujući amino i karboksilnoj skupini aminokiseline mogu reagirati međusobno spajajući se u lance. Ovom reakcijom dehidracije nastaje peptidna veza između amino skupine jedne i karboksilne skupine druge amino kiseline i otpušta se molekula vode [1]. Rezultat vezanja aminokiselina naziva se amid, a odgovarajuća skupina *-C(=O)NH-* amino ili peptidna skupina.



Slika 2.3 Reakcija stvaranja amina i peptidne veze

Ova je reakcija reverzibilna. Peptidne veze u proteinima, koje se definiraju kao metastabilne, mogu se prekinuti hidrolizom odnosno topljenjem u vodi, ali ta je reakcija jako spora i u živim organizmima obično se odvija uz posredovanje enzima.

Zbog prisutnosti i amino i karboksilne skupine aminokiseline se mogu smatrati i kiselinama i bazama. Pri određenoj pH vrijednosti, karakterističnoj za svaku aminokiselinu i poznatoj pod nazivom izoelektrična točka jedna je skupina pozitivno, a druga negativno nabijena i takav se ion naziva zwitter ion (njem. *zwitter* = hibrid, hermafrodit). Ovakvi spojevi nazivaju se amfotermnim spojevima.

2.2 Proteini

2.2.1 Uvod

Riječ *protein* dolazi od grčke riječi *πρώτα* ("prota") koja znači "od primarne važnosti". Prvi ih je opisao švedski kemičar Jöns Jakob Berzelius 1838. godine, ali njihova prava važnost nije uočena do sredine prve polovice 20. stoljeća.

Proteini su makromolekule sačinjene od pedeset ili više aminokiselina međusobno povezanih peptidnom vezom. U ljudskom tijelu, u kojem ih je otkriveno više od 80 000 različitih, imaju brojne funkcije. Za bilo koji dio tijela koji prolazi proces rasta stvaraju se nove tjelesne stanice, koje trebaju proteine za izgradnju i uspostavu svoje funkcije. Kolagen je, primjerice, najčešći protein u ljudskom tijelu i daje snagu vezivnom tkivu mišića, tetiva i kože. Elastin čini kožu rastezljivom, a keratin gradi kosu i nokte. Proteini grade i veliki dio molekule hemoglobina, koja prenosi kisik našim tijelom i omogućuje odvijanje procesa disanja u svim stanicama u kojima je to potrebno. Oni također imaju i strukturne funkcije u održavanju oblika stanice ili mehaničke funkcije u mišićima kao, primjerice, aktin i miozin ili prenose materijale unutar stanice, izgrađuju antitijela za obranu tijela od stranih tvari, bakterija i virusa.

Proteini isto tako sudjeluju u staničnim procesima kao enzimi koji kataliziraju biokemijske reakcije. Bez enzima bi mnoge kemijske reakcije tekle presporo da bi se organizam održao na životu. Osim toga enzimi pomažu stanici pri dobivanju energije iz šećera, pomažu u probavi i imaju još brojne druge funkcije. Neki enzimi iz sline razbijaju velike polisaharidne molekule u manje molekule šećera, što je razlog da nam se čini da kruh postaje slađi, što ga dulje žvačemo.

Broj mogućih različitih proteina u biosferi praktički je neograničen. Uzevši u obzir samo 20 različitih standardnih aminokiselina, može postojati $20^{100} \approx 10^{130}$ različitih proteina dugih 100 aminokiselina (100-mera).

2.2.2 Sinteza proteina

Može se reći da proteini upravljaju svim životnim procesima stanice, ali se ne mogu samoumnažati. Proteini se u stanicama organizma sintetiziraju u ribosomima pomoću informacija zapisanih u genima, pojedinim dijelovima DNA [9]. DNA je nukleinska kiselina u obliku dvostruke spiralne zavojnice čiji lanci su građeni kombiniranjem četiriju mogućih baza: adenina (A), citozina (C), guanina (G) i timina (T). Ona nije jedinstvena molekula nego nekoliko njih povezanih vodikovim vezama [10].

U procesu sinteze proteina pojedini dijelovi DNA najprije se prepisuju u glasničku RNA (mRNA) pomoću proteina koji se naziva RNA polimeraza. Ovaj dio sinteze naziva se transkripcija. Osim RNA polimeraze u transkripciji sudjeluje još i velik broj različitih proteina da bi se osiguralo točno prepisivanje informacije. U nekih organizama mRNA se dodatno obrađuje, da bi napokon postala uzorak na temelju kojega se odvija sinteza u ribosomu. Ovaj dio procesa naziva se translacija. Prilikom translacije mRNA se u ribosomu čita po tri baze odjednom (triplet ili kodon). Od 64 različita kodona (4 različite baze, 3 mjesta) njih 61 kodiraju odgovarajuće aminokiseline (više kodona može kodirati istu aminokiselinu), a kodoni UAA, UAG i UGA služe kao stop-signali koji znače da sintezu treba privesti kraju [9]. Iako je ovo uvriježeno mišljenje, neka novija istraživanja pokazala su da se u nekim slučajevima kodon UAG ne poštuje kao stop-signal, nego kao kod za nestandardnu aminokiselinu pirolizin [11]. Odgovarajuće aminokiseline za sintezu dobavljaju se pomoću glasničke RNA ili tRNA [9].

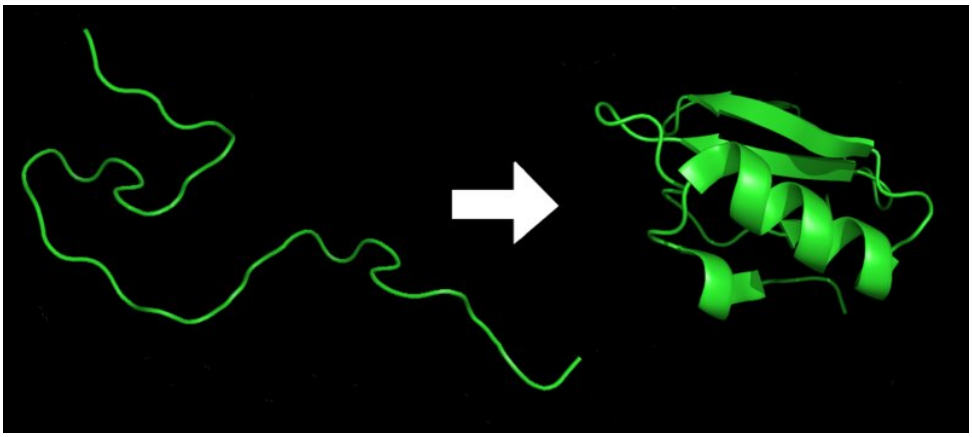
Za vrijeme i nakon translacije često dolazi do svijanja proteina (eng. *protein folding*) kako bi poprimio svoju prirodnu sekundarnu i tercijarnu strukturu. Mnogi proteini podliježu i post-translacijskim promjenama koje mogu uključivati stvaranje disulfidnih mostova ili dodavanje nekih funkcijskih skupina kao što su acetati ili fosfati.

Veličina sintetiziranih proteina može se mjeriti brojem sadržanih aminokiselina ili molekularnom masom koja se obično izražava u jedinicama Daltonima (Da). Ova jedinica sinonim je za unificiranu atomsku jedinicu mase i iznosi $1.660538782 \cdot 10^{-27}$ kg. Najveći poznati proteini su titini duljine gotovo 27 000 aminokiselina i težine oko 3000 kDa.

2.2.3 Struktura proteina

U proteinu niz povezanih aminokiselina čini jedan lanac. Lance možemo dakle smatrati građevnim jedinicama proteina. U ovisnosti o broju lanaca od kojih su građeni, proteini se dijele na monomere i polimere, dakle one građene od samo jednog ili one od više lanaca. Polimeri se mogu dalje dijeliti s obzirom na točan broj lanaca od kojeg su građeni. Sami lanci koji grade neki protein mogu biti slični ili različiti i na temelju toga proteini se dijele na homologne (najčešće s nekim postotkom sličnosti) i heterologne. Za homologne proteine smatra se da su se razvili iz nekoga zajedničkog pretka, a do razlika je došlo djelovanjem sila evolucije. Prema obliku proteine dijelimo na globularne i fibrilarne (vlaknaste).

Svaki protein prilikom nastanka najprije postoji u obliku samo dvočlanog lanca aminokiselina [1]. Elementi ovog lanca nakon njegova formiranja imaju neka svojstva poput hidrofobnosti ili električne nabijenosti. U reakciji sa samim sobom i staničnim okruženjem protein će se tada saviti i poprimiti svoju prirodnu prostornu strukturu. Ovaj mehanizam savijanja i oblikovanja nije još do kraja razjašnjen a zbog njegove važnosti za staničnu fiziologiju i genetiku danas predstavlja jedno od najuzbudljivijih područja biologije.



Slika 2.4 Savijanje proteina u 3-D strukturu

Disulfidne i vodikove veze još su jedan od mogućih pokretača savijanja proteina. Disulfidne veze uspostavljaju se između tiolnih skupina (-*SH*) aminokiseline cisteina. Druga aminokiselina koja sadrži atome sumpora, metionin, ne može stvarati ovakve veze

[12]. Vodikove veze su najjače slabe međuatomske veze, ali slabije od ionskih ili kovalentnih veza. Ostvaruju se asimetričnim kovalentnim vezivanjem atoma vodika na atome izraženije elektronegativnosti, poput fluora, dušika ili kisika. Pozitivno električno polje tako vezanog vodika privlači vanjske elektrone atoma drugih molekula i stvara dodatnu vezu između dijelova makromolekule. Prekidanje samo jedne vodikove veze može izazvati promjenu u strukturi makromolekule. Vodikovim vezama između odgovarajućih parova baza izgrađen je i oblik molekule DNA [13].

Kada govorimo o strukturi proteina, uzimajući u obzir dosad rečeno, hijerarhijski razlikujemo primarnu, sekundarnu, tercijarnu i kvarternarnu strukturu.

