FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
SVEUČILIŠTE U ZAGREBU

DIPLOMSKI RAD br. 76

Alat za prianjanje proteina: modul za utvrđivanje interakcija

Dragana Čolić

Zagreb, lipanj 2010.

Diplomski zadatak

Tema: Alat za prianjanje proteina: modul za utvrđivanje interakcija

Tema eng: Protein docking tool: module for calculation of interactions

Za svaku konformaciju dva proteina potrebno je odrediti: broj i tip interakcije, te njihovu jačinu. Interakcije koje će se promatrati su: vodikove veze, polarne veze, hidrofobne veze, van der Waalsove veze i cistinski mostovi. Isto tako treba promatrati veze koje su nastale neprirodnim prodiranjem jednoga proteina u drugi. Takvim vezama treba pridružiti negativnu jačinu. Prvi korak pri izračunu interakcija je dodavanje vodika za one PDB datoteke koje ne sadrže vodike. Vodike dodati koristeći neke od postojećih aplikacija otvorenoga koda. Za računanje vodikovih veza koristiti informacije o udaljenosti atoma donora i atoma akceptora od vodika, te kutu koji ta tri atoma tvore. Za određivanje polarnih i hidrofobnih veza, te cistinskih mostova koristiti postojeće podatke o pripadnim atomima, njihovoj udaljenosti i koeficijentu jačine veze. Van der Waalsove veze računati koristeći informacije o Van der Waalsovim radijusima pojedinih atoma. Modul je potrebno implementirati u alat za prianjanje proteina na način da je ulaz modula 1000 najboljih konformacija dobivenih prianjanjem, a izlaz je njihovo rangiranje po jačini interakcije.

Modul izvesti u C++, a za konfiguraciju koristiti XML datoteke.

Sadržaj

1. Uvod

U okviru ovog diplomskog rada izrađen je programski modul za analizu proteinskih interakcija. Analiza se izvodi na temelju PDB zapisa geometrijske strukture proteina. Modul je namijenjen za ocjenu rezultata dokiranja proteina pomoću PDT (eng. Protein Docking Tool) alata. Dokiranjem dvaju proteina dobije se kao izlaz tisuću mogućih konformacija tih proteina koje nastaju kao posljedica proteinskih interakcija. Svrha ovog modula je ocijeniti te konformacije na temelju povoljnih i nepovoljnih interakcija koje je moguće očitati iz strukture novonastalog proteina. Povoljne interakcije su one koje sugeriraju veliku vjerojatnost postojanja takve interakcije u prirodi, a nepovoljne one koje sugeriraju malu vjerojatnost. Analiza interakcija provodi se na temelju geometrijskih uvjeta koji se programu zadaju preko XML datoteka.

U prvom dijelu rada dan je uvod u proteinske strukture i interakcije, ukratko su opisane kemijske veze koje tvore proteine te njihove geometrijske aproksimacije korištene u okviru ovog radu. U poglavlju o implementaciji opisani su struktura i način rada programa te mogućnosti prilagodbe programskih postavki u svrhu poboljšanja učinkovitosti. Riječ je o mogućnostima prilagodbe modela interakcija prema kojem se izvodi analiza, smanjivanja prostora pretraživanja postavljanjem praga na udaljenosti atoma i aminokiselina te o mogućnosti paralelizacije. U posljednjem poglavlju predstavljeni su rezultati rada programa. Rezultati sadrže analizu modela interakcija, ocjenu točnosti na temelju ispitnog proteina, te ocjenu ubrzanja koja se postižu paralelizacijom i smanjivanjem prostora pretraživanja.

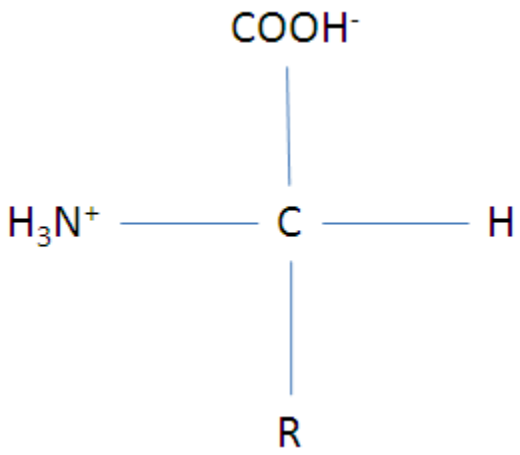
2. Proteinske interakcije

Proteini su velike organske molekule koje imaju važnu ulogu u ljudskom organizmu. Osim što su građevni materijal za žive stanice, tvore enzime koji potiču i ubrzavaju kemijske procese. Funkcija i ponašanje proteina određeno je njegovom strukturom. Najčešće se sastoji od većeg broja organskih molekula, a mogu sadržavati i metale. Primjer proteina koji ima vitalnu važnost za funkcioniranje ljudskog organizma je hemoglobin koji između ostalog sadrži atom željeza. Na njega se veže kisik i kroz krv prenosi do svih stanica u tijelu.

2.1. Struktura proteina

Način na koji se protein povezuje s drugim molekulama ovisi o njegovoj strukturi, a određuje njegovu funkciju u organizmu. Poznavanje strukture proteina omogućuje objašnjavanje i predviđanje funkcionalnosti proteina. Struktura pritom podrazumijeva prostorni raspored građevnih jedinica proteina – atoma i aminokiselina. Prostorni oblik proteina posljedica je kemijskih veza između manjih molekula od kojih je sagrađen. Kemijske veze među pojedinačnim atomima utječu, dakle, na prostorni oblik cijele makromolekule, time i na njezinu funkcionalnost. Zato je preduvjet za analizu proteinskih interakcija razumijevanje strukture proteina i kemijskih reakcija koje ju tvore i održavaju.

Proteini su građeni od lanaca aminokiselina. Aminokiseline su organske molekule koje sadrže amino grupu -NH2 i karboksilnu skupinu –COOH.



Povezivanjem amidne skupine jedne aminokiseline i karboksilne skupine druge aminokiseline nastaje dipeptid. Takva kemijska veza naziva se

peptidna veza. Prema broju povezanih aminokiselina, osim dipeptida, aminokiseline mogu tvoriti polipeptide i proteine.

Lance ljudskih proteina tvore različite kombinacije sveukupno 20 aminokiselina. Različita elektrostatska svojstva atoma u aminokiselinama i međusobno povezanih aminokiselina utječu na prostorni raspored atoma. Istovrsni naboji u atomima i molekulama nastoje se međusobno udaljiti, dok se raznovrsni privlače. Pod djelovanjem odbojnih i privlačnih sila među atomima, oni zauzimaju specifičan raspored u trodimenzionalnom prostoru. Taj raspored opisuje prostornu strukturu proteina koja se može stupnjevati na sljedeći način:

1. primarna struktura – niz aminokiselina koje tvore protein
2. sekundarna struktura – specifični oblici u koje se konformiraju pojedini dijelovi aminokiselinskog lanca
3. tercijarna struktura – trodimenzionalni oblik proteinskog lanca
4. kvartarna struktura – trodimenzionalna struktura proteina koji se sastoji od više proteinskih lanaca.

Tablica 1. Tablica aminokiselina

Naziv aminokiseline	**Hrvatski naziv**	**Troslovna kratica**	**Jednoslovna kratica**
Alanine	Alalnin	Ala	A
Cysteine	Cistein	Cys	C
Aspartic Acid	Asparaginska kiselina	Asp	D
Glutamic Acid	Glutaminska kiselina	Glu	E
Phenylalanine	Fenilalanin	Phe	F
Glycine	Glicin	Gly	G
Histidine	Histidin	His	H
Isoleucine	Izoleucin	Ile	I
Lysine	Lizin	Lys	K

Leucine	Leucin	Leu	L
Methionine	Metonin	Met	M
Asparagine	Asparagin	Asn	N
Proline	Prolin	Pro	P
Glutamine	Glutamin	Gln	Q
Arginine	Arginin	Arg	R
Serine	Serin	Ser	S
Threonine	Treonin	Thr	T
Valine	Valin	Val	V
Tryptophan	Triptofan	Trp	A
Tyrosine	Tirozin	Tyr	C

Protein može imati samo jedan lanac aminokiselina, ali vrlo često ga tvori veći broj lanaca. Protein s dva lanca nastaje evolucijski interakcijom dva proteina građena od jednog lanca. Predviđanje i analiza proteinskih interakcija važna je grana bioinformatike upravo zato jer omogućuje razumijevanje prošlosti i budućnosti evolucije proteina. Odnosno, omogućuje sintezu umjetnih proteina koji će imati željena svojstva. Jednostavan pristup analizi proteinskih interakcija je analiza geometrijske strukture, odnosno analiza prostornog rasporeda atoma koji je za veliko broj proteina eksperimentalno utvrđen i dostupan u RCSB PDB[[1.]] bazi.

2.2. Protein Data Bank

PDB (eng. Protein Dana Bank) je baza eksperimentalno utvrđenih struktura proteina. Pritom se najčešće koriste metode X-ray kristalografije i NMR spektroskopije. Struktura atoma opisana je u posebno definiranom PDB formatu. Glavni dijelovi formata su:

- zaglavlje – opis općih informacija o proteinu
- zapis koordinata atoma u prostoru

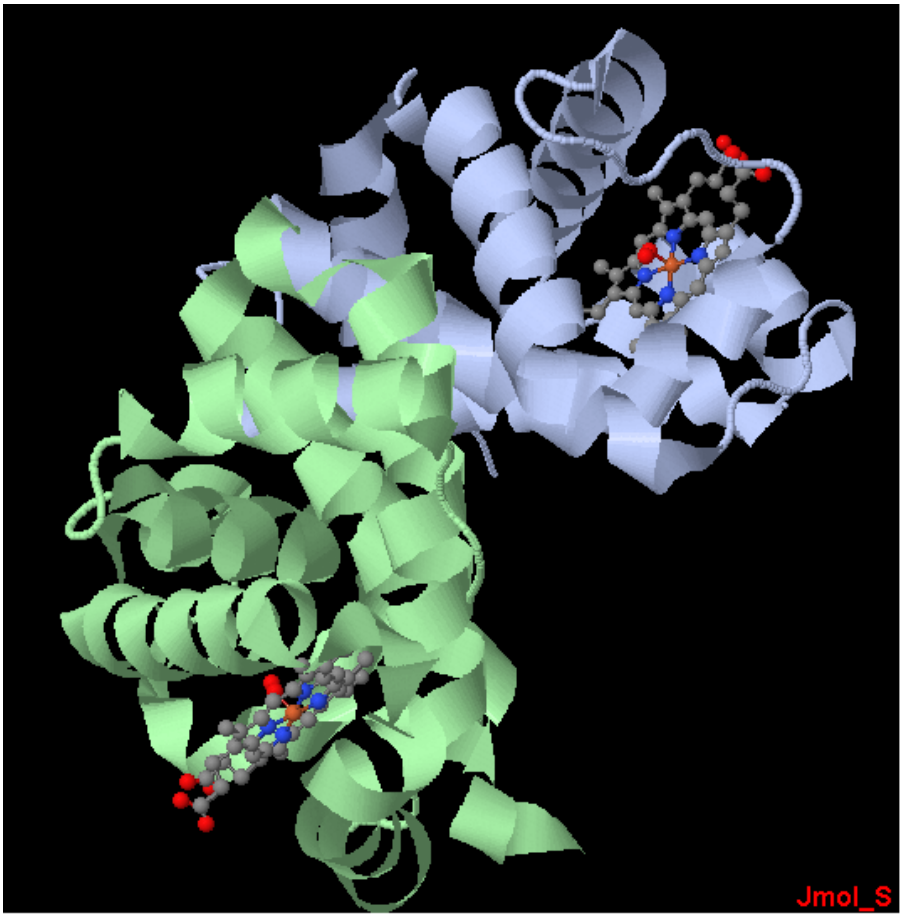
```
HEADER    OXYGEN TRANSPORT  21-MAY-10   3N48
TITLE     STRUCTURAL ANALYSIS OF R-STATE GREYHOUND HEMOGLOBIN
…
ATOM     17  CA  SER A   3       5.051  18.262  10.770  1.00 14.72           C
ATOM     18  C   SER A   3       6.171  19.303  10.734  1.00 14.61           C
ATOM     19  O   SER A   3       7.124  19.155   9.962  1.00 14.60           O
ATOM     20  CB  SER A   3       5.157  17.420  12.046  1.00 14.65           C
ATOM     21  OG  SER A   3       6.487  16.986  12.271  1.00 14.76           O
ATOM     22  N   PRO A   4       6.060  20.368  11.556  1.00 14.63           N
ATOM     23  CA  PRO A   4       7.192  21.281  11.704  1.00 14.44           C
…….
HETATM 2289  CBD HEM B 147       1.740  -8.838  22.713  1.00 19.60           C
HETATM 2290  CGD HEM B 147       1.227  -9.024  24.115  1.00 20.70           C
HETATM 2291  O1D HEM B 147       0.210  -8.378  24.476  1.00 21.12           O
HETATM 2292  O2D HEM B 147       1.840  -9.819  24.870  1.00 21.45           O
HETATM 2293  NA  HEM B 147       4.451 -11.537  17.935  1.00 17.05           N
HETATM 2294  NB  HEM B 147       5.218 -10.459  15.363  1.00 16.44           N
HETATM 2295  NC  HEM B 147       3.198  -8.353  15.645  1.00 16.18           N
HETATM 2296  ND  HEM B 147       2.431  -9.381  18.261  1.00 16.56           N
HETATM 2297 FE   HEM B 147       3.835  -9.905  16.808  1.00 16.35          FE
HETATM 2214  C1A HEM A 142      15.215  10.356 -12.963  1.00 11.79           C
….
```

