SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Neredundatni skupovi proteina u bioinformatici

Juraj Petrović

Zagreb, svibanj 2009.

Sadržaj:

Uvod

Bioinformatika predstavlja svaku primjenu informacijskih tehnologija i računalnih znanosti u organizaciji, povezivanju, pohranjivanju, analizi, vizualizaciji složenih bioloških procesa i molekularnoj biologiji općenito [1]. Sam pojam *bioinformatika* prvi je iskoristio Paulien Hogeweg 1978. godine, a s vremenom je bioinformatika prerasla u široko interdisciplinarno područje. Prema definiciji američkog National Centera for Biotechnology Information [2], postoje tri značajne poddiscipline unutar bioinformatike i to su: razvoj novih algoritama i statistike kojima se određuju veze između članova velikog skupa podataka, analiza i interpretacija različitih tipova podataka uključujući nukleotidne i aminokiselinske sljedove, proteinske domene i strukture proteina, te razvoj i implementacija alata koji omogućuju učinkovit pristup različitim tipovima podataka kao i njihovo učinkovito upravljanje. Primarni je cilj bioinformatike, dakle, objašnjavanje složenih bioloških procesa.

Bioinformatika se između ostalog bavi i predviđanjem strukture proteina (eng. *protein structure prediction*) i predviđanjem proteinskih interakcija (eng. *protein-protein interactions*), koji spadaju u drugu navedenu poddisciplinu bioinformatike. Jedan od ključnih elemenata ovakvih radova svakako je skup podataka koji se obrađuju ili koji se koriste za provjeru točnosti neke metode. Podaci su proteinske strukture prikazane u obliku prihvatljivom računalu. Ovakvi podaci besplatno su dostupni u više baza na internetu, a broj struktura u tim bazama svakodnevno raste. Jedna od najpoznatijih takvih baza svakako je RSCB PDB baza [3], koja je osnovana 1971. godine sa samo sedam, a danas broji preko 57000 proteinskih struktura.

Sve ove strukture, dakako, nisu jednako pogodne za sva istraživanja. Razlikuju se ne samo po konkretnoj proteinskoj strukturi koju opisuju, nego i po tehnici kojom se do tog opisa došlo i rezolucijom odnosno kvalitetom tog opisa i brojnim drugim parametrima. Neki od parametara utječu na to koje će strukture biti odabrane kao dio skupa podataka za neko istraživanje. Vrlo je bitno promatrati i međusobne odnose odabranih struktura, odnosno sličnosti u njihovoj primarnoj strukturi odnosno slijedu aminokiselinskih ostataka. Ovo svojstvo naziva se redundancija ili zalihost. Izvor potrebe za njenim uklanjanjem iz skupova podataka leži u činjenici da je obično u radovima s ovog područja cilj provjera neke teze ili metode na raznolikom skupu podataka, jer bi očekivano, slični podaci trebali davati slične rezultate koji možda nisu opće vrijedeći. Cilj ovog seminarskog rada je dati pregled i analizu dosad najčešće korištenih neredundantnih skupova proteina u bioinfomrmatici, te kreirati vlastiti skup i proanalizirati njegova svojstva.

1. Postojeći skupovi

Generalno govoreći, za znanstvenike redundancija obično predstavlja prednost, jer je promatranjem sličnih struktura obično lakše uočiti funkcionalne ili strukturalne posljedice promjene manjeg djela strukture. To je i razlog zašto se najčešće analiziraju strukture relativno slične već istraženima. Slične strukture također kristaliziraju u sličnim uvjetima, što je vrlo korisna spoznaja pri određivanju struktura tehnikom rendgenske kristalografije (eng. *X-ray crystallography*), zasada najtočnije metode u određivanju molekularnih struktura. Ipak, u skupovima koji se koriste u radovima iz područja bioinformatike redundancija u primarnoj strukturi obično se nastoji izbjeći. Metodu predikcije strukture ili neko svojstvo koje se pokušava dokazati potrebno je ispitati na što raznolikijem testnom skupu. Neki od najčešće korištenih neredundantni testni skupovi proanalizirani su u nastavku rada.

1.1. Ofran i Rost

U svom radu iz 2003. godine, "Predicted protein-protein interaction sites from local sequence information" [4] Ofran i Rost su za predviđanje mjesta interakcije na proteinu koristili testni skup od 1137 lanca iz 333 PDB strukture. Strukture su bile odabrane između tada dostupnih u RCSP PDB baze i to one strukture koje imaju rezoluciju manju od 4Å i bar dva lanca s po najmanje 30 aminokiselinskih ostataka. Za lance iz ovog skupa također vrijedi da tvore tranzijentne interakcije, no to nam u ovom slučaju nije predmet interesa.

Analiza strukturalne sličnosti lanaca iz ovog skupa dala je sljedeće rezultate: za 28.92% od ukupno 408 clustera 30% pokrivenih ovim skupom postoje bar dvije PDB strukture čiji lanci su u tom clusteru zastupljeni. Isto vrijedi i za 23.95% od ukupno 522 zastupljena clustera 50%, 20.80% od ukupno 572 zastupljena clustera 70% i za 19.62% od ukupno 627 zastupljenih clustera 90% pokrivenih ovim skupom. Ovi rezultati pokazuju prisutnost redundancije u primarnoj strukturi. U idealnom slučaju u svakom clusteru bio bi prisutan samo po jedan lanac odnosno lanci iz samo jedne PDB strukture. U ovom slučaju, primjerice, u prvom clusteru 30% prisutno je bar jedan lanac iz čak 121 PDB strukture. Potpuni rezultati analize prikazani su u tablicama 1.1, 1.2, 1.3 i 1.4. Datoteke sa listama lanaca koji pripadaju određenom clusteru preuzete su sa izvora [3].

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	290	17. 1G6V_A
2	83	6. 1FDL_Y; 1MLC_E; 1MLC_F
3	25	7. 1VIW_B; 1LGB_A; 1LOD_A; 1LGB_B; 1LOD_B; ...
4	4	54. 1TMF_2; 1RVF_2; 1QGC_2; 1QQP_2; 1RVF_4
5	3	57. 1BUN_B; 1D0D_B; 1EJM_B; 1EJM_D; 1EJM_F; ...
6	1	182. 1AZS_C; 1CUL_C; 1AGR_A; 1AGR_D; 1GP2_A; ...
24	1	2. 1AVG_H; 1E0F_D; 1E0F_E;...
121	1	1. 1AON_A; 1AUI_A; 1BMF_A;...