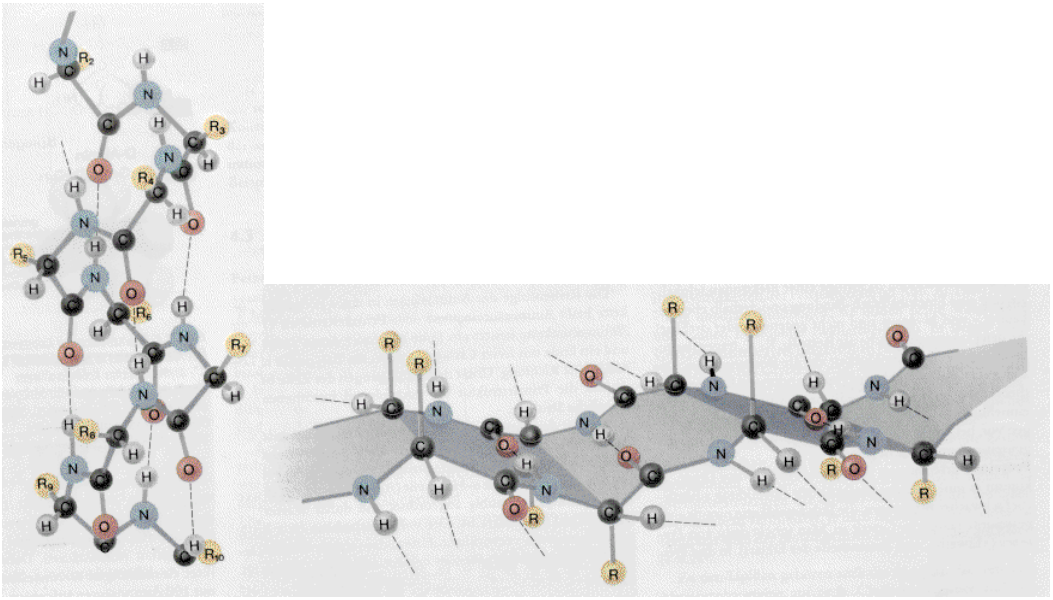
Poput redoslijeda slova u riječi, primarna struktura podrazumijeva raspored ili redoslijed vezanja aminokiselina u polipeptidni lanac [14]. Slika 2.4 prikazuje primarnu strukturu inzulina, proteina za reguliranje razine šećera u krvi. Kao što je moguće primijetiti, protein inzulin se sastoji od dva lanca (dimer), a ujedno je i prvi protein čija je primarna struktura određena [15].

Lanac 1: GLY- ILE -VAL- GLU -GLN -CYS -CYS -THR- SER -ILE -CYS-
SER -LEU - TYR -GLN -LEU -GLU -ASN -TYR -CYS -ASN

Lanac 2: PHE -VAL -ASN-GLN -HIS -LEU -CYS- GLY- ASP -HIS -LEU-
VAL- GLU- ALA -LEU- TYR -LEU- VAL- CYS- GLY- GLU- ARG-
GLY- PHE -PHE -TYR - THR -PRO -LYS -THR

Slika 2.5 Primarna struktura: inzulin [15]

Sekundarna struktura odnosi se na nekoliko uobičajenih, lokalnih struktura u proteinu koje tvore određeni i uvijek isti slijedovi aminokiselina i govori o uzorcima smatanja polipeptidnog lanca [14]. Postoje dva tipa sekundarne strukture koji su posebno stabilni i koji se pojavljuju u vrlo velikom broju proteina, a to su α-uzvojnice i β-ploče. Uz α-uzvojnice postoje i kolagenske uzvojnice, koje se za razliku od prethodnih sastoje ne od jednog, nego od tri međusobno isprepletena polipeptidna lanca. Ove uzorke nalazimo i u globularnim i u fibrilarnim proteinima, iako su znatno češći u fibrilarnim [1]. Nalaze se najčešće blizu jezgre proteina [14]. Slika 2.5 prikazuje izgled α -uzvojnice i β -ploče.



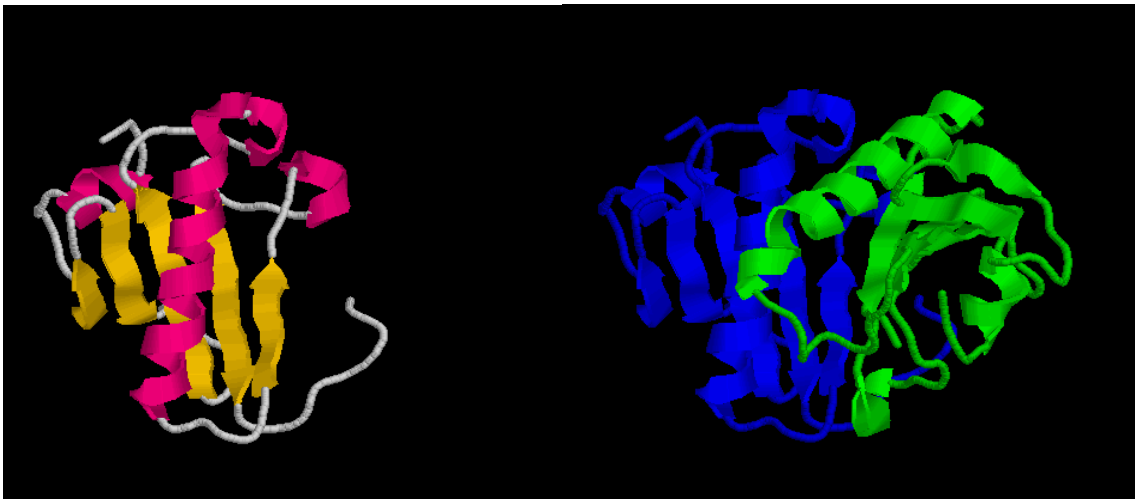
Slika 2.6 Sekundarna struktura: α-uzvojnica (lijevo) i β-ploča (desno) [23]

Tercijarna struktura ili konformacija obuhvaća cjelokupan prostorni raspored atoma u proteinu nastao njegovim savijanjem [14]. Aminokiseline koje su u lancu jako udaljene mogu u tercijarnoj strukturi biti vrlo blizu, jer ona pokazuje prave prostorne odnose među elementima sekundarne strukture. Kada bi, primjerice, vrpcu spiralizirali stružući je oštrim predmetom, dobili bi predodžbu sekundarne strukture proteina. Kada bi istu vrpcu potom omotali oko nekog predmeta, dobili bi tercijarnu strukturu [1].

Kvarternarna struktura proteina postoji samo ukoliko se protein sastoji od više odvojenih polipeptidnih lanaca i njihov raspored u prostoru nazivamo kvarternarnom strukturom. U prethodnom primjeru to bi značilo da za omatanje predmeta ne koristimo samo jednu, nego više vrpci, istih ili različitih [1]. Tercijarna i kvartarna struktura proteina mogu bitno pojasniti na koji način protein obavlja svoju funkciju. Slika 2.6 predstavlja primjer dvaju kvartarnarnih struktura proteina. Za lijevi protein na slici kvartarna struktura identična je tercijarnoj budući da je građen od samo jednog lanca.

Za brojne proteine prirodna trodimenzionalna struktura nužna je da bi pravilno funkcionirali, tako da neuspjeh pri savijanju rezultira neaktivnim ili čak potencijalno opasnim proteinima. Biokemičari su otkrili da ponekad i najmanja izmjena polipeptidnog lanca stvara protein bitno drugačijih svojstva i biološke funkcije, iako nije uvijek nužno tako. Primjerice, proteini iz mišića krave i čovjeka iznimno su slični i unatoč nekim

razlikama obavljaju istu funkciju i imaju istu kvartarnu strukturu. Ali kada se u molekuli hemoglobina hidrofilna aminokiselina valin zamijeni hidrofobnim glutamatom (soli glutaminske aminokiseline), cijeli hemoglobin gubi funkciju i počinje lako kristalizirati. Ovaj poremećaj naziva se bolest srpastih ćelija. Zbog hemoglobina crvene krvne stanice poprimaju srpasti oblik zbog kojega se više ne mogu lako kretati kroz krvne žile nego zapinju stvarajući nakupine. Budući da je bolest genetski nasljedna, potencijalni lijek traži se na području genske terapije. Osobe koje su od roditelja nasljedile samo jedan gen za sintezu lošeg hemoglobina neće oboljeti od ove bolesti, ali će genetsku predispoziciju s vjerojatnošću od 50% prenijeti na djecu.



Slika 2.7 Kvartarne struktura proteina: monomer (lijevo) i dimer (desno) [23]

2.2.4 Prioni

Bolest srpastih ćelija uzrokuje, dakle, "pokvareni" protein. Kada je 1982. Stanley Prusiner skovao riječ *prion* od riječi *proteinacious* i *infectious*, kolege su ga ismijali što smišlja novu riječ za nešto čije postojanje uopće nije dokazano. Njegova teza, dugo neprihvaćena, bila je kako bolest kravljeg ludila koju je proučavao ne uzrokuje ni virus ni bakterija nego upravo pogrešno formirani protein - prion [1].

Za promatrani prion nazvan PrP uskoro je ustanovljeno kako se može saviti u dva različita oblika: normalan oblik, pronađen i u mozgovima zdravih životinja, čija je struktura

sadržavala četiri alfa-uzvojnice, te drugi oblik, dobiven iz mozgova zaraženih životinja koji je umjesto dvije α-uzvojnice sadržavao po dva para paralelnih β-ploča. Štoviše, primjećeno je kako iste te β-ploče mogu u zdravom proteinu izmijeniti α-zavojnice u isti model [1].