Svaki redak PDB datoteke započinje s ključnom riječi. Reci koji započinju s ključnom riječi ATOM sadrže koordinate atoma osnovnih lanaca, polimera, u proteinu. HETATM reci opisuju manje molekule koje također tvore strukturu proteina, tipično se radi o metalima ili molekulama vode. Iza ključne riječi *ATOM/HETATM* zapisuju između ostalog:

- redni broj atoma u lancu
- puni naziv atoma koji jednoznačno određuje njegovu ulogu u molekuli
- naziv molekule,
- naziv lanca proteina u kojem se atom nalazi
- redni broj molekule proteina
- x, y, z koordinate atoma u prostoru u mjernoj jedinici angstrem (10^{-10}m)
- naziv elementa.

PDB format točno definira od koje do koje pozicije u retku se opisuje određeni atribut i kakve znakove može sadržavati čime se olakšava parsiranje. Nomenklatura molekula i atoma definirana je prema IUPAC-u[[2.]].

RCSB PDB baza nudi i mogućnost vizualizacije 3D strukture proteina.



Slika 1. JMol vizualizacija PDB strukture proteina 3N48

2.3. Kemijske veze

Podaci o prostornom rasporedu atoma dobiveni iz PDB datoteka koriste se geometrijsku analizu kemijskih veza između molekula i atoma u proteinu. Prilikom traženja interakcija, ima smisla gledati aminokiseline i atome u različitim lancima. Povezanost aminokiselina u istim lancima uglavnom nije izravna posljedica interakcija između različitih lanaca, odnosno različitih proteina pa se ne radi o proteinskim interakcijama. Osim toga, zbog prostorne blizine susjednih molekula, moguće je krivo zaključiti da je blizina atoma posljedica kemijske veza.

Kemijske veze koje se razmatraju u okviru ovog rada su:

- vodikove i pi veze
- polarne veze
- Van der Waalsove veze (VDW)
- cistinski mostovi
- hidrofobne veze

Kao posljedice polarnosti vode javljaju se hidrofobna “vezanja“ nepolarnih dijelova aminokiselina. Osim toga, PDB strukture mogu sadržavati interakcije koje su malo vjerojatne u prirodi, odnosno nepovoljne kemijske interakcije. Može se raditi o pogreškama prilikom utvrđivanja strukture ili o stvarnim konformacijama.

2.3.1. Vodikove veze

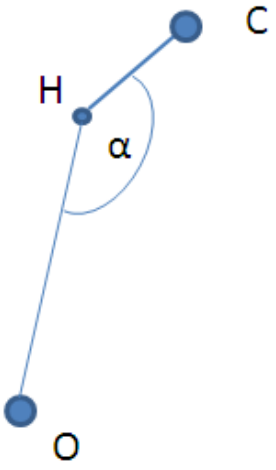
Vodikova veza nastaje između molekula u kojima je atom vodika vezan za drugi atom visoke elektronegativnosti. Elektronegativnost je svojstvo atoma u molekuli da privlači elektrone drugih atoma. Najelektronegativniji elementi iz periodnog sustava elemenata su oni iz gornjeg, desnog kuta: F, N i O. Osim toga vodikove veze mogu tvoriti C, S, Cl i Br.

Povezivanjem vodika i elektronegativnog elementa dolazi do razdvajanja naboja tako što elektronegativni element privuče elektron vodika. Takva molekula naziva se i dipol. Pozitivan kraj jednog dipola privlači negativan kraj drugog dipola i među njima se tvori vodikova veza. Riječ je o jakoj vezi koja se javlja između različitih molekula.

Zanimljivost vodikove veze leži u značaju koji ima prilikom određivanja prostornog rasporeda molekula. Zbog odbijanja elektronegativnih krajeva molekula kao najpovoljniji položaj za atome u vodikovoj vezi procjenjuje se onaj u kojem se svi nalaze na istom pravcu, tj. pod kutom od 180° stupnjeva. U molekuli vode polarnost i elektronegativnost uzrok su specifične konformacije atoma, a time i specifičnih svojstava vode. Također upravo vodikove veze povezuju lance nukleinskih kiselina u DNA i određuju oblik dvostruke zavojnice.

Kod analize proteinskih interakcija, vodikove veze raspoznaju se na temelju geometrijske strukture proteina. Promatraju se udaljenosti između vodika i donora u njegovom lancu i udaljenosti između vodika i akceptora u drugom lancu. Jačina veze određuje se na temelju udaljenosti atoma i kuta kojeg zatvaraju donor, vodik i akceptor.

Pritom se koeficijent jakosti razlikuje za različite atome. Vodikove veze u kojima je donor relativno slabo elektronegativan atom C, smatraju se slabim vodikovim vezama. Također vodikova veza slabija je ukoliko je akceptor kovalentno vezan za vodikov atom.

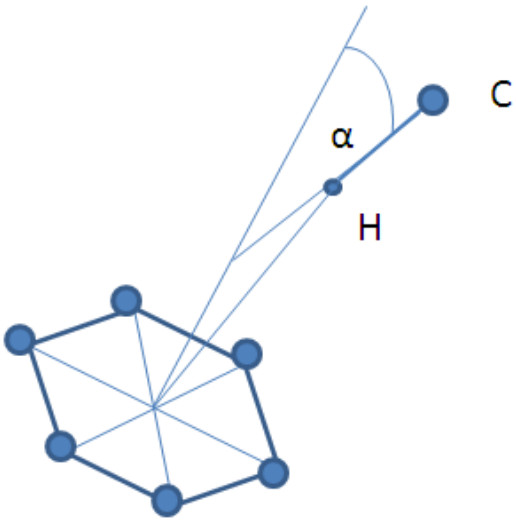


Slika 2. Vodikova veza

2.3.2. Pi veze

Pi veze nastaju između prstenastih struktura ugljikovih atoma i pozitivnog kationa. Prsten predstavlja elektronima bogat akceptor na koji se veže pozitivno nabijen atom. U slučaju aminokiselina, tri su potencijalna akceptora pi veza: TRP, TYR, HIS i PHE.

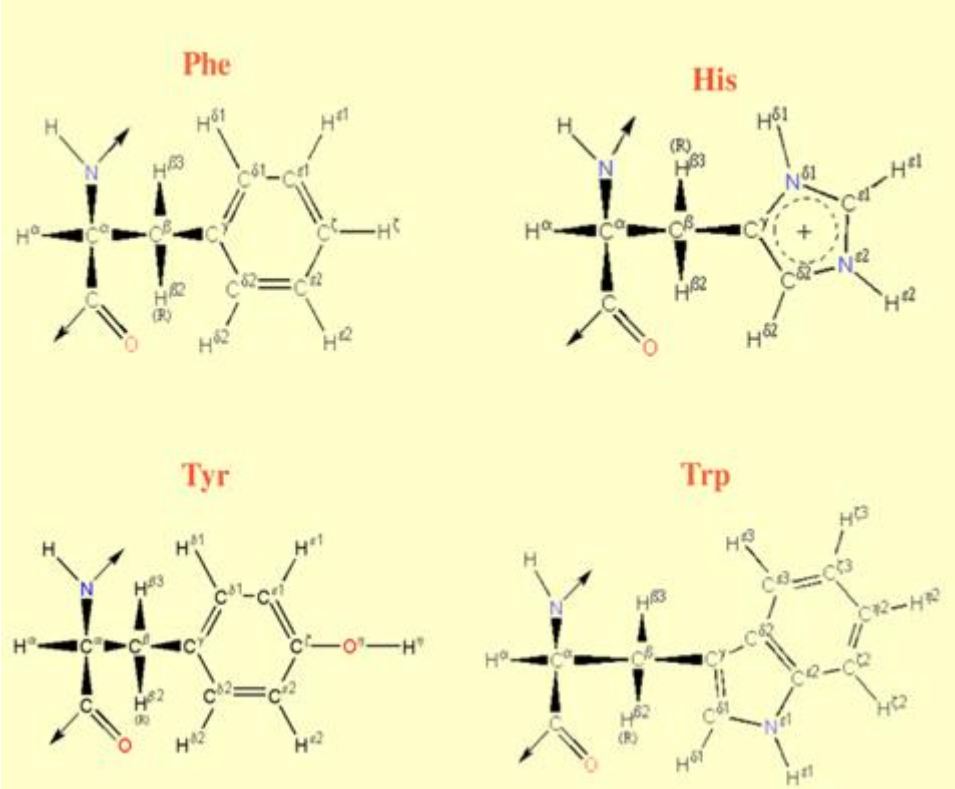
Kation u proteinskim lancima predstavlja atom vodika vezan za elektronegativan element C, N ili O. Povoljne konformacije u slučaju pi veza su one u kojima je kut između ravnine prstena i vektora koji zatvaraju vodik i njegov donor pravi, tj. donor i vodik leže na normali ravnine. Analogno vodikovim vezama, negativan akceptor i negativan donor u povoljnijem su položaju što su međusobno udaljeniji.



Slika 3. Pi veza

Zbog većeg broja atoma u prstenima triptofan i fenilalanin aminokiselina, pi veze koje oni tvore smatraju se jačima od pi veza histidina i tirozina.

Za pronalaženje pi veza važne su vrste atoma donora – elektronegativni atomi koji vodik pretvaraju u kation, udaljenosti između centroida prstena i atoma vodika te kut koji zatvara vektor donor-vodik s normalnom ravnine prstena.



Slika 4. Aminokiseline koje sadrže prstenaste strukture

2.3.3. Van der Waalsove veze

Van der Waalsove veze su slabe međumolekularne veze koje nastaju između nepolarnih molekula. Kretanje elektrona u elektronskom oblaku u određenim trenucima inducira polaritet atoma, tj. pretvara ga u polarnu molekulu. Takav inducirani, nestabilni polaritet molekula povezuje ih Van der Waalsovim vezama. Jačina takvih veza raste s veličinom molekule, odnosno brojem elektrona u njoj i njezinom masom. Geometrijski se Van der Waalsove veze mogu promatraju između bilo koja dva atoma na temelju udaljenosti njihovih Van der Waalsovih radijusa. Ti su radijusi predefinirane, poznate vrijednosti. Van der Waalsova veza slaba je veza čija jačina opada vrlo brzo s udaljenošću.