Tablica 1.1: Cluster 30%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	397	17. 1FM6_A; 1FM6_D
2	90	8. 1FDL_Y; 1MLC_E; 1MLC_F
3	22	10. 1VIW_B; 1LGB_A; 1LOD_A; 1LGB_B; 1LOD_B; 1LOD_D
4	7	11. 1QKZ_A; 1EFN_A; 1FC2_C; 1DEE_G; 1DEE_H
5	2	90. 1C1Y_A; 1GUA_A; 1LFD_B; 1BKD_R; 1HE8_B
6	1	5. 1AUI_A; 1TCO_A; 1A2X_A; 1WDC_C; 1BR4_B; 1G4Y_R
9	1	83. 1E96_A; 1I4T_D; 1CC0_A; 1G4U_R; 1CXZ_A; 1AM4_D; 1FOE_B; 1TX4_B; 1HE1_C
23	1	2. 1AVG_H; 1E0F_D; 1E0F_E; 1E0F_F; 1TBR_H;...
41	1	1. 1QGC_4; 1FG2_A; 1FG2_D; 1FG2_G;...

Tablica 1.2: Cluster 50%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	453	8. 1BAB_B
2	91	26. 1RCX_L; 8RUC_A;
3	15	15. 1YVN_A; 2BTF_A; 1ATN_A
4	8	21. 1BUH_A; 1F5Q_A; 1FQ1_B; 1JSU_A
5	1	566. 1A2K_C; 1A2K_D; 1A2K_E; 1I2M_A; 1IBR_A; 1QBK_C; 1RRP_A; 1RRP_C
6	1	123. 1E96_A; 1I4T_D; 1G4U_R; 1AM4_D; 1FOE_B; 1HE1_C
8	1	10. 1EZX_C; 1AN1_E; 1AVW_A; 1AVX_A; 1C9T_A; 1C9T_B; 1C9T_C; 1C9T_D; 1C9T_E; 1C9T_F; 1EJM_A; 1SBW_A; 1SLU_B
41	1	1. 1QGC_4; 1FG2_A; 1FG2_D; 1FG2_G; 1FG2_J; 1EFX_A;...

Tablica 1.3: Cluster 70%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	504	7. 1BAB_B
2	97	8. 1FDL_Y; 1MLC_E; 1MLC_F
3	16	6. 1AUI_A; 1TCO_A; 1G4Y_R
4	5	19. 1BUH_A; 1F5Q_A; 1FQ1_B; 1JSU_A
5	3	348. 1E96_A; 1I4T_D; 1G4U_R; 1FOE_B; 1HE1_C
7	1	11. 1EZX_C; 1AN1_E; 1AVW_A; 1AVX_A; 1C9T_A; 1C9T_B; 1C9T_C; 1C9T_D; 1C9T_E; 1C9T_F; 1EJM_A; 1SBW_A
19	1	1. 1QGC_4; 1GC1_H; 1FC2_D; 1OAK_H; 1QFU_H;...

Tablica 1.4: Cluster 90%

Pogledajmo još samo što ovi rezultati zapravo znače na primjeru. Prema tablici 1.4, u 97 clustera 90% nalaze se lanci iz bar 2 različite PDB strukture. Jedan od tih 97 clustera je i 8. po redu cluster 90% koji iz promatranog skupa sadrži lance 1FDL_Y, 1MLC_E i 1MLC_F iz PDB struktura 1FDL i 1MLC. To što se lanci nalaze u istom clusteru 90% znači da je njihova sličnost bar 90%. Promotrimo aminokiselinski sastav lanaca 1FDL_Y i 1MLC_F:

1FDL_Y:

```
SEQRES   1 Y  129  LYS VAL PHE GLY ARG CYS GLU LEU ALA ALA ALA MET LYS
SEQRES   2 Y  129  ARG HIS GLY LEU ASP ASN TYR ARG GLY TYR SER LEU GLY
SEQRES   3 Y  129  ASN TRP VAL CYS ALA ALA LYS PHE GLU SER ASN PHE ASN
SEQRES   4 Y  129  THR GLN ALA THR ASN ARG ASN THR ASP GLY SER THR ASP
SEQRES   5 Y  129  TYR GLY ILE LEU GLN ILE ASN SER ARG TRP TRP CYS ASN
SEQRES   6 Y  129  ASP GLY ARG THR PRO GLY SER ARG ASN LEU CYS ASN ILE
SEQRES   7 Y  129  PRO CYS SER ALA LEU LEU SER SER ASP ILE THR ALA SER
SEQRES   8 Y  129  VAL ASN CYS ALA LYS LYS ILE VAL SER ASP GLY ASN GLY
SEQRES   9 Y  129  MET ASN ALA TRP VAL ALA TRP ARG ASN ARG CYS LYS GLY
SEQRES  10 Y  129  THR ASP VAL GLN ALA TRP ILE ARG GLY CYS ARG LEU
```

1MLC_F:

```
SEQRES   1 E  129  LYS VAL PHE GLY ARG CYS GLU LEU ALA ALA ALA MET LYS
SEQRES   2 E  129  ARG HIS GLY LEU ASP ASN TYR ARG GLY TYR SER LEU GLY
SEQRES   3 E  129  ASN TRP VAL CYS ALA ALA LYS PHE GLU SER ASN PHE ASN
SEQRES   4 E  129  THR GLN ALA THR ASN ARG ASN THR ASP GLY SER THR ASP
SEQRES   5 E  129  TYR GLY ILE LEU GLN ILE ASN SER ARG TRP TRP CYS ASN
SEQRES   6 E  129  ASP GLY ARG THR PRO GLY SER ARG ASN LEU CYS ASN ILE
SEQRES   7 E  129  PRO CYS SER ALA LEU LEU SER SER ASP ILE THR ALA SER
SEQRES   8 E  129  VAL ASN CYS ALA LYS LYS ILE VAL SER ASP GLY ASN GLY
SEQRES   9 E  129  MET ASN ALA TRP VAL ALA TRP ARG ASN ARG CYS LYS GLY
SEQRES  10 E  129  THR ASP VAL GLN ALA TRP ILE ARG GLY CYS ARG LEU
```

Detaljnim uspoređivanjem utvrđujemo da su lanci slični ne samo 90% nego 100%, odnosno da su identični. Uvrštavanjem obije ove PDB strukture odnosno oba lanca u skup podataka dobili smo strukturnu redundanciju.