Prusiner je 1997. godine za svoja otkrića dobio Nobelovu nagradu iako su mnogi znanstvenici taj čin smatrali preuranjenim, a neki se još uvijek nisu slagali s njegovim rezultatima i smatrali kako tu bolest ipak uzrokuje tip virusa. Danas je poznato da se kao posljedica priona mogu javiti mnoge neurodegenerativne bolesti, između ostalog i Parkinsonova, Huntingtonova i Alzheimerova bolest. Ideja dizajna lijekova za ovakve bolesti je pronalaženje molekula ili drugih proteina koji će onesposobiti ili promijeniti prione pri reakciji s njima, a to zahtijeva detaljno poznavanje strukture i aktivnih mjesta na površini priona i proteina.

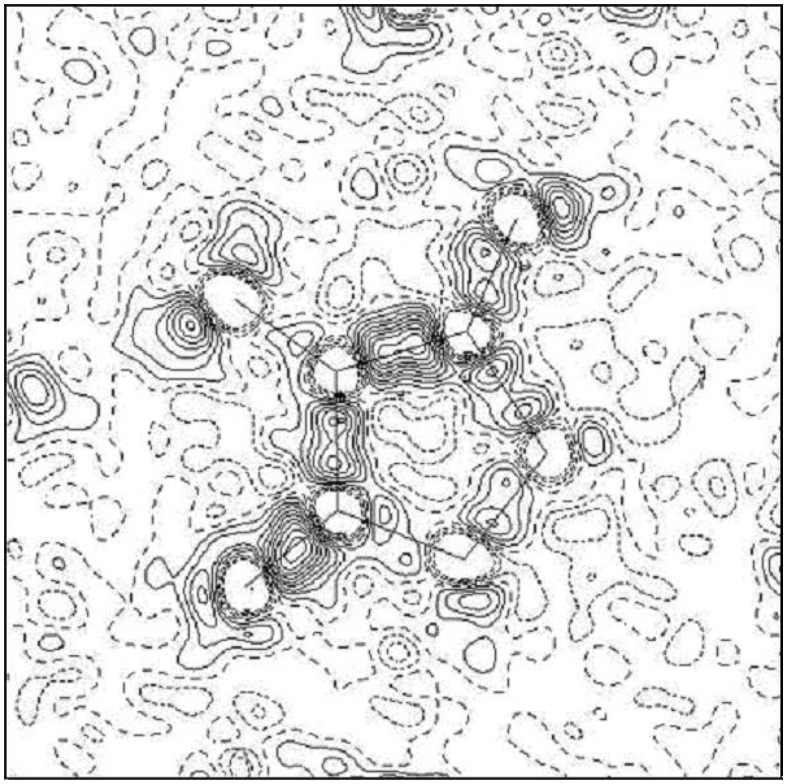
2.2.5 Određivanje strukture proteina

Trenutno najpopularnija metoda za određivanje struktura makromolekula je rendgenska kristalografija (eng. *X-ray cristallography*) i to je zasada najmoćnija i najtočnija metoda za određivanje kristalnih struktura, toliko razvijena, da se strukture veličine 100 do 200 atoma mogu odrediti za svega jedan do dva dana [17].

Da bi analizirali strukturu proteina ili makromolekule treba je najprije zamrznuti i time homogenizirati. Veličine uzoraka obično imaju stranice duljine 0.1 do 0.3 mm. Potom zamrznuti uzorak bombardiramo X zrakama. X zrake su elektromagnetski valovi valne duljine od 10 do 0.01 nm [16]. Običnu svjetlost, to jest elektromagnetske valove valne duljine 380 do 780 nm, ne bi imalo smisla koristiti budući da za prepoznavanje osvijetljenog objekta njegova visina mora biti približno jednaka bar polovici valne duljine svjetlosti koja ga obasjava. Nažalost, X zrake za razliku od vidljive svjetlosti nije moguće usmjeriti lećom, a to je i razlog zašto promatrani uzorak mora biti zamrznut. X zrake odbijaju se od elektrona te dolazi do promjene u njihovoj amplitudi i fazi, a njihovo raspršivanje bilježi se kao otisak. Amplituda zraka koje su se raspršile može se lako dobiti iz intenziteta, kojeg opet dobivamo iz broja fotona u točci otiska. Problem koji ostaje jest faza, koja se može izračunati, ali za njeno određivanje često se koriste i neke heurističke metode. Fouriereovom transformacijom iz dobivenih informacija konstruiraju se grafički

prikazi ili karte gustoće elektrona (eng. *electron density map*), na temelju kojih se pretpostavljaju položaji atoma u uzorku. Ovaj postupak računanja provodi se za više kutova kako bi se na kraju dobila prostorna struktura [18].

Glavni nedostaci ove metode određivanja strukture makromolekula su ograničena rezolucija i problem s prepoznavanjem atoma vodika, jer oko njegove jezgre kruži samo jedan elektron koji se često ne vidi na grafičkom prikazu gustoće elektrona [18].



Slika 2.8 Karta gustoće elektrona [24]

Spektroskopija nuklearnom magnetskom rezonancijom (eng. *NMR spectroscopy*) druga je najčešća metoda za određivanje makromolekulskih struktura. Primjenjiva je na svaku jezgru atoma koja posjeduje posebnu vrstu kutnog momenta - spin. Spin posjeduju sve jezgre s neparnim brojem protona ili neutrona [19].

Kada se ovakve jezgre nalaze u magnetskom polju jakosti specifične za određeni tip jezgre ona apsorbira fotone elektromagnetskog zračenja odnosno rezonira. Budući da je i rezonantna frekvencija zračenja karakteristična za pojedini tip jezgre, njezinom promjenom može se utvrditi prisutnost i raspored pojedinih jezgri [19].

Među ostalim metodama određivanja makromolekularnih struktura su krioelektronska mikroskopija (eng. *cryoelectron microscopy*, *cryo-EM*), koja se obično koristi za dobivanje informacija niže rezolucije o strukturi vrlo velikih proteinskih kompleksa, te elektronska kristalografija (eng. *electron cristallography*), koja može dati informacije visoke rezolucije, pogotovo za dvodimenzionalne kristale membranskih proteina i to znatno uspješnije od rendgenske kristalografije. Postoje i metode predikcije strukture proteina na temelju primarne strukture i poznatih svojstava. Uspješnost tih metoda provjerava se na takozvanom CASP eksperimentu (*Critical Assessment of Techniques for Protein Structure Prediction*), koji se od 1994. godine održava svake druge godine [21].

3. Proteinske interakcije

3.1 Podjela proteinskih interakcija

Proteinske interakcije su vrlo složeni procesi koji ovise o velikom broju parametara. Razumijevanje principa po kojima se one odvijaju ključno je za razumijevanje veze između njihove biološke funkcije i molekularne strukture. Proteinske interakcije možemo razlikovati ne temelju proteina koji u njima sudjeluju ili na temelju njihovih fizikalnih svojstava. Na temelju fizikalnih svojstava proteina, proteinske interakcije ili komplekse koji nastaju njima možemo podijeliti u tri podskupine [2].

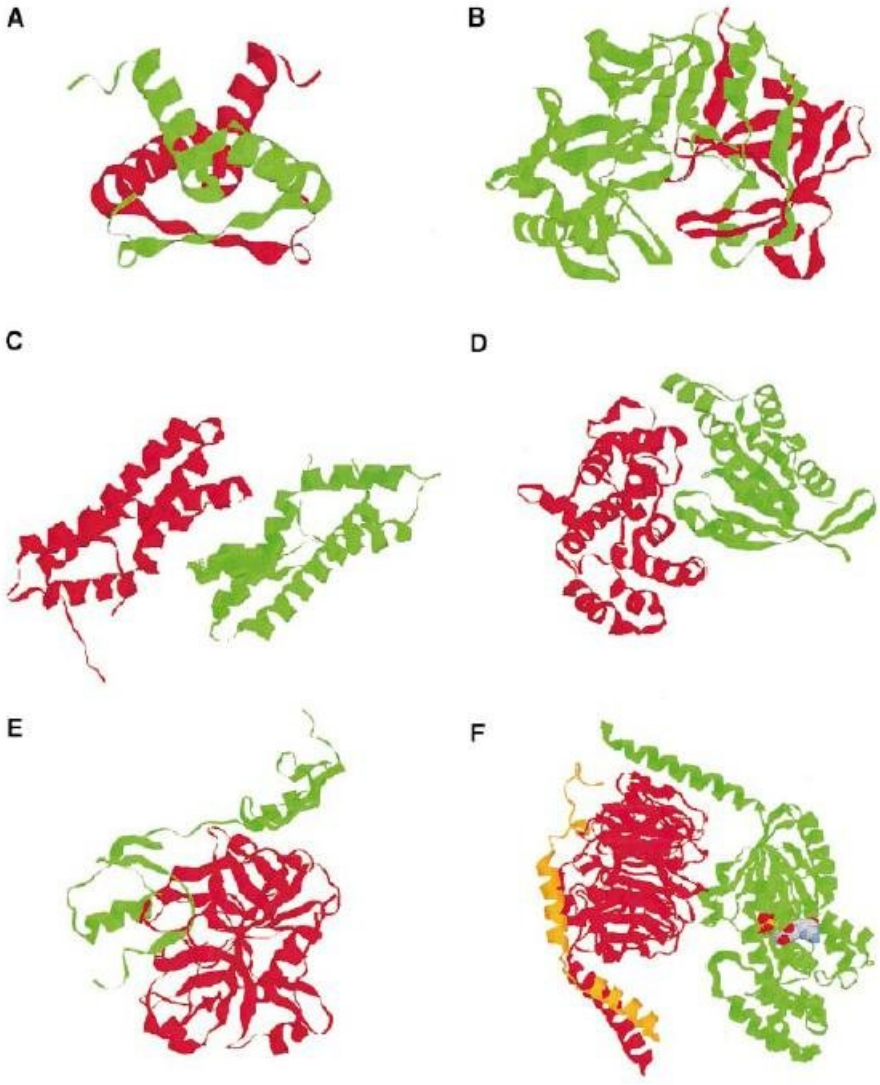
Prvu podskupinu podskupinu čine homo- ili hetero-oligomerski kompleksi. Oligomeri istih ili homolognih proteina mogu se u stvorenom kompleksu organizirati na isti ili drugačiji način (izologno ili heterologno). Izologno pritom znači da reagiraju ista mjesta na površinama odgovarajućih monomera i daljnja oligomerizacija može se dešavati samo na drugim aktivnim mjestima na lancima. U hetero-oligomerskom kompleksu reagiraju različiti aktivni dijelovi površine pa bez kružnog zatvaranja svih aktivnih mjesta može doći do praktički beskonačne agregacije monomera [2].