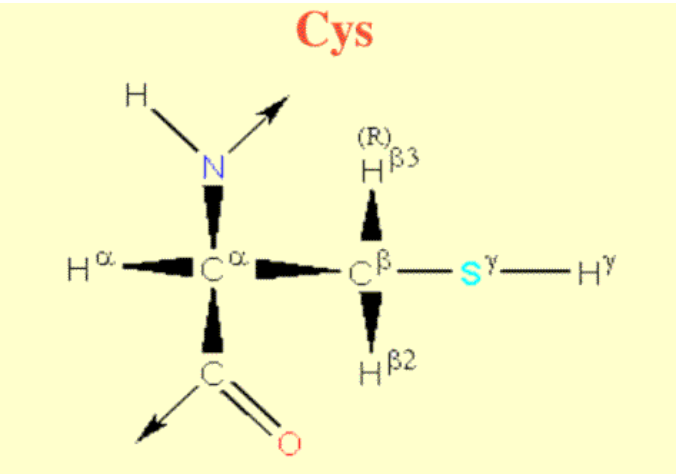
2.3.4. Polarnost i hidrofobnost

Polarne veze nastaju zbog elektrostatskih sila između molekula različitih polariteta. Promatraju se između specifičnih grupa atoma u molekulama koje imaju izrazite pozitivne ili negativne naboje, a njihova jakost opada s kvadratom udaljenosti. Vodikove i pi veze zapravo su poseban oblik polarnih veza

Hidrofobne interakcije nastaju zbog polarnosti vode. Molekule vode kohezivne su i povezuju se polarnim vezama. Nepolarne molekule u vodi ne sudjeluju u takvim interakcijama pa se naizgled odmiču od vode, odnosno skupljaju se u agregate oko kojih interagiraju molekule vode. Hidrofobnost se može detektirati na temelju promjene otapalu dostupne površine molekule. Tj. u hidrofobnoj reakciji dio atoma povući će se u središte agregata. Takvi se atomi lociraju kretanjem molekule vode po površini molekula u interakciji. Ako je neki atom prije interakcije bio dostupan molekuli vode, a nakon interakcije više nije, radi se o potencijalnoj hidrofobnoj interakciji.

2.3.5. Cistinski mostovi

Cistinski mostovi su posebna vrsta veze koja nastaje samo između dvije cistein aminokiseline i to između atoma sumpora (S). Detektira se ako su S atomi međusobno udaljeni za manje od 2.5 A. Čak niti za vrlo male udaljenosti, ova vrsta veza ne smatra se neprirodnim sudarom atoma. Riječ je naime o blizinama koje sugeriraju postojanje kovalentne veze. Budući da kovalentna veza nastaje između atoma unutar jedne molekule, smatra se neprirodnim postojanje takve veze između atoma različitih molekula. U slučaju cistinskih mostova, takva je povezanost prihvatljiva.



Slika 5. Struktura Cistein aminokiseline

2.3.6. Nepovoljne interakcije

Osim nabrojanih interakcija koje sugeriraju prirodnu povezanost molekula, u strukturama proteina, opisanim u PDB bazi, mogu se naći i tzv.

nepovoljne interakcije. To su interakcije koje ukazuju na neprirodnu i malo vjerojatnu konformaciju atoma, odnosno molekula. Najčešće se radi o prostornom rasporedu koji ne podržavaju elektrostatske odbojne sile među promatranim molekulama. Također, uzima se u obzir i neprirodan sudar atoma. Ako su bilo koja dva atoma koja pripadaju različitim molekulama na međusobnoj udaljenosti koja se može uspoređivati s udaljenostima unutar iste molekule, radi s o sudaru molekula. Takva se interakcija smatra negativnom, ali dosta brzo opada s udaljenošću.

2.3.7. Jakosti veza

Jakosti različitih interakcija računaju se u ovisnosti o različitim parametrima. Uvijek se uzima u obzir udaljenost između atoma u interakciji, a za vodikove i pi veze razmatra se i kut između donora, vodika i akceptora. U slučaju pi veza akceptor se definira ravninom i centroidom prstena. Ponašanje jačine interakcija procjenjuje se kao linearno ili kvadratno slabljenje s udaljenošću ili s veličinom kuta.

Interakcije modelirane za potrebe ovog rada tretiraju cistinske mostove i vodikove veze kao vrlo jake interakcije Njihova jakost linearno ovisi o udaljenosti atoma. Kod vodikovih veza, jakost ovisi i o kutu, također linearno.

$$r - d(donor, akceptor)$$

$$\alpha(donor, vodik, akceptor)$$

Pi veze slične su vodikovima, osim što je u njihovom slučaju akceptor atomski prsten. Pi veze procjenjuju se slabijima od vodikovih te njihova jakost opada s kvadratom udaljenosti i kuta.

Polarne, hidrofobne i nepovoljne veze opadaju također s kvadratom udaljenosti. Riječ je zapravo o pozitivnim i negativnim polarnim vezama. Hidrofobnost zapravo nije interakcija, već ponašanje molekule kao posljedica interakcije između atoma i molekule vode. Ta je interakcija posljedica polarnosti pa se kao polarna i tretira. Najslabije od svih veza su Van der Waalsove. One su posljedica sekundarnih pojava polariteta zbog čega su slabe i kratkog dometa. U tablici su prikazane procjene jakosti interakcija kojima se u programu mogu dodati konstante proporcionalnosti.

Tablica 2. Jakosti interakcija

R.br. veze	Naziv veze	Procjena jakosti
1.	Cistinski mostovi	1/r
2.	Vodikove veze	1/r, α/180
3.	PI veze	1/r, (α/90)^2
4.	Polarne veze	1/r^2
5.	hidrofobne	1/r^2
6.	Van der Waalsove	1/r^4
7.	Nepovoljne	1/r^2

3. Implementacija

Modul za analizu proteinskih interakcija prima ulaznu datoteku s popisom putanja do PDB datoteka koje treba ispitati, a ispisuje izlaznu datoteku s rangiranim popisom PDB datoteka prema dobivenim dobrotama.

Primjer ulazne datoteke:

```
/putanja/pdb_1.pdb
/putanja/pdb_2.pdb
….
/putanja/pdb_N.pdb
```

Primjer izlazne datoteke:

```
/putanja/pdb_4.pdb3.456
/putanja/pdb_50.pdb2.111
….
/putanja/pdb_80.pdb0.001
```

Dobrote pojedine PDB datoteke izračunavaju se na temelju modela kemijskih interakcija. Prije analize interakcija dodani su vodici koji nedostaju u PDB datotekama uz pomoću alata *Reduce*[[4.]]. Razlog zbog kojeg vodici u velikom broju PDB datoteka nisu ugrađeni je u metodi X-ray kristalografije kojom se dobivaju 3D strukture proteina, a koja ne detektira vodikove atome. Zato se oni dodaju modeliranjem. Budući da vodici čine oko 50 % atoma u proteinima i 35% atoma u nukleinskim kiselinama, njihovo dodavanje značajno doprinosi definiciji strukture. Dodavanje vodika preporučuje se obaviti prije samog procesa dokiranja, jer dodavanje vodika na svaku od cca. 1000 rezultantnih konformacija predstavlja nepotreban i značajan gubitak vremena u odnosu na dodavanje vodika u dvije izvorne datoteke.

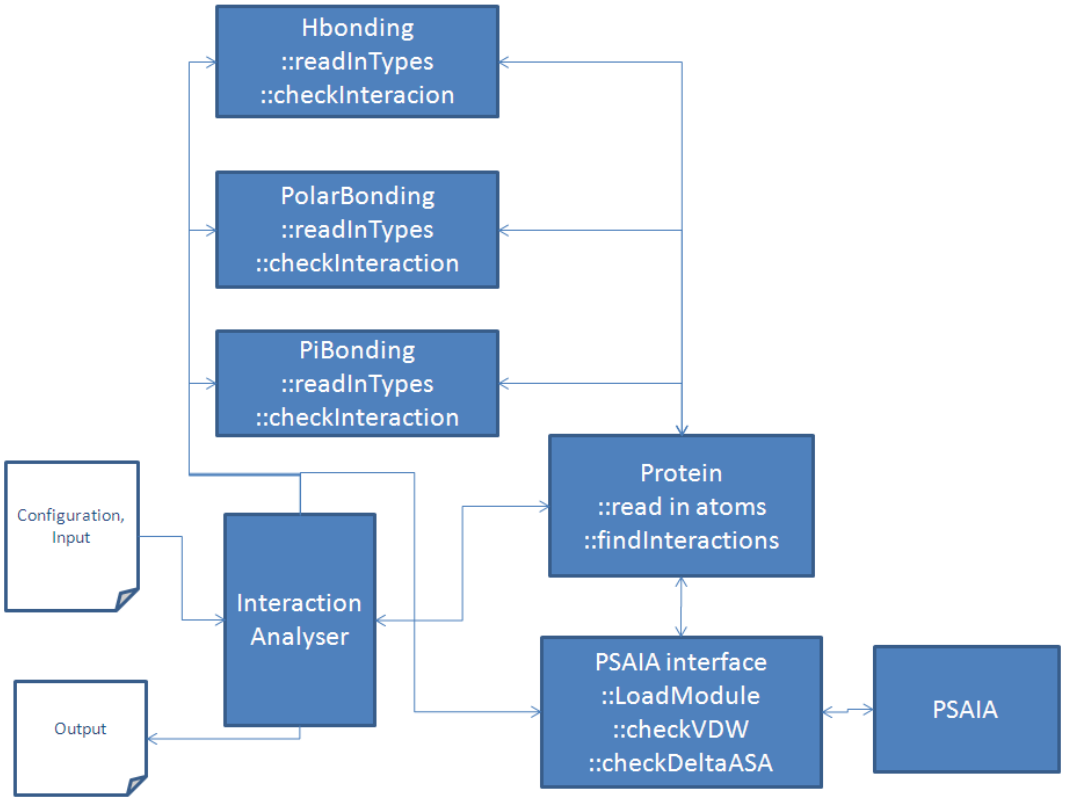
Za analizu hidrofobnih i Van der Waalsovih interakcija dostupne su za korištenje prilagođene funkcije PSAIA alata za analizu proteinskih interakcija[[3.]]. Pristupa im se preko posebno oblikovanog sučelja.

Prilikom pokretanja programa zadaje se putanja do konfiguracijske datoteke.

```
machine$ ./InteractionAnalyser /path/configurationFile
```

U konfiguracijskoj datoteci nalaze se sljedeće definicije:

- putanja do datoteke koja sadrži popisa PDB datoteka koje treba obraditi
- putanja do datoteke koja sadrži definicije vodikovih veza
- putanja do datoteke koja sadrži definicije pi veza
- putanja do datoteke koja sadrži definicije ostalih veza
- putanja do datoteke s VDW radijusima atoma
- zadane veličine za provjeru algoritmima modula PSAIA:
  - veličina sferne sonde za ispitivanje ASA-e
  - debljina sloja za ispitvanje ASA-e
  - kritična promjena ASA-e
  - prag VDW udaljenosti
- maksimalna udaljenost aminokiselina
- maksimalna udaljenost atoma
- udaljenost donora vodika i vodika
- višedretveno izvođenje
  - broj dretvi