1.2. Rost i Sander

Ovaj skup, korišten za ispitivanje i predviđanje površine dostupne otapalu oko proteina korišten u radu "Conservation and prediction of solvent accessibility in protein families" [5] sadrži 129 lanaca iz 119 PDB struktura. Detaljnom analizom lanaca ovog skupa utvrđeno je postojanje relativno male strukturne redundancije. Ispitivanje sličnosti pomoću clustera 30% utvrđeno je da je u prvom clusteru 30% prisutan po jedan lanac iz 16 PDB struktura, u 15. clusteru 30% nalazi se po jedan lanac iz 3 PDB strukture i u 6 clustera 30% zastupljen je po jedan lanac iz dvije različite PDB strukture. Ako lance uspoređujemo tražeći sličnost od bar 50%, tada ćemo u istom clusteru naći najviše dva lanca iz dvije različite PDB strukture, a broj takvih clustera konačno će se svesti na nulu povećanjem zahtijevane sličnosti lanaca na 90%. Detaljni rezultati ispitivanja navedeni su u tablicama 1.5, 1.6, 1.7 i 1.8.

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	92	3. 2LHB_A
2	6	1246. 1MRT_A; 2MHU_A
3	1	15. 1AZU_A; 1PAZ_A; 2PCY_A
16	1	1. 2GLS_A; 8ADH_A; 2AK3_A; 1MCP_L; 1FDL_H; 1GP1_A;...

Tablica 1.5: Cluster 30%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	115	4. 5HVP_A
2	4	1. 1MCP_L; 1FDL_H

Tablica 1.6: Cluster 50%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer
1	119	2. 5HVP_A
2	2	1. 1MCP_L; 1FDL_H

Tablica 1.7: Cluster 70%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	123	1. 1FDL_H

Tablica 1.8: Cluster 90%

1.3. Manesh et al.

Manesh i suradnici kreirali su ovaj skup od 217 lanaca iz 215 PDB struktura u radu “Predicting of protein surface accessibility with information theory”[6]. Prema riječima autora, lanci iz skupa ne bi smjeli imati više od 25% sličnosti u primarnoj strukturi, biti kraći od 50 aminokiselinskih ostataka ili imati rezoluciju veću od 2.5Å. Ispitivanje skupa ipak je dokazalo prisutnost ne tako velike redundancije u primarnoj strukturi. U sedam pokrivenih clustera 30% nalaze se lanci iz bar dvije PDB strukture, a samo jedan zastupljeni cluster ima dva lanca iz različitih PDB struktura koji su slični barem 90%. Detaljni rezultati prikazani su u tablicama 1.9, 1.10, 1.11 i 1.12.

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	170	5. 5PTP_A
2	6	3. 1HLB_A; 1MBD_A
34	1	1. 1BNC_A; 1AFW_A; 2NAC_A; 1SFT_B; 3MDD_A; 1UBY_A; 1CMK_E; 1DXY_A; 1FJM_A; 1ONR_A; 1IRK_A; 1CSN_A...

Tablica 1.9: Cluster 30%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	212	1. 1HSB_A
2	2	320. 1BKS_A; 2TYS_A

Tablica 1.10: Cluster 50%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	214	1. 1HSB_A
2	1	263. 1BKS_A; 2TYS_A

Tablica 1.11: Cluster 70%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	214	5. 119L_A
2	1	305. 1BKS_A; 2TYS_A

Tablica 1.12: Cluster 90%

1.4. Carugo

Ovaj skup koji sadrži 338 proteinskih lanaca iz 338 različitih PDB struktura također je korišten za predviđanje slobodne površine proteina dostupne otapalu i to u radu “Predicting residue solvent accessibility from protein sequence by considering the sequence environment” [7]. Prema riječima autora, skup sadrži samo lance sa sličnosti aminokiselinskog slijeda ne većom od 25% i rezolucijom manjom od 2.5Å. Ispitivanje je pokazalo da je blaža redundancija u vidu sličnosti sljedova koji tvore lance ipak prisutna, iako nikada nije veća od 90%. Detaljni rezultati prikazani su u tablicama 1.13, 1.14, 1.15 i 1.16.

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	276	2. 1A7S_A
2	6	3. 1HLB_A; 1A6M_A
50	1	1. 1A8D_A; 1GSO_A; 1ANF_A; 1KVS_A; 1DXY_A; 1BG2_A; 1PTY_A; 1ADS_A; 1RKD_A; 1IRK_A; 1CSN_A ...

Tablica 1.13: Cluster 30%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	336	1. 1G3P_A
2	1	80. 1ANF_A; 1FNA_A

Tablica 1.14: Cluster 50%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	336	1. 1G3P_A
2	1	46. 1ANF_A; 1FNA_A

Tablica 1.15: Cluster 70%

Broj različitih PDB struktura čiji lanci se nalaze u istom clusteru:	Broj takvih clustera:	Primjer:
1	338	5. 119L_A