Proteinski kompleksi mogu se dijeliti i na obligatne ili neobligatne (eng. *obligate*, *non-obligate*) i to predstavlja podjelu druge skupine. Obligatni proteinski kompleksi u pravilu nastaju odmah po izlasku sastavnih proteina iz ribosoma, a neobligatni podalje od mjesta nastanka. U obligatnom proteinskom kompleksu njegovi protomeri, dakle, *in vivo* ne postoje kao samostalne stabilne strukture. Mnoge PDB strukture ipak uključuju neobligatne interakcije protomera koji postoje i nezavisno od kompleksa [2].

Jedno od implementiranih programskih rješenja za određivanje obligatnosti interakcija kojima je nastao proteinski kompleks je NOXclass program. Ovaj program analizira i interpretira kvarternarnu strukturu proteina te na temelju svojstava koja je moguće zadati određuje najprije je li riječ o biološki funkcionalnom proteinu, a ako da, o kojem tipu bioloških interakcija se radi: obligatnim ili neobligatnim. Ovaj program raspoloživ je za preuzimanje na web adresi noxclass.bioinf.mpi-sb.mpg.de/ i može se pokretati isključivo pod operacijskim sustavima Linux ili Unix, a za rad su mu potrebni i NACCESS i

LIBSVM, koji su također *open-source* programi dostupni na web adresama www.bioinf.manchester.ac.uk/naccess/ i www.csie.ntu.edu.tw/~cjlin/libsvm/.

Treća i posljednja podskupina razlikuje tranzijentne od permanentnih kompleksa odnosno interakcija. Ova podjela temelji se na očekivanom vremenu života kompleksa. Permanentni kompleksi obično su vrlo stabilni, dok tranzijentni tipično asociraju i disociraju *in vivo*. Kod tranzijentnih interakcija još razlikujemo slabe i jake tranzijentne interakcije. Slabe znače da kompleks postoji u točki dinamičkog ekvilibrija, gdje se veze konstantno stvaraju i opet pucaju, dok je kod jakih tranzijentnih interakcija potreban molekularni okidač koji će pomaknuti ravnotežu iz ekvilibrija na neku stranu [2].



Slika 3.1 Grafički prikazi proteinskih interakcija: (A) obligatni homodimer P22 Arc represor, (B) obligatni heterodimer katepsin D, (C) neobligatni homodimer lizin, (D) neobligatni heterodimer RhoA (zeleni) i RhoGAP (crveni), (E) neobligatni permanentni heterodimer trombin (zeleni) i rodniin (crveni), (F) neobligatni tranzijentni heterodimer bovin G [2]

Važno je napomenuti da je prijelaz između svaka dva tipa interakcija iz pojedine skupine kontinuiran. Takav je, primjerice, prijelaz između obligatnih i neobligatnih interakcija, koji ovisi o fiziološkim uvijetima i okolišu proteina. Neke interakcije su *in vivo* tranzijentne, ali promjenom određenih uvjeta unutar stanice postaju permanentne. Konkretne vrijednosti tih uvjeta često nam ostaju nepoznanica, ali nam je sama funkcija proteina ponekad dovoljna da zaključimo o kakvim je interakcijama riječ. Tako je poznato da su proteinske interakcije između proteina i liganda, enzima i inhibitora, antitijela i antigena su najčešće permanentne i ireverzibilne [2].

3.2 Tehnike proučavanja proteinskih interakcija

3.2.1 Analitička ultracentrifuga

Trenutno sve popularnija tehnika proučavanja proteinskih interakcija je analitička ultracentrifuga (eng. *analytical ultracentrifugation*, *ACU*) [6]. Iako dosad vrlo nepopularna zbog zahtijevnosti u pogledu tehničke opreme i teorijskog znanja, ova se tehnika sada počinje sve više primjenjivati, prvenstveno zahvaljujući nekim tehničkim pojednostavljenjima, ali i rastućem interesu za proteinske interakcije.

Promatrani uzorak vrti se u ultracentrifugatoru koji može postići akceleraciju do $9800 km/s^2$, a istovremeno se može promatrati pomoću optičkog sustava za detekciju koji koristi apsorpciju ultraljubičastog zračenja. Obično se provodi jedan od dva tipa eksperimenta: eksperiment brzine i smjera sedimentacije (glavna značajka je brzina) ili eksperiment ekvilibrija sedimentacije (glavna značajka je veća preciznost). Kod ekvilibrija sedimentacije brzina vrtnje je relativno niska tako da je sedimentacija uravnotežena difuzijom. Iz gradijenta koncentracije prema udaljenosti u epruveti u kojoj se uzorak nalazi računa se molekularna masa, a na temelju različitih svojstava apsorpcije elektromagnetskog zračenja razlikuju se primjerice homo- ili hetero-oligomeri [6].

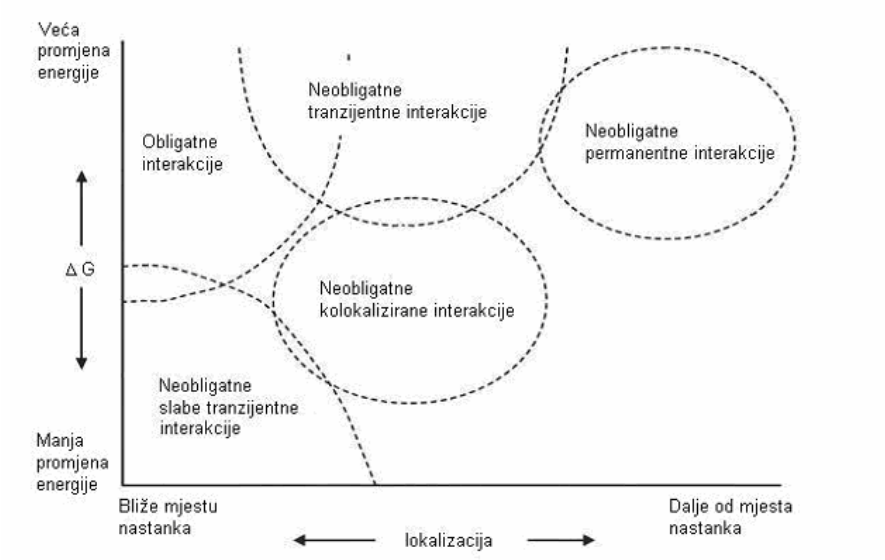
3.1.2 Raspršenje svjetlosti

Statičko i dinamičko raspršenje svjetlosti (eng. *static/dinamic light scattering*) još su dva pristupa proučavanju proteinskih interakcija i kompleksa [6]. Do raspršenja dolazi kad čestice skreću sa svojih inače ravnih trajektorija uslijed neuniformnosti medija po kojem

putuju. Kod statičkog raspršenja je intenzitet raspršenja povezan sa molekularnom masom i koncentracijom proteina, dok se dinamičko raspršenje autokorelira na temelju fluktuacije raspršenja. Ova metoda korisna je kod opisivanja stohiometrije kompleksa pri visokim koncentracijama [6].

3.2 Reguliranje interakcija

Sve proteinske interakcije regulirane su prvenstveno koncentracijom proteina i slobodnom energijom kompleksa. Okoliš proteina u živim stanicama i organizmima iznimno je raznolik pa, generalno govoreći, reguliranje ostvarivanja interakcije možemo podijeliti u tri kategorije [2]: lokalizaciju, lokalnu koncentraciju reaktanata i fiziokemijsku okolinu. Grafički prikaz ovisnosti tipa interakcije proteina o promjeni energije kompleksa i lokalizaciji nalazi se na slici 3.2.



Slika 3.2 Tip proteinske interakcije u ovisnosti o lokalizaciji i promjeni energije kompleksa [2]

Da bi neki proteini reagirali nužno je ne samo da se oni sretnu, nego se svakako moraju susresti njihove odgovarajuće aktivne površine. Ono što se pod lokalizacijom proteinske interakcije podrazumijeva jest da se taj susret može dogoditi odmah pri izlasku iz ribosoma, ili negdje drugdje u ili izvan stanice. Među mehanizmima omogućavanja susreta najvažniji su unutarstanični transport i difuzija. Lokalna koncentracija reaktanata veća od nule svakako je nužan uvjet, ali ne možemo zaključiti da je što veća lokalna koncentracija bolja za odvijanje proteinskih interakcija, jer je poznato da iznad određene vrijednosti

koncentracija reaktanata može i kočiti samu reakciju. Jedan od mehanizama osiguravanja lokalne koncentracije je i usidravanje proteina u membrane ili druge strukturne komplekse. U fiziokemijsku okolinu ubrajaju se koncentracije iona i kemijskih spojeva, promijene pH vrijednosti i temperature, te kovalentne promjene poput fosforilacije. Kod fiziokemijski reguliranih tranzijentnih interakcija može doći do promjena u međusobnom afinitetu protomera za nekoliko redova veličine [2].