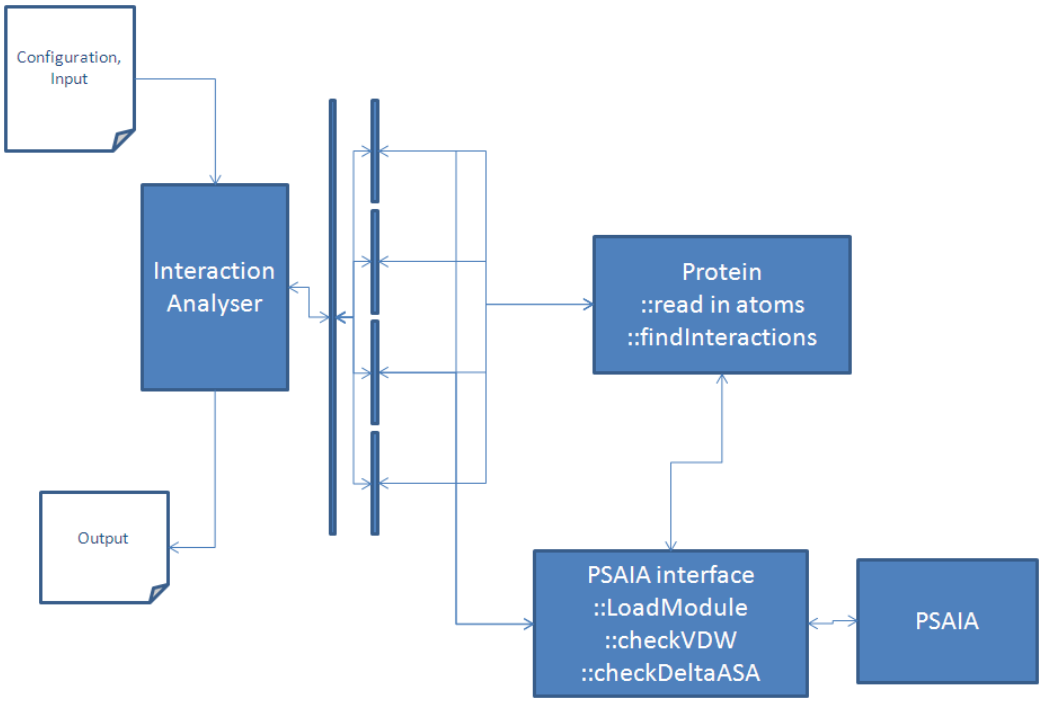


Slika 6. Blok shema sustava

Sustav se u osnovi može podijeliti na sljedeće podsustave:

- PSAIA sučelje
  - koristi se ako su zadani VDW radijusi
  - učitava se za svaki protein posebno
  - pristupa mu se preko sučelja i osnovnih funkcija za
    - provjeru ASA-e,
    - provjeru VDW udaljenosti
- Sustav za pretraživanje prostora proteinskih interakcija
  - učitava atome iz PDB datoteke
  - učitava definicije veza iz konfiguracijskih datoteka
  - pokreće pretraživanje interakcija u proteinu
- Sustav za paralelizaciju posla
  - raspodjeljuje datoteke za obradu po većem broju dretvi
  - sakuplja rezultate u zajedničku listu interakcija

Nakon što su svi rezultati učitani sortiraju se i upisuju u izlaznu datoteku.



Slika 7. Shema paralelizacije posla

Objekti koji sadrže definicije tipova veza (Hbonding, PolarBonding i PiBonding) instanciraju se i učitavaju iz XML datoteka jednom, na početku rada programa. Objekt koji sadrži listu atoma i pretražuje interakcije, učitava se također jednom, na početku rada programa. Prilikom obrade različitih datoteka:

- briše se stara lista interakcija, interakcijskih parova i atoma
- učitava se nova lista atoma
- određuju se iznova interakcijski parovi za ispitivanje
- ispituju se iznova interakcijski parovi
- popunjava se iznova lista interakcija

Instanciranje novih objekata prilikom pretraživanja oduzima relativno puno vremena pa su za čuvanje podataka o interakcijama u memoriji korištene strukture podataka.

Prilikom učitavanja konfiguracijskih postavki, za numeričke pragove i numeričke vrijednosti postavljeni su predefinirani iznosi. Oni se koriste ako u konfiguracijskim podacima ti podaci nisu zadani ili su neispravno zadani.

Podaci koji se učitavaju iz datoteka, atomi proteina, tipovi veza te tipovi grupa aminokiselina i atoma, zanemaruju se ako su neispravno formatirani. U slučaju atoma očekuje se isključivo PDB format. U slučaju grupa i tipova veza, očekuju se predefinirani XML formati opisani u nastavku poglavlja.

Ako datoteka koja sadrži definicije određenog tipa veze nije dostupna, taj se tip veze ne pretražuje u programu, npr. vodikove veze.

Ako datoteka za obradu koja sadrži konformaciju u PDB formatu nije dostupna, ta se konformacija zanemaruje. U slučaju neispravnog formata pojedinog atoma, samo se taj atom u proteinu izostavlja.

Uz pogreške koje nisu fatalne, npr. nedostatak datoteke s definicijom VDW radijusa, samo se javlja informacija o pogrešci, a eventualno instanciranje objekta, odnosno korištenje funkcionalnosti se izostavlja. U navedenom slučaju nedostatka VDW radijusa, to bi značilo ne korištenje provjera ASA-e i VDW udaljenosti pomoću PSAIA modula. U slučaju fatalne pogreške kao što je nedostatak datoteke s popisom PDB putanja, program se prekida.

Rad programa, dojave o fatalnim i nefatalnim pogreškama moguće je pratiti kroz ispise na standardni izlaz ili u elektronički dnevnik rada (eng. log).

3.1. Ubrzavanje algoritma smanjenjem prostora pretraživanja interakcija

Udaljenost između aminokiselina iznad koje se ne ispituju kemijske veze procjenjuje se udaljenošću njihovih CA atoma. CA je oznaka središnjeg atoma aminokiselina na koji su vezane amidna, karboksilna skupina, vodik i ostatak molekule. Podešavanjem ovog parametra, skupa s parametrom maksimalne udaljenosti među parovima atoma koji se uzimaju u obzir, reducira se prostor pretraživanja i skraćuje vrijeme izvođenja programa.

Pretraživanje interakcija u cijelom proteinu moguće je, ali u ovu svrhu nije korišteno jer veze unutar jednog lanca nisu bitna informacija za dokiranje proteina. Novi proteini nastaju povezivanjem različitih proteina koji u novostvorenom proteinu ostaju vidljivi kao odvojeni lance. Pritom se kod

dokiranja proteina dvije ulazne makromolekule mogu promatrati kao dva lanca novog proteina. Traženje interakcija unutar jednog lanca nema važnost za dokiranje, ali vremenski opterećuje program, zato se izbjegava.

3.2. Ubrzavanje algoritma paralelizacijom posla

Budući da je zadaća programa analizirati veći broj datoteka, cca. 1000 njih, kao mogućnost ubrzavanja obrade nameće se paralelizacija. Radi se o trivijalnom problemu za paralelizaciju jer nema potrebe praktički za nikakvom komunikacijom između procesa, a računski su procesi zahtjevni. Za paralelizaciju na višeprocesorskim računalima koristi se višedretveni rad. Primjerice, ako se zadaje veći broj dretvi tada se broj proteina koji se obrađuju u jednoj dretvi izračunava kao:

$$n - broj\ dretvi$$

$$k - ukupno\ pdb$$

$$pdbPoDretvi = \lceil n/k \rceil \ , \ dretva = 1, \dots, k-1$$

$$pdbPoDretvi = k - n\%k \ , \qquad dretva = k$$

3.3. Definiranje vrsta veza

Vrste veza koje se ispituju definiraju se kroz XML datoteke. Za vodikove i pi interakcije definiraju se udaljenosti na temelju kojih se uočavaju donori i akceptori.

Tablica 3. Udaljenosti donora i akceptora

	H-veza	**PI-veza**
Udaljenost(donor, H) / angsterm	1.2	1.2
Udaljenost(akceptor, H) / angstrem	2.7	4.0

Tablica 4. Vodikove interakcije

Donor	**Akceptori**	**Vrsta veze**	**Akceptor kovaletno vezan na H**	**Faktor jačine**
C	O,S,N	slaba	NE	2.0
C	OS	slaba	NE	1.5
O,N,S	O,N,S,Cl, Br	jaka	NE	3.0

O,N,S	O,S,Cl, Br	srednja	DA	2.5

Za pi interakcije definiraju se prsteni koji mogu biti akceptori vodika. Za potrebe diplomskog rada uzeti su u obzir samo prsteni navedeni u donjoj tablici.

Tablica 5. Pi interakcije

Donor	Akceptor		Faktor jačine
	Aminokiselina	Atomi	
C, O,N	HIS	CG-ND1-CD2-CE1-NE2	1.5
C, O,N	TRP	CD2-CE2-CE3-CZ2-CZ3-CH2	2.0
C, O,N	PHE	CG-CD1-CD2-CE1-CE2-CZ	2.0
C, O,N	TRP	CG-CD1-CD2-NE1-CE2	2.0
C, O,N	TYR	CG-CD1-CD-CE1-CE2-CZ	1.5

U XML formatu pi i vodikove interakcije definirane su kroz glavne atribute prikazane na sljedećem isječku iz datoteke:

```
<donor_distance>1.2</donor_distance>
<hydrogen_bonds>
    <acceptor_distance>2.7</acceptor_distance>
    <DHAangle>90-180</DHAangle>
    <hydrogen_bond id = "1">
      <name>weak</name>
      <donor>C</donor>
      <acceptor>O</acceptor>
      <acceptor_occupied>0</acceptor_occupied>
      <factor>2</factor>
    </hydrogen_bond>
          .....
<hydrogen_bonds>
<pi_bonds>
     <angleDHC>60-180</angleDHC>
     <angleDCN>40-90</angleDCN>
     <rings>
          <ring id="1">
               <residue_name>HIS</residue_name>
               <atoms>
                    <atom>CG</atom>
                    <atom>ND1</atom>
                       ....
               </atoms>
          </ring>
            .....
     </rings>
     <bonds>
          <pi_bond id = "1">
```

```
            <donor>C</donor>
            <acceptor_id>1</acceptor_id>
            <distance_CH>4</distance_CH>
            <distance_CD>5</distance_CD>
            <factor>1.5</factor>
        </pi_bond>
          ….
    </bonds>
</pi_bonds>
```

Sve ostale vrste interakcija (polarne, cistinske, hidrofobne, nepovoljne, Van der Waalsove) definiraju se između dva atoma. Za njih je pretraživanje interakcija ostvareno po grupama atoma.

Na sljedećem primjeru XML datoteke prikazan je format pomoću kojeg se modelira takva vrsta interakcije, a preko kojeg je moguće definirati interakciju između bilo koje dvije vrste molekula i atoma na zadanoj udaljenosti.

```
<interaction_settings>
      <z_slice>1.4</z_slice>
      <delta_asa>1</delta_asa>
      <r_slovent><r_slovent>
      <interaction_groups>
            <interaction_group id="11">
                  <group_type>polar+</group_type>
                  <residue>
                      <name>ASN</name>
                      <atom_type>ND2</atom_type>
                  </residue>
                       …
            </interaction_group>
                 …
      </interaction_groups>
      <interaction_types>
            <interaction_type id="1">
                  <type>van der waals</type>
                  <group id="1">any</group>
                  <group id="2">any</group>
                  <distance>4</distance>
                  <distance_factor>1/r^4</distance_factor>
            </interaction_type>
                    …
       </interaction_types>
<interaction_settings>
```

Parametri zadani preko XML konfiguracijskih datoteka, mogu se dodavati, brisati i mijenjati kako bi se poboljšali rezultati.

Kod Van der Waalsovih i hidrofobnih interakcija atomi na zadanim udaljenostima tretiraju se kako kandidati za veze. Radi li se zaista o vezama procjenjuje se:

- uspoređivanjem udaljenosti Van der Waalsovih radijusa za Van der Waalsove veze te
- uspoređivanjem površine dostupne otapalu u lancu i u cijelom proteinu.

U tu svrhu koriste se prilagođene funkcije spomenutog programskog alata za analizu interakcija PSAIA.

4. Rezultati

Rad modula ispitan je na računalu s procesorom Intel(R) Xeon(R) CPU E5530, @2.40GHz sa 16 jezgri i s 50GB RAM-a. Budući da se definicije veza i pretraživanje interakcija zadaje parametarski, a potrebno je obraditi velik broj datoteka, veći naglasak u implementaciji programa dan je na brzinu rada.

4.1. Analiza modela interakcija

Na devet nasumično odabranih proteina iz RCSB PDB baze izvedena je kvantitativna i kvalitativna analiza modela interakcija koji je korišten za potrebe diplomskog rada. Budući da za nasumično odabrane proteine nisu poznate informacije o stvarnim interakcijama, ocjena točnosti nije bila moguća. Osnovni podaci o testnim proteinima navedeni su u donjoj tablici.