Tablica 1.16: Cluster 90%

1.5. Cuff i Barton

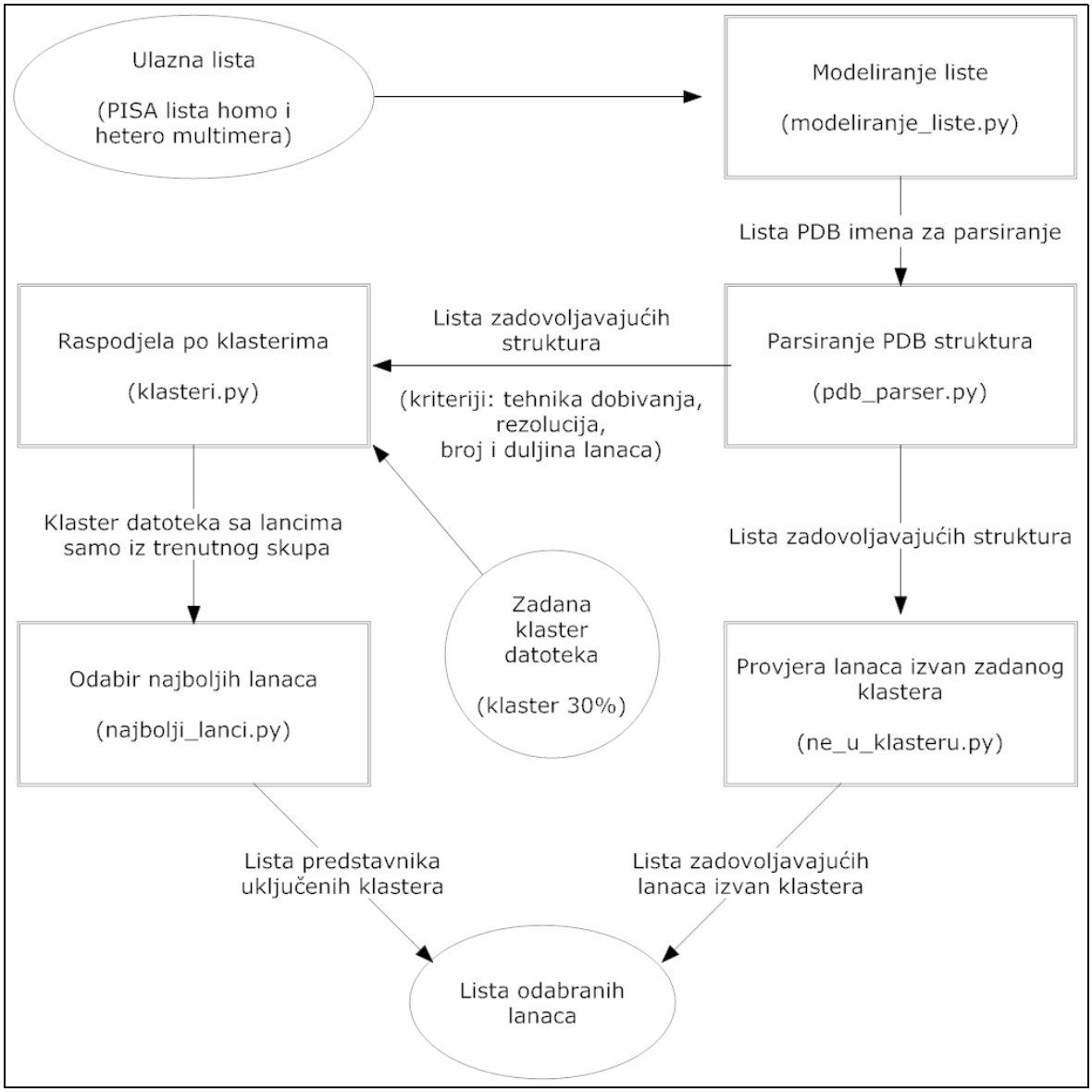
Ovaj skup, korišten u radu "Application of multiple sequence alignment profiles to improve protein secondary structure prediction" [8] sastoji se od 513 segmenta lanaca iz PDB struktura. S obzirom na to da lanci nisu korišteni cijeli nego samo njihovi dijelovi, analizu ovakvog skupa nije moguće provesti na temelju već gotovih cluster datoteka. Ipak, međusobna usporedba svih 513 struktura pokazala je kako se radi o skupu s vrlo niskom prisutnošću redundancije u obliku primarne strukture. Niti jedan lanac (odnosno segment lanca) nije po aminokiselinskom slijedu sličniji niti jednom drugom lancu iz skupa više od 30%. Štoviše, traženjem sličnosti od barem 20% dobiven je rezultat da se ona javlja samo u tri slučaja, gdje su sva tri uzrokovana duljinom lanca. U skupu su, naime, prisutna dva lanca duljine 21 aminokiselinski ostatak (1edn-1-AS, 1bpha-1-DOMAK) i jedan lanac duljine 20 aminokiselinskih ostataka (1atpi-1-DOMAK). Sličnost od bar 20% u tom slučaju znači da se u nekom drugom članu skupa pojavljuje isti 4 odnosno 5-člani uređeni podskup iz sekvence koja čini neki od navedenih lanaca. S obzirom da u skupu postoje i relativno dugi segmenti lanaca, vjerojatnost da će se neki takav podskup ponoviti i nije tako mala. Još smanjiti minimalan postotak sličnosti bilo bi besmisleno, jer bi tada pretraživali sličnosti već i u tročlanim ili dvočlanim podskupovima lanaca, a takvih sličnosti bi s obzirom na ograničeni broj mogućih kombinacija zasigurno bilo mnogo. Clusteri segmenata lanaca koji su slični bar 20% prikazani su u tablici 1.17. Svi ostali clusteri sadrže od samo jedan lanac.

1	1atpi-1-DOMAK; 1vnc-1-JAC; 2ebn-1-AS
2	1bpha-1-DOMAK; 7rsa
3	1edn-1-AS; 1mdam-1-DOMAK

Tablica 1.17: Clusteri 20% s više od jednog lanca

2. Izrada novog skupa

Budući da se RSCB PDB baza, izvor proteinskih opisa koji su glavni izvor informacija za izradu neredundantnog skupa često nadopunjava novim strukturama i poboljšanim opisima starih, stvorila se potreba za automatiziranjem postupka izrade takvog skupa. Složenost ponovne izrade skupa čak i za drugačije parametre time je svedena na minimum. Za potrebe ovog rada, postupak sam implementirao u programskom jeziku Python u obliku nekoliko skripti koje se sljedno pokreću. Graf cjelog postupka dobivanja neredundantnog skupa prikazan je na slici 2.1. Svaki od procesa kratko je opisan u nastavku rada.



Slika 2.1: Dijagram postupka dobivanja neredundantnog skupa

2.1. Početna lista

Podaci potrebni za dobivanje skupa su u prvom redu PDB baza, dostupna za preuzimanje [3], te lista struktura među kojima se biraju lanci za skup. U ovom radu, kao početna lista odabrana je lista homo i u drugom slučaju lista hetero biološki funkcionalnih multimera. Ova lista dobivena je s PISA Web servera dostupnog na adresi [9].

2.2. Modeliranje liste

Modeliranje liste drugi je korak pri dobivanju neredundantnog skupa. U idealnom slučaju ovaj korak bio bi suvišan, ali potreba za njim realno proizlazi iz više razloga. Najbitniji od njih je da je lista preuzeta s PISA Web servera zapravo lista pohranjena u bazi podataka koja može, ali ne mora biti usklađena s trenutnim sadržajem RSCB PDB baze. S vremenom, neke strukture se prestaju koristiti (eng. *obsoleted*) ili neke dobivaju nova imena. Isto može vrijediti za bilo koju početno zadanu listu. Modeliranje liste stoga znači provjeru postojanja svih struktura iz liste u lokalnoj kopiji PDB baze. Ukoliko neka struktura ne može biti pronađena u lokalnoj kopiji baze, skripta će web serveru RSCB PDB uputiti upit o imenu nedostupne strukture. Odgovor servera nosit će informaciju o tome da li je struktura u međuvremenu možda promijenila ime, više se ne koristi ili je još uvijek dostupna pod istim imenom, samo je iz nekog razloga nema u lokalnoj kopiji baze. Rezultat izvršavanja skripte je lista imena PDB struktura usklađena glede promijenjenih imena i više ne korištenih struktura.