3.3 Struktura aktivnih mjesta na proteinima

Aktivna mjesta na površini proteina, koja se mogu okarakterizirati parametrima kao što su veličina površine, polarnost i geometrijski oblik, mogu nam pružiti uvid u vrstu proteinskih reakcija kojima će poslužiti. Statistička ispitivanja dovela su do nekoliko zaključaka na ovom području. Mjesta na kojima proteini reagiraju ili aktivna mjesta površine proteina obično su veća od 1100 $Å^2$. Zbog ostvarivanja kontakata na tim mjestima, svaki protein koji sudjeluje gubi oko 800 $Å^2$ inače dostupne površine koja odgovara dvadesetak aminokiselinskih ostataka. Doprinos dimera, trimera i tetramera površini interakcije u prosijeku iznosi redom 12%, 17,4% i 20% njihovih vlastitih dostupnih površina [22].

Nezavisna istraživanja pokazala su da su mjesta interakcije u oko 84% slučajeva manje ili više ravna. Uz nekoliko iznimki, mjesta interakcije su u pravilu kružnog oblika i u permanentnim i u tranzijentnim kompleksima, s tim da su u permanentnim kompleksima obično veća, manje planarna i međusobno bliže smještena. Uz izgled površine vežemo i pojam komplementarnosti, koji označava koliko si dva oblika aktivnih mjesta međusobno odgovaraju kao trodimenzionalni modeli. Mjesta interakcije kod homodimera, enzim-inhibitor kompleksi i permanentni heterokompleksi su obično najkomplementarniji, dok suprotno svojstvo imaju mjesta interakcije kod antitijelo-antigen kompleksa i neobligatnih heterokompleksa [22].

Neka istraživanja pokazala su da su aktivna mjesta u obligatnim kompleksima u pravilu veća i hidrofobnija od onih u neobligatnim kompleksima, dok su ona neobligatnih kompleksa, čije sastavnice mogu postojati u stabilnom obliku i zasebno, više polarna. Ali samo ovi parametri, bez obzira na to što se nalaze na pravom putu, još uvijek nisu dovoljni da bi njima s sigurnošću mogli definirati afinitete aktivnih mjesta na površini i tako

predvidjeti tip interakcije koji će se desiti [2]. Važno je također primijetiti i da mjesta na površini na kojima dolazi do interakcija između dva proteina A i B, ne moraju biti mjesta interakcija između proteina A i nekoga drugog proteina C [3].

Barem s određenom točnošću moguće je predvidjeti aktivna mjesta na proteinu analizirajući samo njegovu primarnu strukturu. Ovo je pokazao rad pod referencom [3]. Autori ovog rada kreirali su neuronsku mrežu koja je nakon treniranja s testnim skupom postizala rezultate s najviše 94% točnosti predviđanja mjesta na kojima će doći do interakcije, isključivo promatrajući slijedove aminokiselina u lancima.

3.4 Evolucija proteinskih interakcija

Kod proteinskih interakcija često se može primjetiti i utjecaj evolucije i to u smislu optimiziranja biološke funkcionalnosti, fiziološke okoline i kontrolnog mehanizma. Ne samo obligatne, nego svakako i jake i slabe tranzijentne interakcije vrlo su važne za razne stanične procese, te i kod njih postoji potreba za efikasnom kontrolom interakcija. Obligatni kompleksi odražavaju težnju za stabilnošću ili evoluciju funkcije koja zahtijeva više od jednog protomera koji bi odvojeno bili neaktivni. Postoje i primjeri proteina koji su tijekom evolucije promijenom oligomerskog stanja promijenili i funkciju, poput homolognih monomera metionina aminopeptidaze i homodimera kreatinaze. Kako se tijekom evolucije dešavaju takve varijacije u oligomerskoj strukturi unutar obitelji homolognih proteina još nije do kraja razjašnjeno, ali postoji nekoliko različitih ideja. Mala promjena u slijedu može rezultirati promjenom afiniteta i energetski optimalnog oligomerskog stanja. Ipak treba imati na umu da ovo ne znači da su sve proteinske reakcije i oligomerski spojevi od velike važnosti: neki su evoluirali i uspjeli se održati čisto zahvaljujući nedostatku pritiska da nestanu [2].

3.5 Neredundantni skupovi za istraživanje proteinskih interakcija

Stariji skupovi za istraživanje proteinskih interakcija i svojstava proteina općenito su bili vrlo nepraktični jer su uglavnom sadržavali mali broj ručno odabranih PDB datoteka i zbog nemogućnosti razlikovanja pravih proteinskih interakcija od kristalno pakiranih kontakata

bili orjentirani na proučavanje aktivnih mjesta površine unutar jednog lanca nasuprot aktivnim mjestima površine između različitih lanaca [4]. Unaprijeđenje metoda za prepoznavanje kristalno pakiranih kontakata, tranzijentnih i permanentnih kompleksa, te znatno povećanje PDB baze podataka omogućilo je izradu nešto praktičnijih skupova.

Jedan takav skup dobili su Yanay Ofran i Burkhard Rost [4]. Njihov skup na temelju parsiranja datoteka zadovoljava uvjet visoke rezolucije i izdvajanja NMR struktura i ne sadrži lance sa manje od 75% razlike u strukturi. Ova je svojstva zadovoljavalo oko tisuću tada ispitivanih PDB struktura. Razlika između biološki funkcionalnih multimera i kristalno pakiranih monomera određena je pomoću PQS servara. Ovaj problem preciznije je opisan u poglavlju 3.4. Razlika između tranzijentnih i permanentnih kompleksa određena je na temelju na temelju zapisa u SWISS-PROT bazi. Autori su pretpostavili da ako neki lanac postoji u više zapisa u bazi, to znači da je eksperimentalno potvrđeno njegovo postojanje kao zasebnog monomera, što opet povlači da je multimerski kompleks u kojem je on također pronađen tranzijentan a ne permanentan. Dobiveni testni skup autori su iskoristili za analizu šest tipova proteinskih interakcija.

Drugi primjer je neredundantni skup koji su dobili Keskin, Tsai, Wolfson i Nussinov [3]. Njihov skup sadrži lance koji su na temelju obilježja reaktivnih dijelova površine raspodijeljeni u 3799 clustera. Ovi clusteri dodatno su razvrstani u tri skupine na temelju sličnosti ili razlika izgleda lanca i reaktivnog dijela pripadajuće površine. Ovaj skup dostupan je za preuzimanje na Web adresi http://protein3d.ncifcrf.gov/~keskino/nonred-interface.list2 i izrađen je odabirom između svih u vrijeme izrade dostupnih struktura PDB baze podataka. Do sada je skup poslužio za istraživanja kemijskih i fizikalnih svojstava reaktivnih dijelova površine, uspoređivanje broja vodikovih veza u jednom lancu nasuprot njihovu broju na mjestima reakcije, a očekuje se da bi mogao biti koristan i u otkrivanju molekula koje mogu blokirati određene proteine od reagiranja.

4. Podaci

4.1 PDB baza podataka

4.1.1 Uvod

Za izradu testnog skupa proteina koji zadovoljavaju tražena svojstva korištene su dvije baze podataka: PDB i Uniprot. Protein Data Bank (PDB) [25], zbirka javno dostupnih trodimenzionalnih struktura velikih bioloških molekula (makromolekula), uključujući proteine i nukleonske kiseline. Te strukture i ostali dostupni podaci, dobiveni obično eksperimentalno rendgenskom kristalografijom ili nuklearnom magnetskom rezonancijom pohranjuju se u datoteke ekstenzije najčešće .pdb ili .ent, ali postoje i brojni drugi formati. Ova baza podataka, koja je osnovana u laboratoriju Brookhaven National Laboratory 1971. godine i na početku sadržavala samo sedam struktura proteina, danas broji preko 50 000 struktura.

4.1.2 PDB format

Ne postoji jedinstveno standardizirani način zapisa podataka u pdb datoteku. Trenutno važeća verzija PDB datotečnog formata je verzija 3.1 definirana 11. veljače 2008, ali brojne datoteke su pisane i na temelju starijih verzija.

Prvi dio datoteke naziva se naslovni dio (*title section*). Zapisani podaci u cijeloj datoteci obično su organizirani u redove koji započinju riječju koja približno opisuje sadržaj tog retka. Ovaj dio datoteke sadrži zapise s informacijama o samom eksperimentu i prisutnim biološkim makromolekulama. Tu je između ostaloga moguće pronaći ime autora datoteke (AUTHOR), ključne riječi vezane uz ovu strukturu (KEYWDS), način na koji je struktura određena (EXPDTA) i popis datuma izmjena datoteke (REVDAT). Nakon toga slijedi niz oznaka REMARK: rezolucija odnosno preciznost dobivenog opisa (REMARK 2 RESOLUTION), verziju PDB datotečnog formata kojim je pisana datoteka (REMARK 4). Ukoliko je eksperimentalna tehnika određivanja proteinske strukture bila lom rentgenska kristalografija, tada će pod REMARK 200 biti navedeni detalji i parametri korišteni u eksperimentu.

Drugi dio datoteke opisuje primarnu strukturu svih lanaca makromolekule, a treći sekundarnu strukturu makromolekule. Pod oznakom SEQRES u drugom dijelu navedene su redom sve aminokiseline koje grade lance makromolekule. Nakon toga u datoteci slijede dijelovi koji opisuju lokaciju i postojanje disulfidnih veza i drugih elemenata za povezivanje, prostorni raspored atoma makromolekule i još drugin informacija.

4.1.3 Ostali formati

FASTA je naziv za još jedan tekstualni format datoteka za pohranu nizova nukleinskih ili amino kiselina, koje se prezentiraju s po jednim slovom. Ovaj format je vrlo jednostavan i sadrži samo primarnu strukturu makromolekule i komentare, što ga čini vrlo pogodnim za parsiranje i obradu u skriptnim jezicima kakvi su Python ili Pearl. Uobičajene ekstenzije FASTA datoteke mogu biti fa, .mpfa, .fna, .fsa, .fas ili .fasta

Stockholm je format sličan formatu FASTA. Na početku datoteke definira se zaglavlje u kojem se navodi verzija formata, a nakon toga se u svakom retku navodi ime niza aminokiselina te potom sam niz.