Tablica 6. Popis ispitnih proteina

Naziv proteina	Ukupna duljina	Broj lanaca	Broj pronađenih veza između atoma (aminokiselina)
3bn1	373	4	3530(448)
3hmb	157	3	1044(122)
1ecp	238	6	6778(869)
3hmv	378	2	662(118)
2qs9	194	2	73(15)
3n55	127	2	1352(100)
3i3b	1023	4	17706(1768)
3knl	4204	26	12335(2324)
3if2	444	2	3612(400)

U sljedećim tablicama su brojevi i jačine pronađenih interakcija prikazani po vrstama. Prikazani su i kvalitativni (prema jačini), odnosno kvantitativni (prema količini) udjeli različitih vrsta interakcija među ukupno pronađenim interakcijama za sve proteine.

Tablica 7. Rezultati analize po tipovima interakcija

Naziv proteina	Polarne veze (broj jačina)	Pi veze (broj jačina)	H veze (broj jačina)	Van der Waalsove veze (broj jačina)	Cistinski mostovi (broj jačina)	Hidrofobne (broj jačina)	Nepovoljne (broj jačina)	Sudar (broj jačina)	Ukupna jačina
3bn1	1089 191,672	0 0	103 68,302	1263 8,509	0 0	890 51,060	185, -20,495	0 0	299,048
3hmb	486 82,772	0 0	17 10,395	321 3,22	0 0	83 4,869	137 -16,272	0 0	84,984
1ecp	2243 358,687	0 0	149 92,313	2193 13,573	0 0	1557 87,885	636 -79,760	0 0	472,698
3hmv	155 29,794	0 0	11 6,251	218 1712	0 0	246 12,979	30, -3,361	2 -0,534	1757,129
2qs9	31 4,403	0 0	1 0,483	22 0,127	0 0	14 0,829	5 -0,601	0 0	5,241
3n55	416 85,427	0 0	32 19,812	582 4,332	0 0	159 8,870	163 -19,734	0 0	98,707
3i3b	5983 1059,869	0 0	275 185,828	7304 49,475	0 0	2285 131,986	1859 -227,212	0 0	1199,946
3knl	2058 323,678	0 0	1264 765,309	4375 29,022	0 0	4199 241,915	439 -58,607	0 0	1301,317
3if2	1132 202,803	0 0	94 63,505	1344 8,766	0 0	788 45,266	254 -28,482	0 0	291,858

Tablica 8. Kvalitativna i kvantitativna važnost interakcija

Vrsta veze	Udio u ukupnom broju pronađenih veza	Udio u ukupnoj jačini pronađenih veza
Polarne	0,289	0.420
PI	0.000	0.000
H	0,041	0,220
Van der Waals	0,370	0.332
Cistinski mostovi	0.000	0.000
Hidrofobne	0,217	0.106
Nepovoljne interakcije	0,080	-0.082
Sudari	0,000	0.000

Slika 8. Grafikon kvantitativnih udjela pojedinih vrsta interakcija

Slika 9. Grafikon kvalitativnih udjela pojedinih vrsta interakcija

Pregledavanjem tablica pronađenih interakcija vidljivo je kako su Van der Waalsove (VDW) i hidrofobne veze često pronađene, ali su relativno slabe, tj. kvantitativan udio u svim pronađenim interakcijama im je veći od kvalitativnog udjela.

Cistinski mostovi i pi interakcije, kao specifične vrste interakcija, nisu pronađene niti u jednom ispitanom proteinu. Za razliku od VDW veza, broj pronađenih vodikovih veza je malen, ali su značajno doprinosile ukupnoj povezanosti proteina. Vodikove veze su rjeđe od ostalih polarnih i VDW veza jer imaju stroži prag na udaljenost. U proteinima su pronađena i dva sudara atoma koja ukazuju na malu vjerojatnost postojanja pronađenih konformacija atoma u prirodi. Isto tako iz usporedbe interakcija, vidljivo je kako su nepovoljne interakcije pronađene u znatno manjoj mjeri nego povoljne. To može biti posljedica prirodne strukture proteina u kojoj nepovoljnih interakcija ima malo, ali može biti i posljedica nedostatno definiranih interakcija među proteinima.

Budući da broj i jačina veza izračunatih na ovaj način ovisi o veličini proteina, potrebno ih je normalizirati s veličinom proteina, odnosno s brojem mogućih interakcijskih parova, kako bi bile usporedive. U općem slučaju broj promatranih parova atoma proporcionalan je s kvadratom broja atoma u proteinu.

$$k - broj\ atoma\ u\ proteinu$$

$$brojParova = k^2$$

Tablica 9. Normalizirane jačine interakcija

Protein	Ukupna jačina (broj jačina)	Ukupan broj veza (AA)	Normalizirani broj veza	Normalizirana jačina veza	Veličina proteina (broj AA)
3bn1	299,048	448	0,322%	0,00215	373
3hmb	84,984	122	0,495%	0,00345	157
1ecp	472,698	869	1,534%	0,00835	238
3hmv	27,985	118	0,083%	0,00020	378
2qs9	5,241	15	0,040%	0,00014	194
3n55	98,707	100	0,620%	0,00612	127
3i3b	1199,946	1768	0,169%	0,00115	1023
3knl	1301,317	2324	0,014%	0,00075	4204
3if2	291,858	400	0,203%	0,00150	444

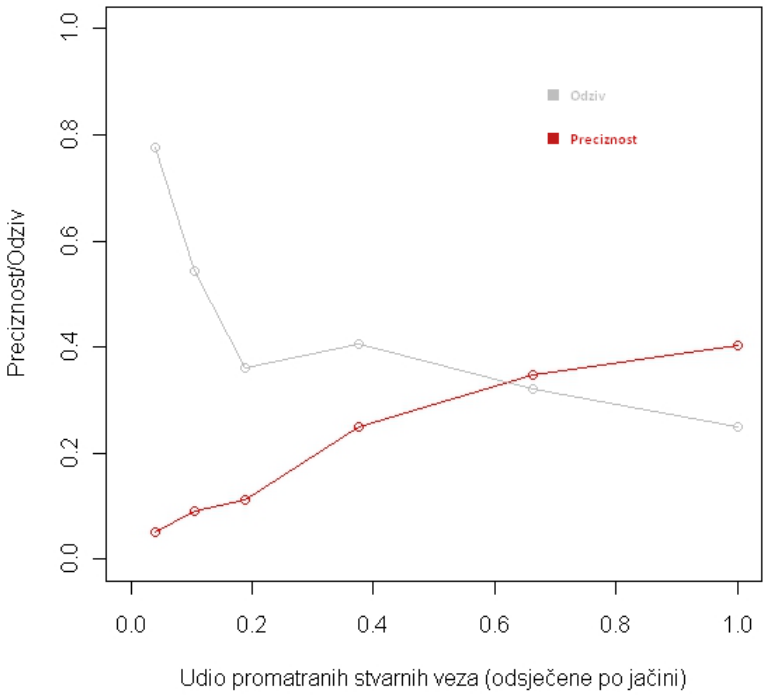
Normalizacijom postaje vidljivo kako proteini s apsolutno najvećom jačinom su zapravo među najslabije povezanima u odnosu na svoju veličinu. Broj pronađenih interakcijskih parova u postocima se kreće oko 0,1% što znači da je prosječno na 1000 interakcijskih parova pronađena 1 veza.

4.2. Analiza preciznosti i odziva sustava

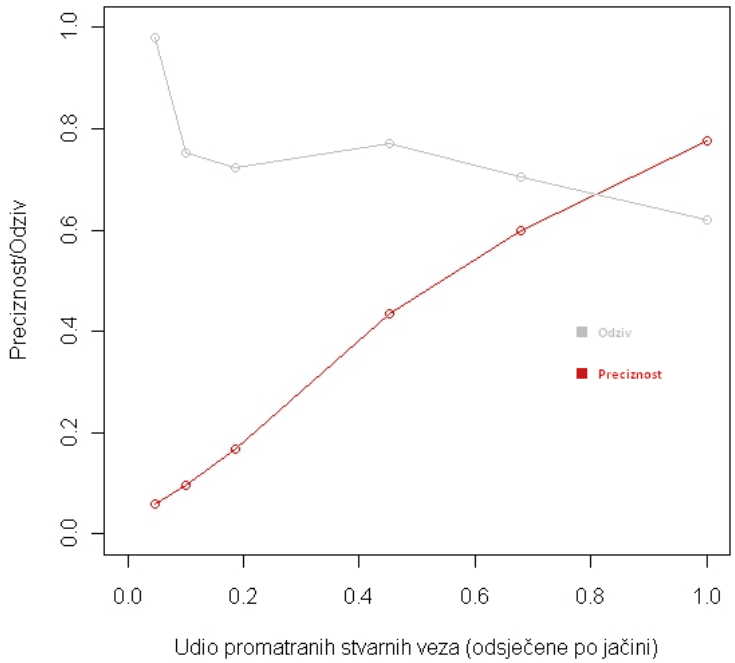
Ponašanje sustava opisano preciznošću i odzivom ispitano je na testnom proteinu. Riječ je o proteinu sa 6 lanaca (heksamer) i 1422 aminokiseline. Interakcija su pretraživane prema modelu za koji je u prethodnom poglavlju analiziran kvalitativno i kvantitativno.

Na sljedećim slikama prikazano je ponašanje preciznosti i odziva sustava u odnosu na promatrani skup stvarnih interakcija. Pritom se mogu

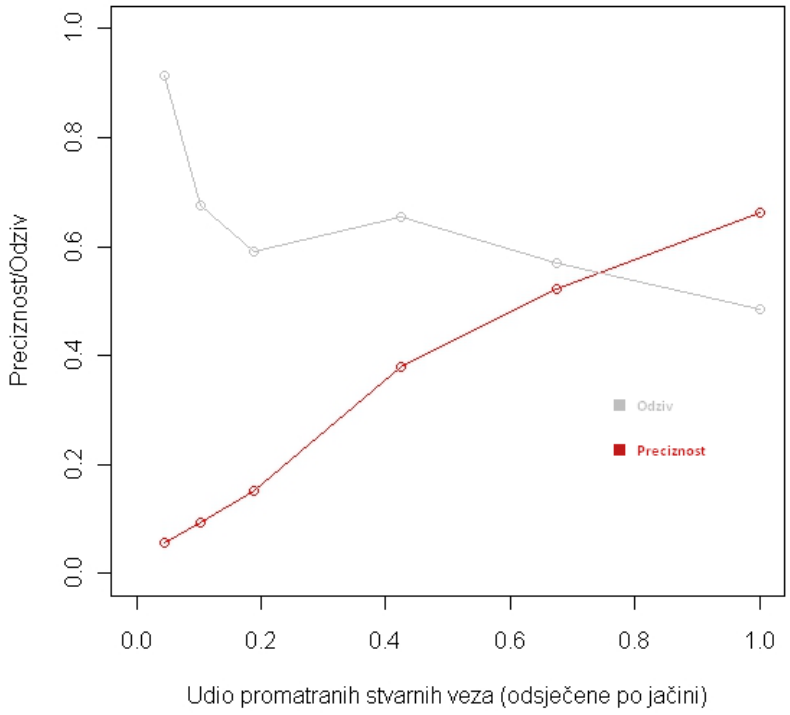
promatrati sve stvarne interakcije zabilježene u proteinu ili određeni dio najsnažnijih, npr. polovica, četvrtina, desetina najsnažnijih interakcija, itd. Alternativno je bilo moguće koristiti apsolutni prag na jačine interakcija, ali budući da se jačine definiraju proizvoljno takav prag ne bi imao stabilno, konstantno značenje. U sljedećim primjerima za određivanje interakcija nisu korištene provjere ASA-e i VDW radijusa PSAIA funkcijama.