2.3. Parsiranje PDB struktura

Jednom kada imamo listu PDB struktura koje su u lokalnoj bazi odnosno dostupne za provjeru, možemo započeti sa ispitivanjem željenih svojstava među strukturama. U ovom slučaju, željena svojstva predstavljaju ograničenje glede tehnike kojom je struktura određena (nepoželjni su teoretski modeli i strukture određene NMR tehnikom). Informacija o tehnici kojom je struktura određena nalazi se u "EXPDTA" retku PDB datoteke. Idući uvjet na proteinske lance odnosno strukture jest rezolucija. Rezolucija predstavlja preciznost određivanja strukture i mjeri se u Angstremima. Manja rezolucija znači kvalitetniju odnosno precizniju strukturu. U ovom slučaju, željena rezolucija je manja ili jednaka 3.5Å. Informacija o rezoluciji nalazi se u "REMARK 2 RESOLUTION" retku PDB datoteke. Posljednji uvjet koji je postavljen na strukture je onaj o postojanju barem dva lanca sačinjena od barem po 30 aminokiselinskih ostataka. Informacije o lancima nalaze se u "SEQRES" odjeljku PDB datoteke. Izlaz skripte koja obavlja PDB parsiranje je lista PDB struktura koje zadovoljavaju navedena svojstva.

2.4. Raspodjela po clusterima

Sljedeći korak u dobivanju skupa je uočavanje i potom uklanjanje strukturne redundancije među lancima iz parsiranjem dobivene liste struktura. U ovom koraku se izlučuju lanci iz dosad zadovoljavajućih struktura i sortiraju se u clustere prema željenom maksimalnom stupnju sličnosti koji u ovom slučaju iznosi 30%. Rezultat ovog koraka je tekstualna datoteka po strukturi identična zadanoj cluster datoteci, ali samo s vrijednostima lanaca koji su bili sadržani u PDB strukturama ulazne liste.

2.5. Odabir najboljih lanaca

Ovo je pretposljednji korak pri dobivanju željenog skupa i sastoji se u odabiru po jednog lanca iz svakog pokrivenog clustera. Kao predstavnika tog clustera odabire se onaj lanac, koji unutar iste PDB strukture ima najviše lanaca u ostalim clusterima za hetero, odnosno najviše lanaca u istom clusteru za homo strukture. Koliki je taj broj određuje se pomoću rangirane cluster datoteke zadanog postotka sličnosti u kojoj uz svaki lanac stoji traženi broj. Takva pomoćna datoteka generira se iz originalne cluster datoteke za zadani postotak sličnosti jednostavnim prebrojavanjem broja traženih lanaca.

2.6. Provjera lanaca koji nisu u clusteru

Neki lanci se zbog određenih svojstava ne nalaze u cluster datotekama ni za koji postotak sličnosti. Takvi lanci bili su prisutni i u listi PDB struktura nakon parsiranja, ali smo ih zanemarili daljnjim promatranjem lanaca isključivo preko njihova prisutstva u clusteru. U ovom koraku ipak provjeravamo postoji li u PDB strukturama neki lanac koji je dulji od 30 aminokiselinskih ostataka, ali istovremeno da ga ne čine vrijednosti poput (“A”, “G”, “C”, “U”, “T”, tj. baze koje grade DNA, dalje “DA”, “DG”, “DC”, “DU”, “DT” ni “UNK” odnosno nepoznate unose). Potraga za takvim lancima ipak je za ispitivanu listu završila bez rezultata, stoga lista dobivena u 5. koraku nije doživjela nikakve promjene.

Spajanjem rezultata iz 5. i 6. koraka (koja je u ovom konkretnom slučaju ispala prazna lista) dobili smo konačni neredundantni skup lanaca visoke rezolucije, odgovarajuće tehnike dobivanja i duljine koji inače tvore biološki funkcionalne jedinke i procjenjuje se da bi trebali tvoriti veći broj interakcija nego njima slični lanci.

3. Svojstva novog skupa

3.1 Analiza kvantitativnih i kvalitativnih svojstava skupa

Novodobiveni skup sastoji se od dvije liste koje broje 1777 lanaca koji tvore hetero interakcije, te 2880 lanaca koji tvore homo interakcije. Ovih lanaca očekivano je znatno više nego lanaca iz prethodno analiziranih radova, otprilike 3 do 10 puta, jer novi opisi struktura svakodnevno se dodaju u RSCB PDB bazu. Automatizirani postupak omogućava ponovno kreiranje skupa u vrlo kratkom vremenu i dodatno povećanje skupa uz moguću promjenu nekih parametara. Tako bi, primjerice, da dođe do znatnog povećanja preciznosti pri opisivanju novih struktura vrlo lako mogli smanjiti najveću dopuštenu rezoluciju i tako dobiti još bolje lance u skupu.

Što se svojstava samih lanaca tiče, možemo tvrditi da ovaj skup ne sadrži redundanciju u vidu primarne strukture veću od 30%, što se nije pokazalo potpuno točnim za prethodno analizirane skupove. Zadani postotak maksimalne sličnosti lanaca također je vrlo lako moguće promijeniti. Predefinirana rezolucija skupa s zadanom tehnikom dobivanja osigurava da je struktura samih lanaca precizno određena.

3.2 Statistička analiza skupa

Prvo statističko ispitivanje provedeno nad dobivenim skupom analizira učestalost pojavljivanja pojedinih tipova aminokiselinskih ostataka. U tablicama 3.1 i 3.2 prikazani su rezultati ovog ispitivanja za hetero i homo lance. Vrijednosti u stupcima prikazuju redom koliko se puta pojedini tip aminokiselinskog ostatka javlja u skupu, zatim koliko se puta javlja u interakcijama, zatim omjer broja pojavljivanja u interakcijama i pojavljivanja u skupu i na kraju omjer broja pojavljivanja u interakcijama i ukupnog broja pojavljivanja svih tipova aminokiselinskih ostataka u interakcijama.

	Ukupni broj pojavljivanja u skupu:	Ukupni broj pojavljivanja u interakcijama:	Omjer broja pojavljivanja u interakcijama i u skupu:	Omjer broja pojavljivanja u interakcijama i broja svih koji sudjeluju u interakcijama:
ALA	32312	4941	0,1529	0.0572
ARG	32312	6739	0,2085	0.0780
ASN	17720	3724	0,2101	0.0431
ASP	23240	4157	0,1788	0.0481
CYS	6270	1189	0,1896	0.0138
GLN	16486	3750	0,2274	0.0434
GLU	29437	5184	0,1761	0.0600
GLY	28067	4319	0,1538	0.0500
HIS	9754	2353	0,2412	0.0272
ILE	24884	4933	0,1982	0.0571
LEU	41422	8501	0,2052	0.0984
LYS	25166	4426	0,1758	0.0512
MET	9429	2515	0,2667	0.0291
PHE	18026	4751	0,2635	0.0550
PRO	19247	3880	0,2015	0.0449
SER	24880	4447	0,1787	0.0515
THR	22813	4391	0,1924	0.0508
TRP	6389	2022	0,3164	0.0234
TYR	15960	4935	0,3092	0.0571
VAL	29553	5252	0,1777	0.0608