Za ovaj rad, zbog potrebnih informacija korištena je baza podataka s datotekama PDB formata.

4.2 UniProt baza podataka

Druga korištena baza podataka je UniProt [27]. UniProt je baza nastala suradnjom između Europskog instituta za bioinformatiku (EBI), Švicarskog instituta za bioinformatiku (SIB), te američkog Protein Information Resource-a (PIR). Ova baza sastoji se od četiri dijela od kojih je svaki optimiziran za druge potrebe pretrage. UniProt Knowledgebase (UniProtKB) je ključni dio za opće informacije o proteinima uključujući njihovu funkciju i klasifikaciju. UniProt Reference Clusters (UniRef) pohranjuje slične nizove aminokiselina u isti zapis sa ciljem ubrzavanja pretrage po sličnosti. UniProt Archive (UniParc) je repozitorij svih jedinstvenih proteinskih slijedova, a poslijednji UniProt Metagenomic and Environmental Sequences (UniMES) je posebni repozitorij za metagenomske podatke. Metagenomika je

relativno nova grana genetike koja se bavi genetskim istraživanjem organizama koje nije jednostavno uzgojiti u laboratorijskim uvjetima.

4.3 Cluster 30%

Slijedovi aminokiselina koje čine lanac ili lance pojedinog proteina često se mogu potpuno ili djelomično ponavljati. Na temelju tih sličnosti unutar lanca proteini se katalogiziraju u takozvane *cluster* datoteke (eng. *cluster* = grozd). *Cluster* datoteka obično sadrži sortirane sve proteinske lance koji imaju sličnost veću ili jednaku zadanoj. U našem slučaju, budući da su nam područje interesa proteini čiji lanci ne smiju imati veću sličnost od 30%, od pomoći će nam biti upravo *cluster* 30%. Uobičajeno se cluster datoteke organiziraju na način da su u pojedinom retku svi lanci uz ime pripadajućeg proteina čija sličnost je veća ili jednaka zadanom postotku. Drugi, ali u načelu vrlo sličan način organizacije je u obliku tri stupca: u prvom se nalazi broj *clustera*, u drugom broj lanca unutar tog *clustera*, a u trećem ime proteina i pripadajućeg lanca. *Cluster* datoteke za postotke sličnosti od 30, 40, 50, 70, 90, 95 i 100 dostupne su za preuzimanje na stranicama RSC PDB organizacije (www.rcsb.org). Ovakvo katalogiziranje omogućava znatna ubrzanja pri pretrazi sličnosti odnosno različitosti proteina, te izradi ovakvoga neredundantnog skupa. Sortiranje unutar *cluster* datoteke može se obavljati i po nekom drugom kriteriju.

4.4 PISA poslužitelj

Kada postupkom rentgenske kristalografije određujemo strukturu nekog proteina, kao što je navedeno, uzorak koji promatramo najprije je potrebno zamrznuti. Odredivši potom strukturu samog uzorka ipak nam ostaje problem ispravne interpretacije rezultata: u tako smrznutom uzorku ne možemo zaključiti koja su mjesta na kojima dolazi do prividne interakcije između proteinskim lancima stvarna mjesta interakcije, a koja su nastala samo kao posljedica kristalizacije (eng. *crystal packing contacts*) i ne realiziraju se *in vivo*. S druge strane, za sve ne-monomere razumno je za pretpostaviti da im lanci neće disocirati za vrijeme kristalizacije, te bi se oni trebali sastojati od više sastavnica. Dodatne se komplikacije pojavljuju ako pretpostavimo postojanje dvaju ili više kompleksa koji se nalaze u stanju dinamičke ravnoteže.

Rješavanje ovog problema vrlo je netrivijalno i postoji nekoliko mogućih pristupa. PQS poslužitelj i program PITA kao metodu koriste analizu i ocjenjivanje mjesta proteinske interakcije za koja se pretpostavlja da su nastale kao rezultat kristalizacije monomera. PQS poslužitelj najviše pažnje posvećuje veličini aktivne površine (u nekim istraživanjima primijećeno je da je površina mjesta kristalnih pakiranja manja od pravih aktivnih površina), ali pazi i na moguće vodikove i sulfidne veze, te slane mostove. Program PITA ocjenjuje potencijalna aktivna mjesta koristeći sofisticirani statistički potencijal i konačno rješenje traži iterativnim biparticioniranjem najveće moguće sastavnice proteina.

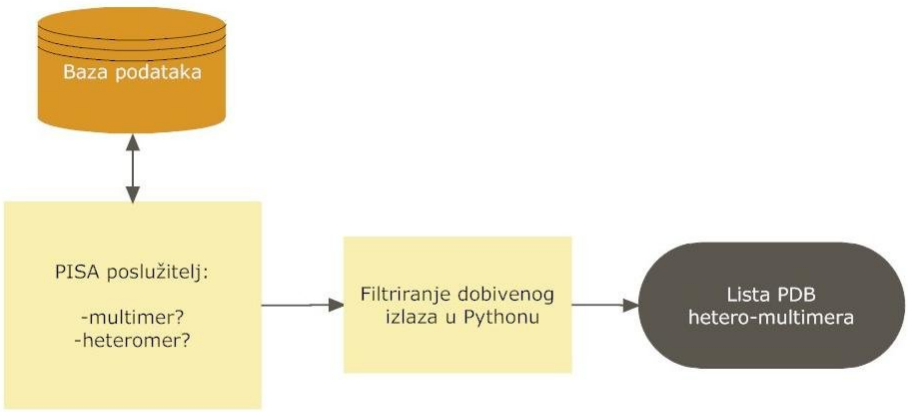
Nešto noviji pristup realiziran je na PISA poslužitelju. Implementirana metoda najprije određuje sve moguće sastavnice proteina, a potom njihova obilježja. Pokazalo se da je ova metoda uspješnija od prethodne dvije i da daje rezultate s točnošću od 90%.

PISA poslužitelj [26] omogućuje pretraživanje vlastite baze u kojoj postoje pohranjeni podaci dobiveni analizom proteina iz PDB baze podataka. Moguće je i zadavanje određene ili neke vlastite PDB datoteke, ali dostupni podaci su unaprijed pohranjeni, jer njihovo dobivanje zbog složenosti PISA algoritma nerijetko zna potrajati. Za većinu proteina rezultati se ipak mogu dobiti u granici od nekoliko minuta. Među parametrima koji se mogu zadati su željeno multimerno stanje, homomernost odnosno heteromernost, postojanje ili nepostojanje disulfidnih veza i slanih mostova, promjena slobodne energije uslijed disocijacije, veličina površine aktivnih mjesta i još neki. Za svaki protein njegovu izračunatu strukturu moguće je vidjeti i u obliku prostornog modela.

5. Metode

5.1 PISA lista hetero-multimera

Lista datoteka od koje sam započeo izradu neredundantnog testnog skupa dobivena je popisivanjem imena svih datoteka PDB baze podataka. Tu listu potom sam usporedio s listom biološki funkcionalnih multimera PISA poslužitelja. Pretraga baze podataka na prije navedenoj Web adresi uz zadane uvjete da želim proteine koji su multimeri (sastoje se od više građevnih jedinica) i heteromeri (građevne jedinice se međusobno razlikuju) rezultirala je listom koja uz imena zadovoljavajućih PDB datoteka sadrži i još neke nepotrebne podatke u vezi aktivnih mjesta na lancima multimera i njihovih karakteristika. Izbacujući iz početne PDB liste sve PDB datoteke koje se ne nalaze u PISA listi dobio sam listu od 7260 biološki funkcionalnih hetero-multimera, što iznosi 14.82% od početnog broja struktura. Ovaj proces grafički je prikazan na slici 5.1.



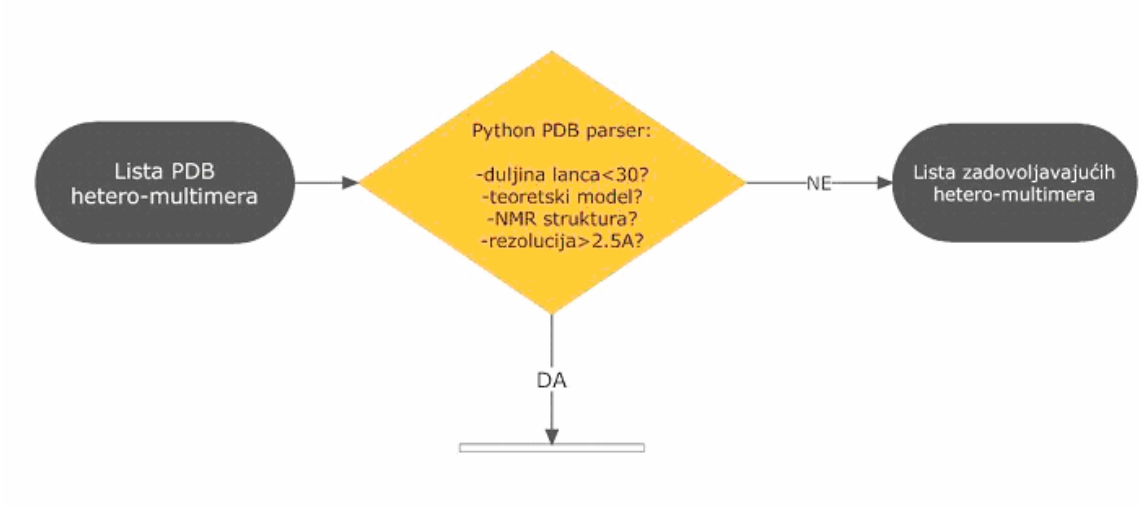
Slika 5.1 Dobivanje liste biološki funkcionalnih hetero-multimera

5.2 Parsiranje u programskom jeziku Python

Da bih provjerio zadovoljava li određeni protein, odnosno PDB struktura svojstva zadana u zadatku potrebno je svaku od njih otvoriti te pronaći vrijednosti ključnih parametara. Dio

koda koji sam koristio za to preuzeo sam se Web izvora http://xed.ucsd.edu/PDB/index.html. Autor ovog PDB parsera Chris X. Edwards pokušao je, kao i mnogi, napraviti PDB parser u programskom jeziku Python, koji će moći na jednostavan način pružiti korisniku sve moguće informacije sadržane u zadanoj datoteci. Izvedba takvog parsera vrlo je složena zbog čestog nestandardnog zapisivanja u PDB datoteke. Ovaj parser pruža stoga i brojne mogućnosti koje mi nisu bile potrebne. U praksi se za potrebe parsiranja često koristi i BIOPython, skup datoteka koje čine jedinstveni parser PDB datoteka, a koristi se u istu svrhu kao i prije navedeni parser.