Slika 10. Ponašanje preciznosti i odziva sustava za nepovoljne interakcije u odnosu na različite veličine skupa najjačih stvarnih interakcija



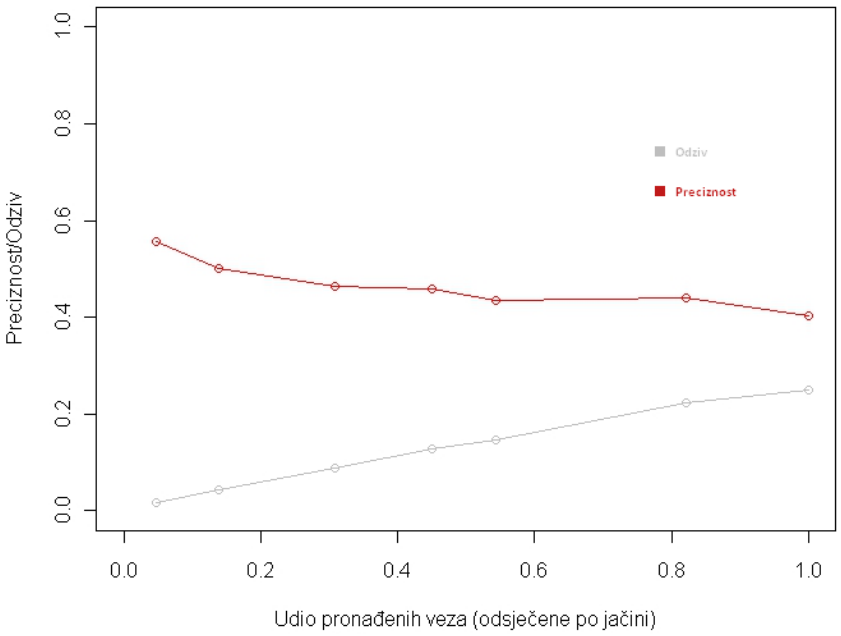
Slika 11. Ponašanje preciznosti i odziva sustava za povoljne interakcije u odnosu na različite veličine skupa najjačih stvarnih interakcija



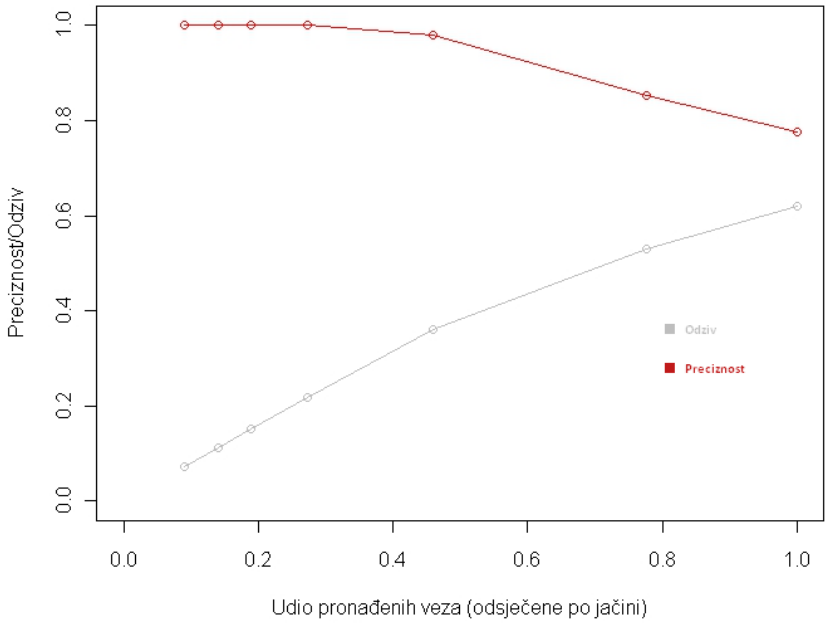
Slika 12. Ponašanje preciznosti i odziva sustava za sve interakcije u odnosu na različite veličine skupa najjačih stvarnih interakcija

Na slikama je vidljivo povoljnije ponašanje, odnosno veći iznosi preciznosti i odziva za povoljne nego za nepovoljne interakcije. To je povezano s oskudnijim definicijama nepovoljnih interakcije koje se definiraju samo kao negativne polarne veze. Za identifikaciju povoljnih interakcija, osim pozitivnih polarnih veza, koriste se vodikove, pi, VDW i hidrofobne interakcije.

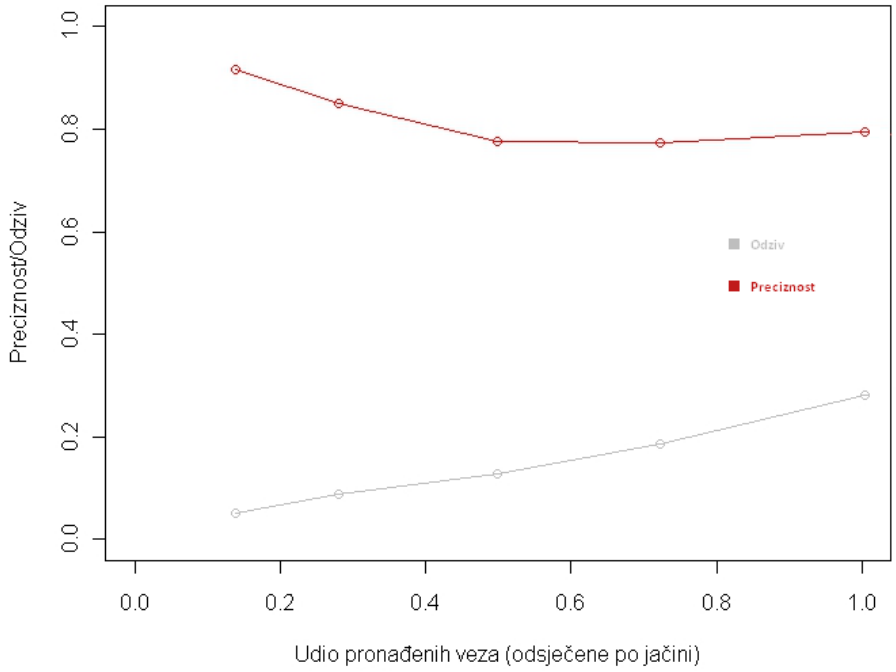
Na sljedećim primjerima ispitivanje preciznosti i odziva izvedeno je u odnosu na skup najboljih pronađenih interakcija, tj. prilikom ispitivanja odziva i preciznosti sustava, nisu uzete u obzir sve detektirane interakcije, već samo određeni dio najjačih prema ocjeni sustava.



Slika 13. Ponašanje preciznosti i odziva sustava za nepovoljne interakcije u odnosu na različite veličine skupa najjačih detektiranih interakcija



Slika 14. Ponašanje preciznosti i odziva sustava za povoljne interakcije u odnosu na različite veličine skupa najjačih detektiranih interakcija



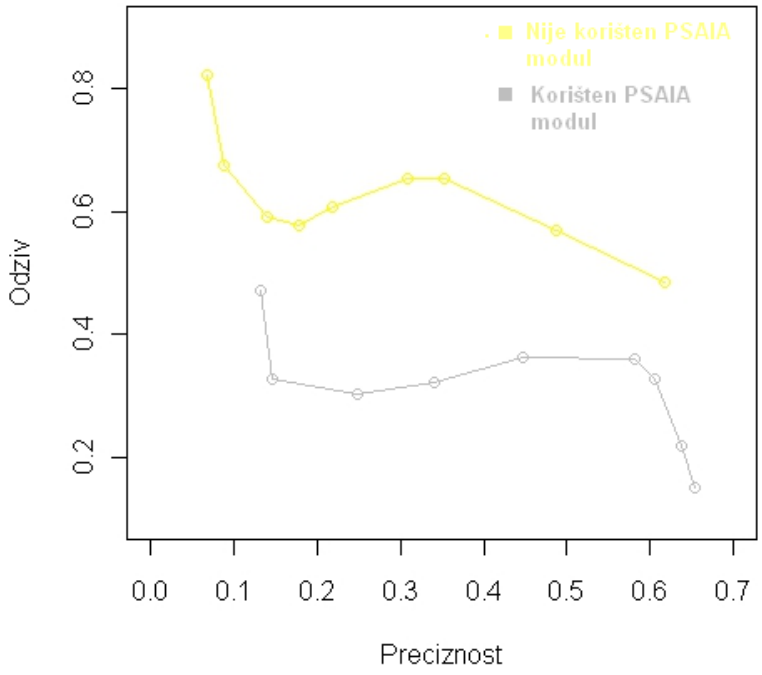
Slika 15. Ponašanje preciznosti i odziva sustava za sve interakcije u odnosu na različite veličine skupa najjačih detektiranih interakcija

Rezultati pokazuju kako odabir detektiranih interakcija prema jačini, odnosno zanemarivanje slabijih detektiranih interakcija također utječe na ponašanje sustava na sličan način kao kad se sustav ispituje u odnosu na jači dio stvarnih interakcija. Razlika je u tome što zanemarivanje većeg dijela detektiranih interakcija smanjuje odziv, za razliku od smanjivanja većeg dijela postojećih interakcija koje povećava odziv sustava. Ignoriranje slabijih detektiranih interakcija povećava preciznost jer je veća vjerojatnost da su slabije pronađene interakcije lažno detektirane. Kod smanjivanja skupa promatranih stvarnih interakcije efekt je suprotan jer se iz „vidnog polja“ gube stvarne interakcije. Pritom sustavu koji ih detektira stvaraju naizgled grešku u procjeni, odnosno smanjuju preciznost.

Povoljne se interakcije općenito bolje detektiraju, s većom preciznošću i većim odzivom u odnosu na nepovoljne. Rezultat je sličan bez obzira na to koji se skup varira prema jačini (pronađene ili stvarne interakcije).

Na sljedećoj slici dana je usporedba ponašanja odziva sustava u odnosu na preciznost kad se koriste PSAIA provjere i kad se ne koriste PSAIA

provjere. Graf je dobiven varijacijom skupa stvarnih interakcija prema jačini.



Slika 16. Usporedba grafa preciznost/odziv uz korištenje i bez korištenja modula PSAIA

Iz usporedbe grafova vidljiv je bolji odziv sustava kad se ne koriste PSAIA provjere. Budući da PSAIA eliminira dio kandidata za hidrofobne i VDW veze daje bolje preciznosti jer je manja pogreška. S druge strane kandidati za hidrofobne i VDW interakcije široko su definirani:

- VDW – između bilo koja dva atom,
- hidrofobne – između bilo koja dva C atoma

pa se na malim udaljenostima koje se uzimaju u obzir (eksplicitno postavljeni prag sustava na udaljenost atoma, u ovom slučaju 10 angstrema) može raditi o nekoj drugoj vrsti interakcije koja nije uzeta u obzir u modelu kao poseban tip interakcije.

Rezultati ispitivanja pokazuju da se prema korištenom modelu interakcija pronalazi relativno mali broj stvarnih interakcija, ako se promatra skup svih interakcija u proteinu. S druge strane velika većina pronađenih interakcija jesu stvarne interakcije. Zaključak je taj da je model interakcija prema kojima sustav pretražuje točan, ali nedovoljno opširan. Isto tako u korištenom modelu očito su definirane pretežito jake interakcije dok su

slabije zanemarene. To dovodi do poboljšanja odziva kad se veličina skupa stvarnih interakcija varira prema njihovoj jačini. Ipak, budući da model detektira i manji broj slabih interakcija, one će kod smanjivanja skupa stvarnih interakcija, dovesti do lošije ocjene preciznosti. Model interakcija potrebno je dakle proširiti dodatnim definicijama tipova interakcija kako bi se polučili bolji rezultati.

4.3. Usporedba s rezultatima PDT-a

Usporedba ocjena konformacija između PDT alata i modula za analizu interakcija dana je na primjeru dokiranja liganda i polimera proteina 3hfl. Redni brojevi konformacija dodijeljeni su na temelju osnovnog PDT ocjenjivanja strukture koje se temelji na detektiranju neprirodnog prodiranja jednog proteinskog kompleksa u drugi u prostoru. U tablici su prikazani rezultati za prvi 10 interakcija.