Tablica 3.1: Statistika pojavljivanja tipova aminokiselinskih ostataka - hetero

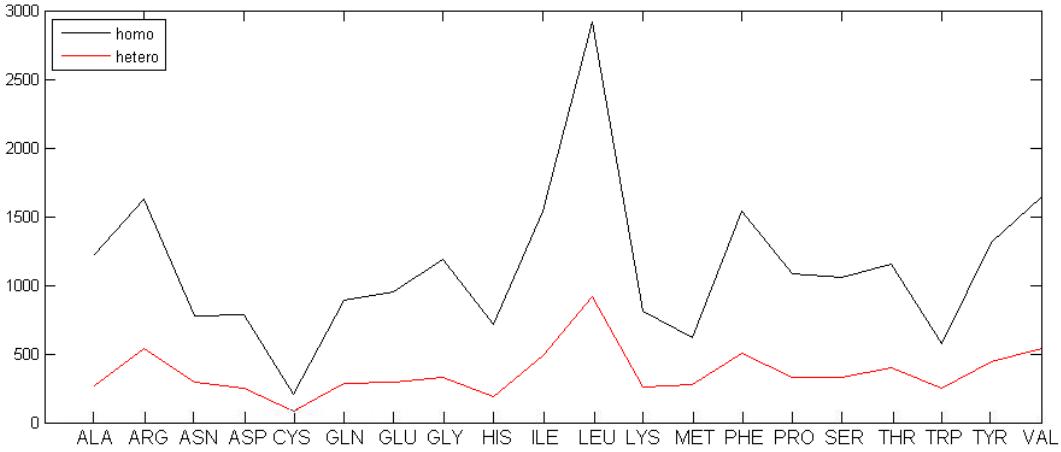
	Ukupni broj pojavljivanja u skupu:	Ukupni broj pojavljivanja u interakcijama:	Omjer broja pojavljivanja u interakcijama i u skupu:	Omjer broja pojavljivanja u interakcijama i broja svih koji sudjeluju u interakcijama:
ALA	70246	18532	0,2638	0.0596
ARG	53561	24809	0,4631	0.0797
ASN	36805	12430	0,3377	0.0399
ASP	49621	14165	0,2854	0.0455
CYS	10248	2869	0,2799	0.0092
GLN	33401	12730	0,3811	0.0409
GLU	61332	18735	0,3054	0.0602
GLY	60282	15292	0,2536	0.0491
HIS	23737	9828	0,414	0.0316
ILE	55334	18959	0,3426	0.0609
LEU	90693	33694	0,3715	0.1083
LYS	49126	14923	0,3037	0.0480
MET	16645	7320	0,4397	0.0235
PHE	41903	18094	0,4318	0.0581
PRO	41749	14666	0,3512	0.0471
SER	50038	15288	0,3055	0.0491
THR	47827	15772	0,3297	0.0507
TRP	14069	6437	0,4575	0.0207
TYR	36209	16491	0,4554	0.0530
VAL	64353	20132	0,3128	0.0647

Tablica 3.2: Statistika pojavljivanja tipova aminokiselinskih ostataka - homo

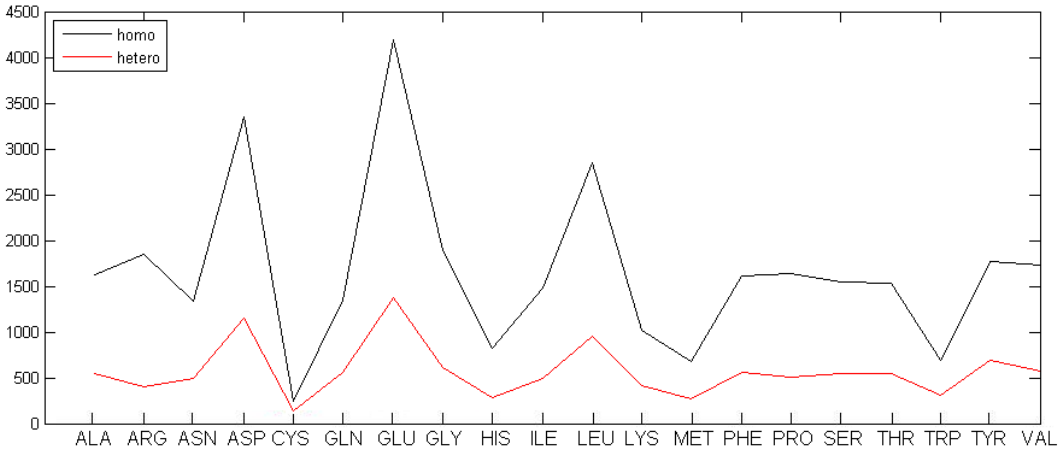
U tablici 3.3 dana je usporedba učestalosti pojavljivanja (omjer broja pojavljivanja u interakcijama za pojedini tip aminokiselinskog ostatka i broja svih aminokiselinskih ostataka koji sudjeluju u interakcijama) između homo i hetero lanaca preuzeta radi preglednosti sa tablica 3.1 i 3.2. Može se primijetiti da su odgovarajući tipovi aminokiselinskih ostataka u homo i hetero slučaju u listi međusobno udaljeni za najviše dva mjesta. Na slikama 3.1 – 3.20 dalje su prikazane vrijednosti koliko je puta koji pojedini tip aminokiselinskog ostatka u interakciji s kojim drugim tipom.

Homo		Hetero	
LEU	0.0984	LEU	0.1083
ARG	0.0780	ARG	0.0797
VAL	0.0608	VAL	0.0647
GLU	0.0600	ILE	0.0609
ALA	0.0572	GLU	0.0602
ILE	0.0571	ALA	0.0596
TYR	0.0571	PHE	0.0581
PHE	0.0550	TYR	0.0530
SER	0.0515	THR	0.0507
LYS	0.0512	GLY	0.0491
THR	0.0508	SER	0.0491
GLY	0.0500	LYS	0.0480
ASP	0.0481	PRO	0.0471
PRO	0.0449	ASP	0.0455
GLN	0.0434	GLN	0.0409
ASN	0.0431	ASN	0.0399
MET	0.0291	HIS	0.0316
HIS	0.0272	MET	0.0235
TRP	0.0234	TRP	0.0207
CYS	0.0138	CYS	0.0092