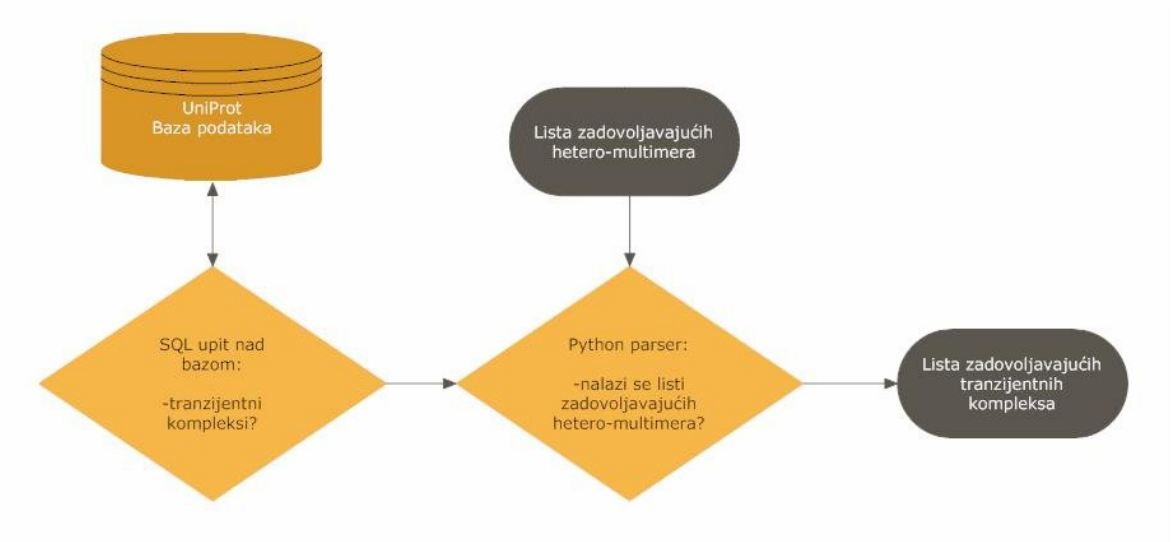
Kreirani vlastiti PDB parser, koji koristi navedeni Edwardsov parser uz još par funkcija kojih u njemu još nema poslužio mi je u svrhu izdvajanja svih onih proteina koji su NMR strukture, teoretski modeli, i imaju rezoluciju veću od 2.5 Å ili nemaju bar dva lanca dulja od 30 aminokiselina. Posljednje svojstvo ispitano je pomoću metode *seqres_list* Edwardsovog PDB parsera. Ova metoda iz datoteke određuje imena svih elemenata i njihovu pripadnost pojedinom lancu. Parsiranjem, koje je grafički pokazano na slici 5.2 dobio sam novu listu od 3213 zadovoljavajućih hetero-multimera, što iznosi 6.55% od početnog broja struktura.



Slika 5.2 Parsiranje u programskom jeziku Python

5.3 UniProt baza i tranzijentni kompleksi

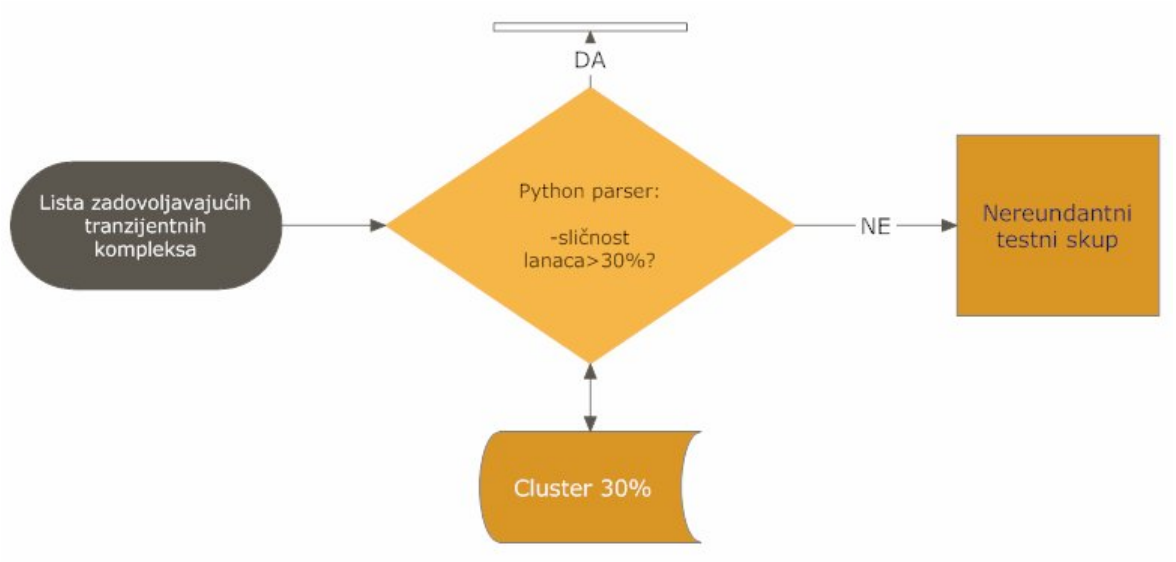
Sljedeći korak u izradi bilo mi je izdvajanje permanentnih kompleksa iz skupa. Ovo sam napravio koristeći UniProt bazu u kojoj permanentni kompleksi imaju sve lance pohranjene u jednom zapisu, a tranzijentni imaju neke lance u različitim zapisima. Dobiveni izlaz iz UniProt baze podataka isparsirao sam u programskom jeziku Python da bih od tih rezultata izabrao samo one koji su se već nalazili u listi hetero-multimera dobivenoj u prethodnom koraku. Ovaj postupak grafički je prikazan na slici 5.3



Slika 5.3 Postupak odabira tranzijentnih kompleksa

5.4 Eliminacija proteina sa sličnim lancima

U zadnjem koraku izrade još je bilo potrebno pomoću *cluster* 30% datoteke eliminirati proteine čiji se lanci od lanaca bilo kojeg proteina po slijedu razlikuju za manje od 70%. Svi takvi proteini su u *cluster* 30% datoteci nalaze organizirani u isti redak. Moj postupak sveo se stoga na odabir po jednog proteina koji ima svoje lance u nekom retku, te eliminacija tog retka i svih ostalih u kojima se nalaze lanci istog proteina. Ovaj postupak ponavljao sam dok iz datoteke nisu izbrisani svi retci, odnosno izabran predstavnik svakog *clustera*. Grafički je ovaj postupak prikazan na slici 5.4.

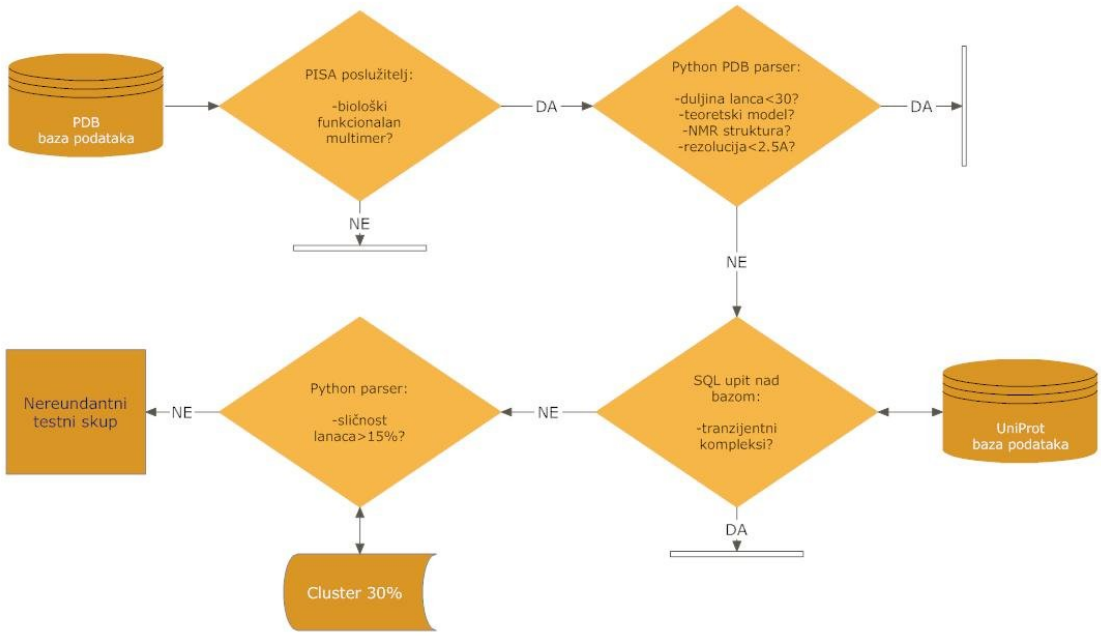


Slika 5.4 Uklanjanje strukturne redundancije

6. Rezultati

Prvi rezultat ovog rada je kreirani postupak dobivanja neredundantnog skupa proteina za predviđanje proteinskih interakcija. Budući da se PDB baza, koja se koristi za izradu istog skupa stalno obnavlja i dopunjuje novim informacijama i strukturama, važnost ovog postupka je u tome da u svakom trenutku omogućava dobivanje najboljeg neredundantnog skupa koji se na temelju dostupnih informacija može dobiti.