Tablica 10. Usporedba s PDT rangiranjem

PDT redni broj konformacije – osnovni scoring	Rangiranje po ukupnoj jačini
1	45
2	17
3	97
4	93
5	29
6	40
7	42
8	67
9	51
10	87
…	…
100	64

Iz tablice je vidljivo kako rezultati PDT rangiranja i rangiranja na temelju ukupne jačine interakcija nisu usklađeni. PDT rangira konformacije prema (ne)postojanju neprirodnih prostornih prodiranja jednog proteina u drugi, a modul za analizu interakciju traži kemijske veze među proteinima.

Rezultati se mogu objasniti na način da PDT ocjenjuje moguće konformacije s obzirom na prostornu usklađenost formi proteina u interakciji. Modul za analizu interakcija promatra konformacije s aspekta lokalnih odnosa i kemijskih značenja konformacija pojedinačnih atoma. Na taj način unosi se dodatnu razinu uvida u stabilnost strukture.

4.4. Brzina rada modula

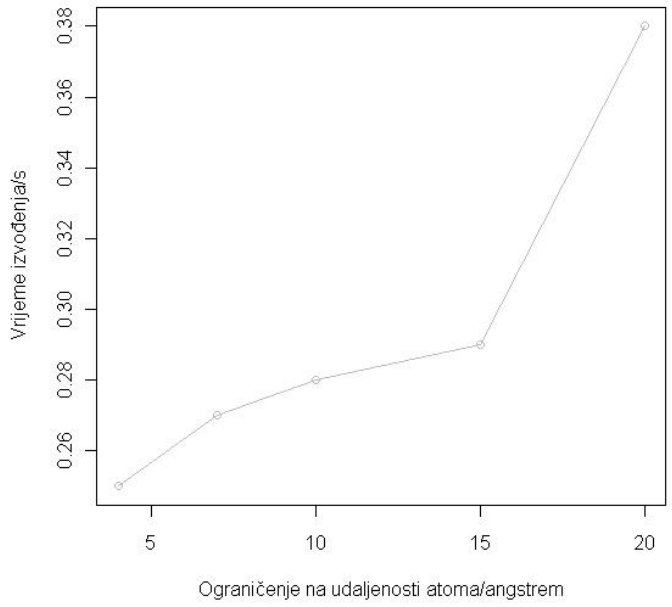
Na brzinu rada modula moguće je eksplicitno utjecati na dva načina:

- zadavanjem maksimalne udaljenosti aminokiselina i atoma koji se uzimaju u obzir za ispitivanje interakcija
- paralelizacijom

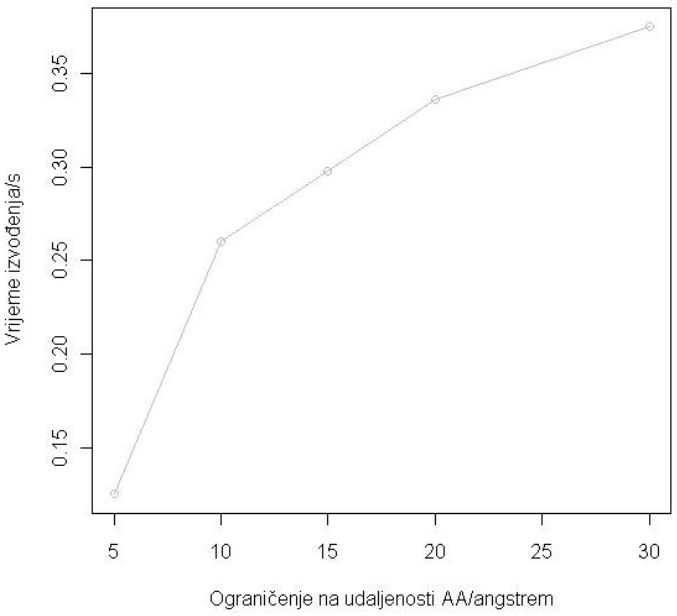
U iduća dva potpoglavlja, opisuju se rezultati optimizacije ovim dvjema metodama.

4.4.1. Podešavanje parametara

Podešavanjem parametara maksimalnih udaljenosti atoma i aminokiselina smanjuje se prostor pretraživanja interakcija. Budući da je u konfiguraciji za svaku vrstu interakcije određena maksimalna udaljenost na kojoj se ona ispituje, zadavanjem parametra udaljenosti atoma, tj. aminokiselina, koji je jednak maksimalnoj udaljenosti zadanoj za pojedinu interakciju u konfiguracijskim datotekama, neće se izgubiti ništa na pronađenim rezultatima. Rezanjem parametra ispod te udaljenosti mogući su gubici.



Slika 17. Utjecaj parametra maksimalne udaljenosti atoma na vrijeme obrade proteina 2qs9



Slika 18. Utjecaj parametra maksimalne udaljenosti aminokiselina na vrijeme obrade proteina 2qs9

Uz ograničenje 14 na udaljenost CA atoma i 7 na udaljenost svih ostalih atoma, dobiva se vrijeme izvođenja 0,23s što je ubrzanje od 1,65 u odnosu na inicijalno izmjereno vrijeme od 0,38 s.

Tablica 11. Utjecaj ograničenja na broj pronađenih veza

Ograničenje na udaljenost CA atoma/angstrem	Broj pronađenih veza
30	194
20	194
15	194
10	194
8	136
5	0

4.4.2. Paralelizacija

Već je spomenuta brzina kao važan aspekt rada modula. Budući da je sam proces traženja proteinskih interakcija složenosti k^2, raste vrlo brzo s veličinom proteina. Osim toga, za potrebe PDT alata potrebno je analizu provesti za 1000 konformacija istog proteina. Uzme li se za primjer protein 3knl koji ima 4204 aminokiselina i 26 lanaca, čija brzina obrade traje preko 300s, dobiva se vrijeme analize 1000 konformacija od 300000s, odnosno 833h.

Tablica 12. Vremena obrade ispitnih proteina

Naziv proteina	Ukupna duljina	Broj lanaca	Vrijeme obrade
3bn1	373	4	1,106
3hmb	157	3	0,229
1ecp	238	6	1,014
3hmv	378	2	0,512
2qs9	194	2	0,262
3n55	127	2	0,093
3i3b	1023	4	7,500
3knl	4204	26	301,161
3if2	444	2	0,660

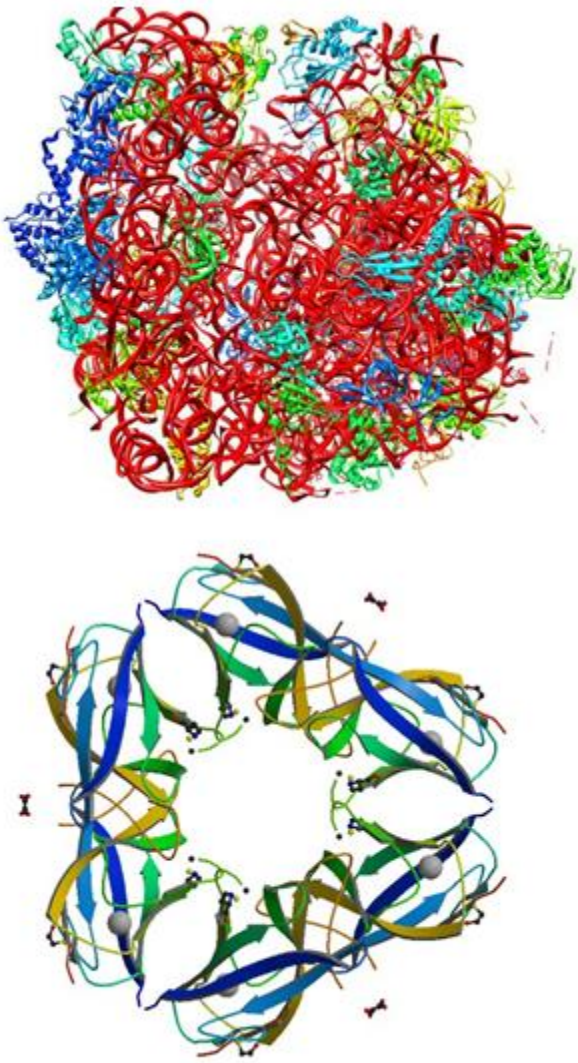
Ipak, treba uzeti u obzir kako je u ovom slučaju uspoređivana struktura cijelog proteina, odnosno 26 lanaca. To znači da je analiziran veći dio strukture proteina. Za potrebe dokiranja proteina PDT alatom, analiziraju se uvijek samo 2 lanca – prvi i drugi protein. Znači da se uspoređuje prosječno ¼ interakcijskih parova. Kod većeg broja lanaca maksimalni dio proteinskih interakcija koje se pretražuju određen je s

$$\frac{k-1}{2k}$$

Za 26 lanaca to znači gotovo polovicu proteinskih parova, odnosno dvostruko dulje vrijeme izvođenja. Bez obzira na to, vrijeme izvođenja reda veličine 10^2 sati, nije prihvatljivo. Ipak budući da se radi o kvadratnoj složenosti programa, s manjim proteinima, vrijeme analize bitno opada. Za proteine duljine 10^2 aminokiselina, vrijeme izvođenja programa je reda veličine sekunde. Za obradu 1000 konformacija proteina, čija se analiza obavi u 1s, potrebno je 1000s, odnosno 16,67 minuta. Dodatno ubrzanje postiže se paralelizacijom.

Budući da su analize zasebnih datoteka potpuno odvojeni poslovi, prigodni su za paralelizaciju. Komunikacija nije potrebna, izlazni rezultat

je malen – samo ocjena kvalitete interakcija, a računski je proces zahtjevan.



Slika 19. Vizualizacija najmanjeg i najvećeg testiranog proteina: 3knl i 3n55

Zbog jednostavnosti odabrana je paralelizacija dretvama. Svaka dretva obrađuje proporcionalni dio posla, a rezultat zapisuje u posebnu strukturu u memoriji. Rezultati dokiranja sadrže isti protein u različitim prostornim konformacijama. To znači da nema rizika od pretjeranog opterećenja jednog dijela dretvi. Na kraju se rezultati rada programa sortiraju i ispisuju u izlaznu datoteku.

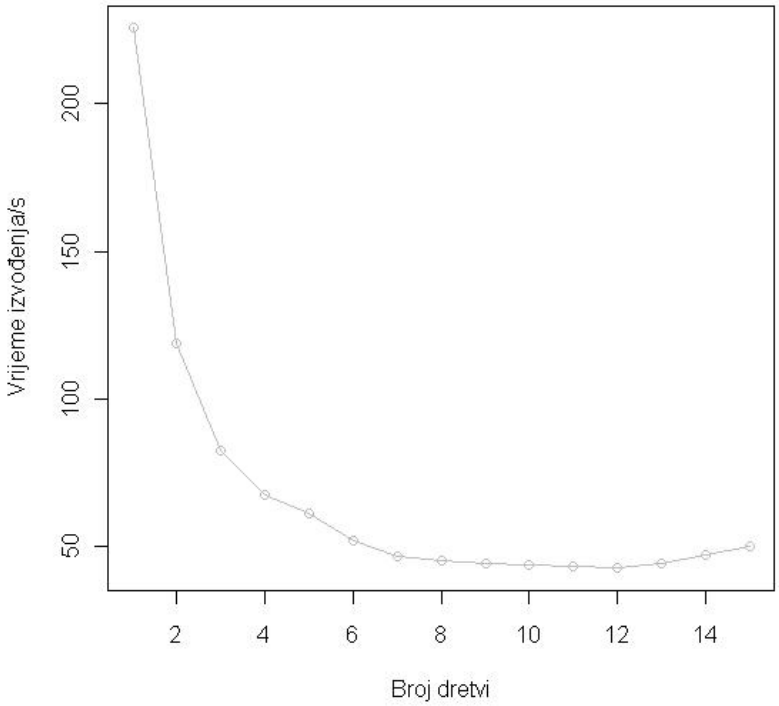
Tablica 13. Vremena izvođenja programa

Broj obrada/broj dretvi	**1**	**10**	**15**
Protein 3hmb			
1	0,229s	#	#
100	22,010s	4,883s	4,811s
1000	222,801s	33,480s	32,309s
Protein 1ecp			
1	1,014s	#	#
100	49,518s	16,130s	26,660s
1000	1020,59s	155,102s	180,217s