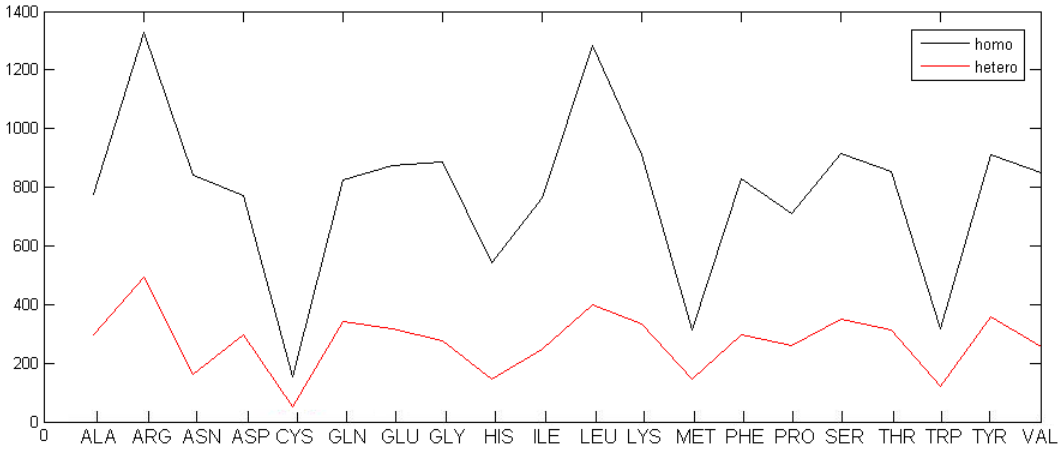
Tablica 3.3: Usporedba između homo i hetero lanaca