Sažeti postupak izrade skupa prikazan je na slici 6.1. Od svih struktura iz PDB baze najprije sam odvojio one s rezolucijom manjom od 2.5 Å, sve NMR strukture i teoretske modele, te sve strukture koje nemaju bar dva lanca dulja od 30 aminokiselina. Filtriranje homolognih proteina i svih onih koji nisu biološki funkcionalni multimeri obavio sam uspoređujući ih s listom istih, dobivenom s PISA poslužitelja. Tranzijentne komplekse eliminirao sam na temelju liste dobivene iz UniProt baze, gdje se tako definiraju svi kompleksi koji nemaju sve lance u istom zapisu. U posljednjem koraku pomoću clustera 30% osigurao sam u rezultantnom skupu bar 30 postotnu razliku svaka dva lanca proteina.



Slika 6.1 Sažeta shema izrade testnog skupa

Drugi rezultat ovog rada je dobiveni neredundantni skup. Restrikcije glede rezolucije znače da su svi dobiveni proteini eksperimentalno određeni vrlo precizno, a budući da je NMR tehnika manje pogodna za proteine kao realtivno velike molekule, iz skupa su izdvojene sve NMR strukture. Biološki funkcionalni multimeri određeni su najtočnijom poznatom tehnikom (PISA). Ovi uvjeti garantiraju kvalitetu ali i maksimalan opseg ovog skupa koji će primjenu naći u ispitivanju ili predviđanju proteinskih interakcija.

Dobiveni skup sadrži 1066 PDB datoteka koje zastupaju neke od ukupno 11733 strukturna *clustera*. Ovaj broj očekivano je veći od onoga kod Ofranovoa i Rosta [3] jer se PDB baza svakodnevno povećava, a time i broj struktura koje zadovoljavaju tražena svojstva. Keskin i suradnici [5] u svom su radu postavljali sve poznate dimere raspodijelili u ukupno 3799 *clustera* i to na temelju izgleda reaktivnog dijela površine. 103 od tih *clustera* sadrže bar pet nehomolognih članova. Taj skup ipak je veći, između ostalog i zbog izostanka uvjeta rezolucije, tehnike dobivanja i broja aminokiselinskih ostataka u lancu.

7. Zaključak

Proteinske interakcije vrlo su složeni procesi i za njihovo razumijevanje potrebno je dobro predznanje iz molekularne biologije, a interes za ovo područje je visok ne samo zbog želje čovjeka da objasni svijet oko sebe, nego i praktične primjene u vidu otkrivanja ljekova za neke bolesti. Za kvalitetno ispitivanje svojstava proteinskih interakcija zato je kao materijal koji se ispituje potreban najveći moguć, ali još uvijek neredundantni testni skup proteina koji bi sadržavao svu otkrivenu raznolikost mogućih aktivnih dijelova proteinske površine.

U svom sam radu stoga najprije pokušao objasniti dio pozadine ovog širokog područja: strukturu aminokiselina koje izgrađuju proteine i način njihova vezivanja, potom obilježja i uloge samih proteina, tehnike dobivanja informacija o njihovoj strukturi, te tipove i obilježja proteinskih interakcija. Potrebne materijale i tehnike za izradu neredundantnoga testnog skupa opisao sam u dva poglavlja. Uz sve korištene materijale naveo sam i njihove izvore i grafički skicirao njihovu uporabu.

Očekujem, i nadam se, da će rezultati i postupci koje sam opisao u ovom radu imati praktičnu primjenu u predviđanju proteinskih interakcija kao i pri analizi aktivnih mjesta proteinske površine ili pak u izradi sličnih skupova iste ili slične namjene.

8. Literatura

[1] Tobin, A. J., Dusheck, J., Asking About Life: *Chemistry and cell biology*, Second Edition, Philadelphia: Harcourt College Publishers, pp. 46-69, 2001.

[2] Nooren, I. M. A., Thornton, J. M., Diversity of protein-protein interactions, Cambridge, European Bioinformatics Institute, 2003.

[3] Ofran, Y., Rost, B., Predicted protein-protein interaction sites from local sequence information, 2003.

[4] Ofran, Y., Rost, B., Analysing Six Types of Protein-Protein Interfaces, 2003.

[5] Kekesin, O., Thai, C.-J., Wolfson, H., Nussinov, R., A new, structurally nonredundant, diverse set of protein-protein interfaces and its implications, 2003.

[6] Royer, C., Protein-Protein Interactions, Biophysics Textbook Online, 2004.

[7] Ofran, Y., Rost, Predicted PP interaction sites from local sequence information, 2003.

[8] Wikipedia, Aminokiselina, 2. lipanj 2008., http://hr.wikipedia.org/wiki/Aminokiselina, 31. svibanj 2008.

[9] Osnovni pojmovi molekularne biologije, http://www.biotehnologija.net/uvod%20u%20biotehnologiju.html, 24. travanj 2008.

[10] Wikipedia, DNA, *Physical and chemical properties*, 31. ožujak 2008., http://en.wikipedia.org/wiki/DNA, 3. svibanj 2008.

[11] Winstead, E. R., New amino acid discovered in Methanosarcina barkeri, 24. ožujak 2002., http://www.genomenewsnetwork.org/articles/05_02/amino_acid.shtml, 4. travanj 2008.

[12] Wikipedia, Disulfide bonds, Disulfide bonds in proteins, 29. travnja 2008., http://en.wikipedia.org/wiki/Disulfide_bond, 10. svibnja 2008.

[13] Medicinski fakultet Sveučilišta u Zagrebu, Zavod za fiziku i biofiziku, http://physics.mef.hr/Predavanja/Struktura_molekule/main11.html, 13. svibanj 2008.

[14] Dokmanić, I., Algoritam za predviđanja proteinskih kompleksa koristenjem razvoja površine u redove funkcija, diplomski rad, Fakultet elektrotehnike i računarstva, 2007.

[15] Decelles, P., Protein structure, 15. srpanj 1999., http://staff.jccc.net/PDECELL/biochemistry/protstruc.html, 2. svibanj 2008.

[16] X-ray Crystallography, http://www.stolaf.edu/people/hansonr/mo/x-ray.html, 2. svibanj 2008.

[17] Wikipedia, X-ray crystallography, 18. svibanj 2008., http://en.wikipedia.org/wiki/X-ray_crystallography, 23. svibanj 2008.

[18] Dobenesque, M., X-ray Cristallography, prezentacija s predavanja, 11. prosinac 2002.

[19] Wikipedia, NMR spectroscopy, 25. svibanj 2008., http://en.wikipedia.org/wiki/NMR_spectroscopy, 29. svibanj 2008.

[20] Rowe, M. D., Smith, J. A. S., Mine detection by nuclear quadrupole resonance, 9. listopad 1996.

[21] Wikipedia, CASP, 16. svibanj 2008., http://en.wikipedia.org/wiki/CASP, 29. svibanj 2008.

[21] Uetz, P., Vollert, C. S., Protein-Protein Interacitions, Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine

[22] Slika preuzeta sa http://lectures.molgen.mpg.de/ProteinStructure/Levels/index.html

[23] Slika preuzeta sa http://www.stolaf.edu/people/hansonr/mo/fig4b.gif

[24] RSCB PDB, 3. lipanj 2008., http://www.rcsb.org/pdb/home/home.do, 15. travanj 2008.

[25] PISA, 4. ožujak 2008., http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html, 23. travanj 2008.

[26] UniProt, www.uniprot.org, 2. svibnja 2008.

Sažetak

Naslov: Automatska izrada testnog skupa za predviđanje proteinskih interakcija

Izrada neredundantnog testnog skupa postupak je koji prethodi mnogim istraživanjima strukture proteina i aktivnih dijelova njihove površine koji imaju funkciju mjesta interakcije. Traženi neredundantni testni skup u ovom radu definiran je kao skup tranzijentnih biološki funkcionalnih multimerskih kompleksa sa bar dva lanca duljine više od 30 aminokiselina, rezolucije najviše 2.5Å, koji nisu teoretski modeli ili NMR strukture i nemaju međusobnu sličnost lanaca veću od 30%. Početni skup struktura dobiven je iz RSCB PDB baze i analizom svojstava u odnosu na zadana u njemu su ostavljene samo zadovoljavajuće strukture. Osim parsiranja PDB datoteka, primjene metode obuhvaćaju i propuštanje kroz PISA poslužitelj, te korištenje dodatnih informacija o tranzijentnosti kompleksa iz UniProt baze. U rezultantnom skupu još su samo uklonjeni proteini sa sličnošću lanaca većom od 30% pomoću *cluster* 30% datoteke sa UniProt Web sjedišta.

Summary

Title: The automated construction of a testing set for predicting of protein interactions

Construction of a non-redundant testing set is a procedure that precedes to many of researches of the protein structure and interaction sites. The non-redundant testing set was defined as the largest set of transient, biologically functional multimers, having at least two chains of more then 30 residues, with resolution not higher then 2.5Å. The NMR structures and theoretical models were not taken into consideration, as well as structures having more then 30% chains similarity. The starting set was obtained from RSCB PDB and analysing its properties we removed all non-satisfying structures. Beside parsing the PDB files, the applied methods have also involved usage of PISA server for differing biologicaly functional multimers from crystal packing contacts, and UniProt database for retrieving information about transient and permanent complexes. From the resoulting set using the UniProt 30% cluster all protein complexes with more then 30% chain sequences similarity were removed.

Ključne riječi:

- neredundantni testni skup
- proteini
- proteinska struktura
- proteinske interakcije
- amino kiseline
- tranzijentni kompleksi
- biološki funkcionalni multimeri

Keywords:

- non-redundant testing set
- proteins
- protein structure
- protein interactions
- amino acids
- transient complexes
- biologically functional multimers