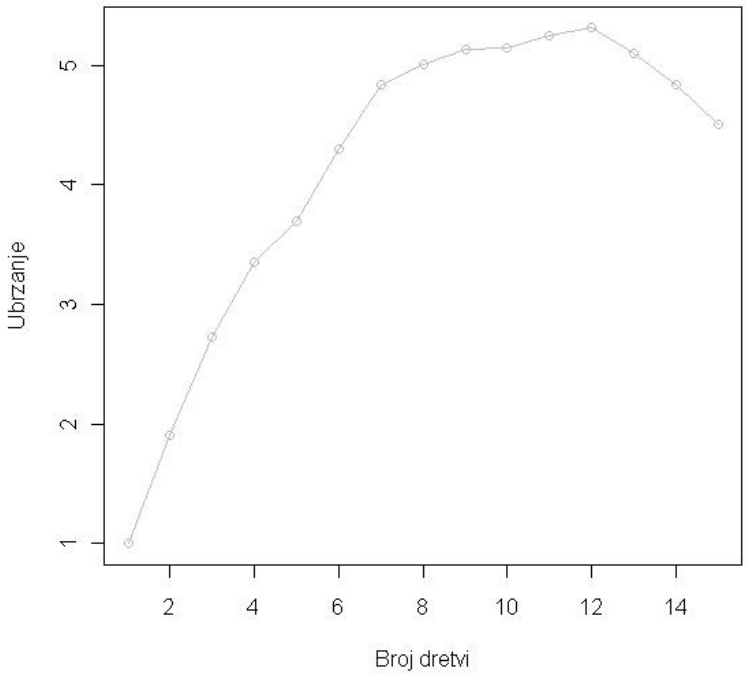
Tablica 14. Ubrzanja kod višedretvenog rada

Protein 3hmb	**Broj obrada**	**Broj procesora**	**Ubrzanje**
	100	10	4,508
	1000	10	6,655
	100	15	4,585
	1000	15	6,896
Protein 1ecp	**Broj obrada**	**Broj procesora**	**Ubrzanje**
	100	10	3,070
	1000	10	6,560
	100	15	1,860
	1000	15	5,663

Navedeni primjeri u tablicama sugeriraju značajno ubrzanje kod korištenja većeg broja dretvi. Isto tako, s više od 10 dretvi, ubrzanje se ne mijenja ili čak pada. Budući da je korišteno računalo sa 16 jezgri, vjerojatno se radi o preopterećenosti pojedinih procesora. Na sljedeća dva grafa prikazano je vrijeme izvođenja i ubrzanje (ukupno i po svakoj dretvi).



Slika 20. Vrijeme izvođenja ovisno o broju dretvi -
- 1000 obrada proteina 3hmb



Slika 21. Apsolutno i ubrzanje po dretvi -
-1000 obrada proteina 3hmb

5. Diskusija

Definicije geometrijskih uvjeta i vrsta atoma koji sudjeluju u pojedinim interakcijama određuju se proizvoljno kao ulazni podaci u sustav. Pritom je moguće modificirati vrste atoma koji tvore interakcije i udaljenosti na kojima interagiraju. Osim toga moguća su i ograničenja na pojedine kutove. Cilj ovakvih podešavanja je ostavljanje otvorenom mogućnosti poboljšavanja rezultata dodatnim modeliranjem interakcija.

Prema obavljenim ispitivanjima na testnim uzorcima i testnim postavkama, pokazuje se kako su korištene definicije interakcija točne, ali oskudne. tj. pronalazi se mali broj veza, ali i s malom pogreškom. Također, pronalaze se pretežito jače interakcije. Rezultati se mogu poboljšati proširenjem definicija interakcija i ugađanjem numeričkih parametara.

Složenost postupka analize proteinskih interakcija je n^2 gdje je n duljina proteina. To znači da vrijeme potrebno za obradu pojedinog proteina raste s kvadratom veličine proteina. Ako se za analizu jednog proteinskog para troši mikrosekunda, za analizu proteina s 10^3 atoma, trošit će se 1 sekunda, za analizu proteina s 10^4 atoma trošit će se 10^2 sekundi, a za protein sa 10^5 atoma, trošit će se 10^4 sekundi. Što znači da, bez obzira na tehničku izvedbu, s porastom broja atoma u proteinu, vrijeme obrade neminovo značajno raste. Kad se tome pridoda brojka od 10^3 proteina koje je potrebno analizirati, problem postaje izraženiji.

Mjere podešavanja parametara i paralelizacije koje su primijenjene kao eksplicitne metode ubrzavanja programa, maksimalno linearno doprinose ubrzanju. S obzirom na složenost algoritma, njihov doprinos gubi značaj s porastom veličine proteina.

Budući da je potrebno obraditi 10^3 proteina u razumnom vremenu, kao prihvatljivo vrijeme obrade jednog proteina uzima se 1 sekunda. Obrada 10^3 proteina u tom slučaju trajat će 16,67 minuta. Uz primjenu paralelizacije to se vrijeme može dodatno smanjiti. Pregledom rezultata vremena izvođenja, očito je da je za 7/9 proteina vrijeme obrade od 1s postignuto. Riječ je o proteinima reda veličine do 10^4 atoma. Osim eksplicitnih mjera sužavanja prostora pretraživanja, ubrzanja su

postignuta i minimalnim instanciranjem novih objekata te optimizacijom koda - prilikom dizajna i prilikom prevođenja.

6. Zaključak

Prilikom, a i prije izrade rada, kao najveći problem naglašena je brzina izvođenja. Budući da se definicije kemijskih veza velikim dijelom zadaju parametarski, moguće ih je podešavati naknadno. Taj posao može ostati na stručnjacima domene. Mogućnosti zadavanja veza pritom su ograničene za zadavanje vrste aminokiselina i atoma koji u njima sudjeluju, udaljenosti na kojima veza djeluje te do određene mjere načina na koji se ocjenjuje jakost veze.

Brzina analize predstavlja problem jer se radi o kvadratnoj složenosti koja raste brže nego što ikakva linearna ubrzanja mogu pratiti. Linearna ubrzanja u programu uvode se eksplicitnim zadavanjem maksimalne udaljenosti na kojoj se ispituju interakcije između atoma. Na taj način smanjuje se prostor pretraživanja, dakle ubrzava obrada. Kao eksplicitna metoda ubrzavanja obrade koristi se i višedretveni rad na višejezgrenom računalu.

U inačici programa koja je korištena za potrebe ovog rada, postignute su brzine izvođenja programa povoljne za proteine veličine do 10^4 atoma. Kako bi se modul mogao koristiti za veće proteina, potrebno je uvesti dodatna ubrzanja. Apsolutna primjenjivost programskog koda na ovaj problem nije moguća dok god se koristi algoritam kvadratne složenosti.

7. Reference

[1.] RCSB Protein Dana Bank, http://www.pdb.org/pdb/home/home.do, svibanj 2010.

[2.] IUPAC-IUBMB-IUPAB INTER-UNION Task Group On The Standardization Of Dana Bases Of Protein And Nucelic Acid Structures Determined By NMR Spectroscopy, Recommendations For The Presentation Of NMR Structures Of Proteins And Nucleic Acids, Pure & Appl. Chem., Vol. 70, No. 1, pp. 117-142, 1998., svibanj 2010.

[3.] Josip Mihel, Mile Šikić, Sanja Tomić, Branko Jeren, Kristian Vlahoviček, PSAIA – Protein Structure and Interaction Analyzer, BMC Struct Biol, doi: 10.1186/1472-6807-8-21., travanj 2008, svibanj 2010.

[4.] 3D Analysis::Reduce Software for Adding Hydrogens, http://kinemage.biochem.duke.edu/software/reduce.php, svibanj. 2010.

[5.] Mile Šikić, Nino Antulov-Fantulin, Igor Čanadi, Matija Piškorec, Ivan Sović, PDT – Protein Docking Tool, http://complex.zesoi.fer.hr/PDT.html, lipanj 2010.

Sažetak

Alat za prianjanje proteina: modul za utvrđivanje interakcija

U okviru rada programski je ostvaren modul za analizu proteinskih interakcija između proteina nastalih kao posljedica dokiranja dva proteina PDT programskim alatom. Prilikom analize, uzete su u obzir vodikove, pi, polarne, hidrofobne cistinske i Van der Waalsove interakcije. Također su razmatrane nepovoljne interakcije koje sugeriraju malu vjerojatnost pronađene konformacije atoma. Njima je pridružena negativna vrijednost. Programska definicija veze uključuje definicije atoma koji je mogu tvoriti, udaljenosti i kutova na kojima veza može postojati te njezine jakosti. Vrste veza koje se pretražuju definiraju se preko XML datoteka. Osim toga parametarski se zadaje i maksimalna udaljenost atomskih i aminokiselinskih parova koji će se promatrati prilikom pretraživanja interakcija.

Izlaz PDT alata oko 1000 mogućih konformacija atoma, cilj je pronaći najvjerojatnije među njima. Ukupna kvaliteta konformacije ocjenjuje se zbrojem svih pozitivnih i negativnih jakosti pronađenih veza. Kako bi se ubrzala obrada tako velikog broja datoteka korištena je paralelizacija uz pomoć dretvi. Rezultati podjelom posla na 10 dretvi doveli su do ubrzanja programa oko 5 puta. Brzini je također doprinijelo u manjoj mjeri podešavanje parametara koji smanjuju prostor pretraživanja veza.

Budući da je sam algoritam pretraživanja veza kvadratne složenosti, brzo raste s veličinom proteina. Ovakva implementacija, primjenjiva je na proteine s 10^4 atoma. Kako bi se mogla primijeniti i na veće proteina, potrebna je dodatna optimizacija. Budući da su tehnološka ubrzanja linearna ne mogu u konačnici potpuno riješiti problem kvadratne složenosti.

KLJUČNE RIJEČI:

Proteini, interakcije, PSAIA, PDT, paralelizacija, PDB, vodikove veze, polarnost, Hidrofobnost, Van der Waals, cistinski mostovi, pi interakcije

Abstract

Protein Docking Tool: Module For Interaction Detection

Protein docking is a process of merging two protein chains into a single new protein. PDT (eng. Protein Docking Tool) is a software which does protein docking on two chains. As a result, PDT gives a certain number of possible chain conformations, which represent the new protein. These output conformations were generated and scored internally in the PDT tool based only on the 3D geometry of the two input chains.

Module for interaction detection is a tool for scoring and ranking these output conformations based on geometrical and chemical properties of atoms and molecules which build the chains. Interactions are detected based on predefined geometrical conditions which approximately describe different chemical bonds between atoms and molecules. Such bonds, considered in the design of this module, are hydrogen bonds, pi bonds, polar bonds, cystein bonds and Van Der Waals bonds. A possibility to analyze hydrophobic behavior of molecules is also implemented.

If found in a result conformation, positively scored interactions suggest higher likelihood of such conformation in nature. Negative interactions are also considered. They are based on incompatible polarities of atoms found geometrically close to each other. Naturally, negative interactions suggest lower likelihood of a chain conformation.

Because of relatively large amount of conformations that have to be examined and scored in the PDT tool, and the quadratic complexity of the algorithm, time of execution is a critical property. Module has several ways of explicit speed improvement. First way is by limiting the distance between atoms which are considered as possible interaction pairs. The same limitation can be applied to molecules, or more accurately to amino acids. The module can also work in parallel mode. The parallelization is based on complete independency of processes that score individual output conformations. Therefore, the scoring can be done simultaneously and independently by different processes, on different processors.

It should be noticed that these methods for speed improvement, can contribute only on a linear level. Therefore, they do not solve the problem absolutely.

KEYWORDS:
protein, interaction, PSAIA, PDT, parallelization, PDB, hydrogen bonds, polarity, hydrophobicity, Van der Waals, cystein, pi interactions