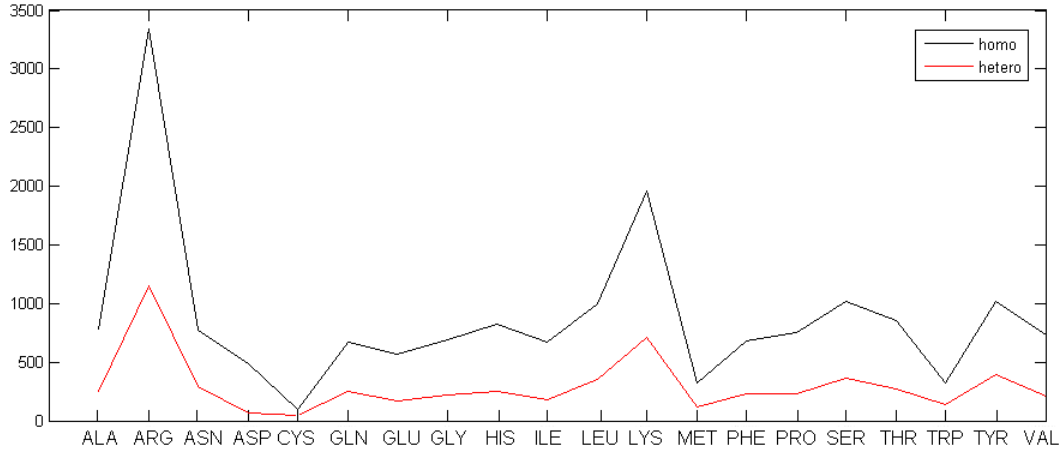
Slika 3.1: Interakcije s ALA



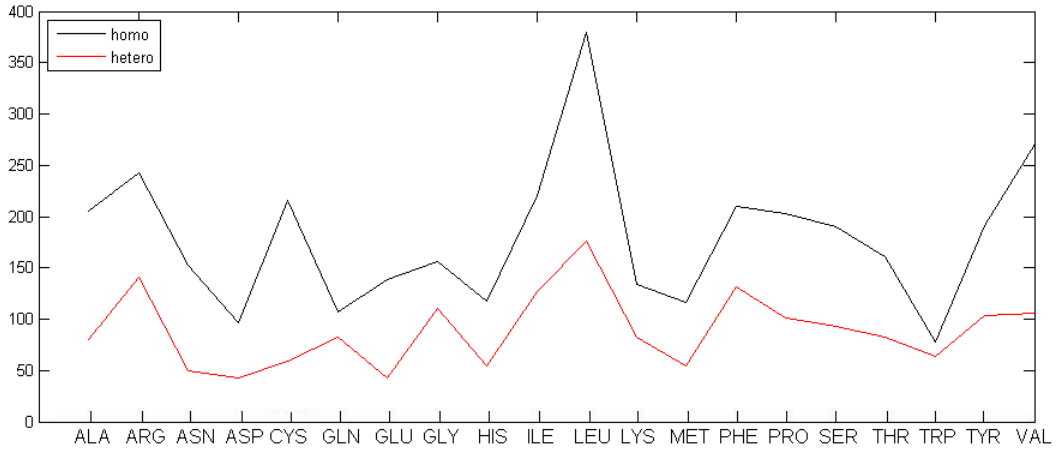
Slika 3.2: Interakcije s ARG



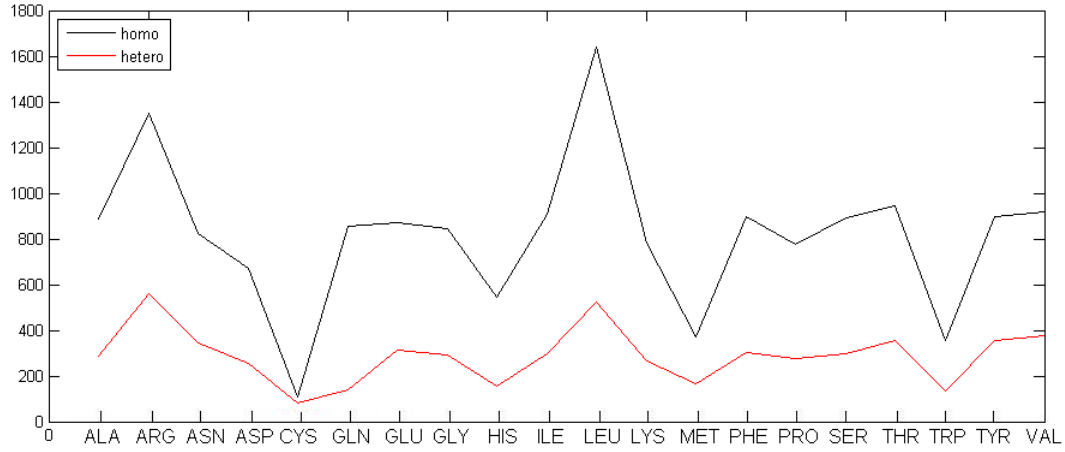
Slika 3.3: Interakcije s ASN



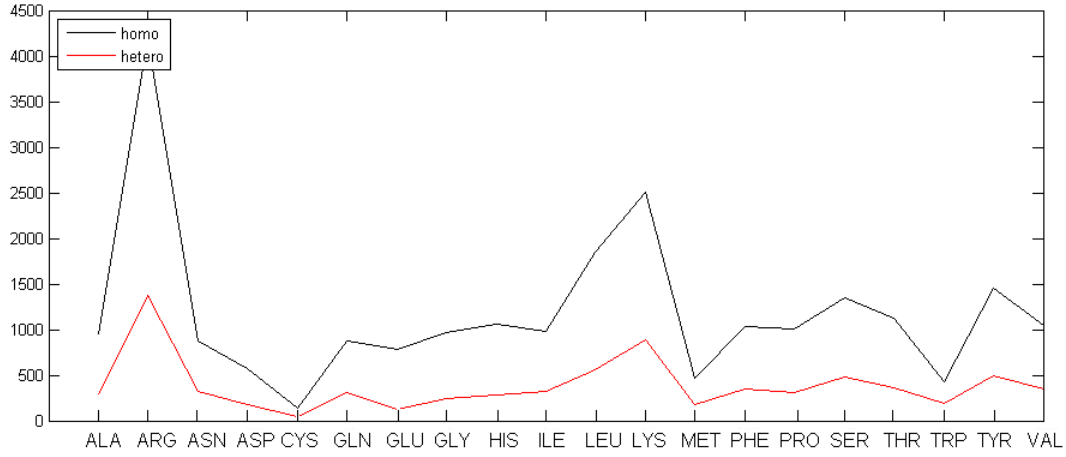
Slika 3.4: Interakcije s ASP



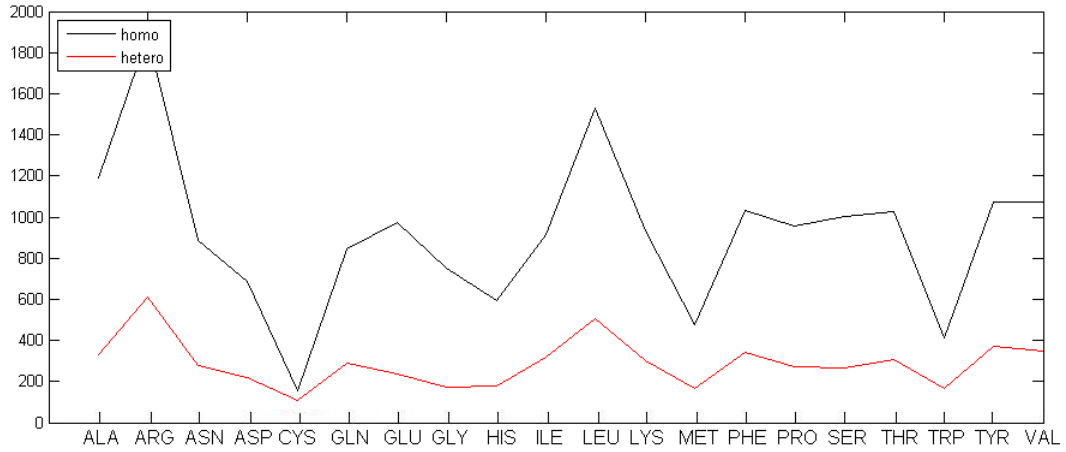
Slika 3.5: Interakcije s CYS



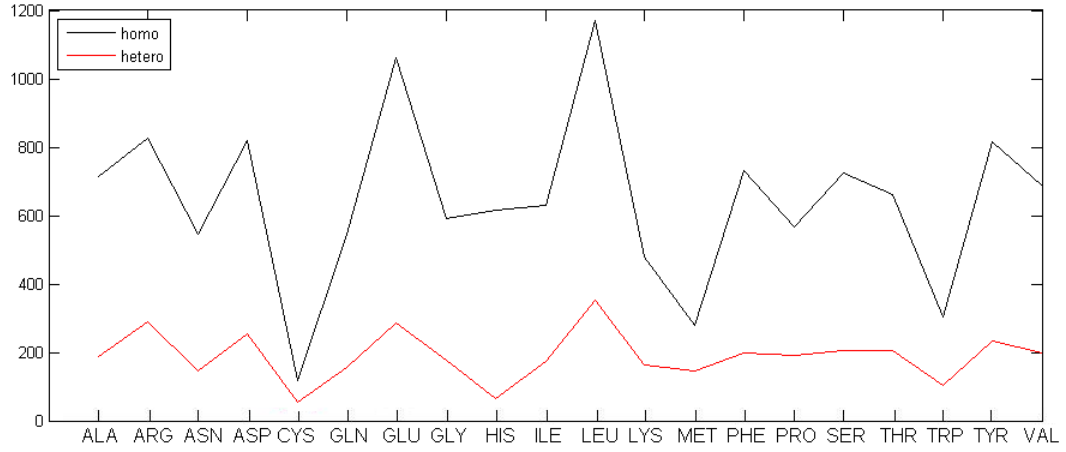
Slika 3.6: Interakcije s GLN



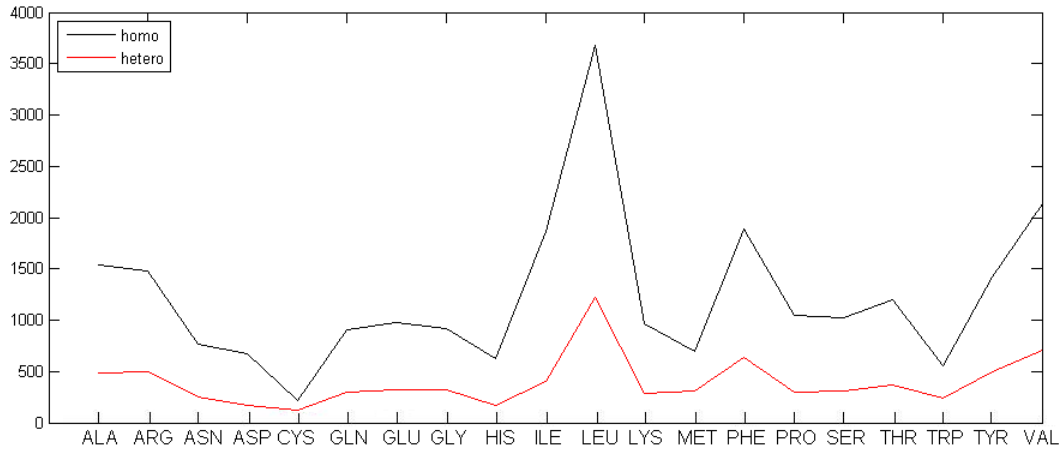
Slika 3.7: Interakcije s GLU



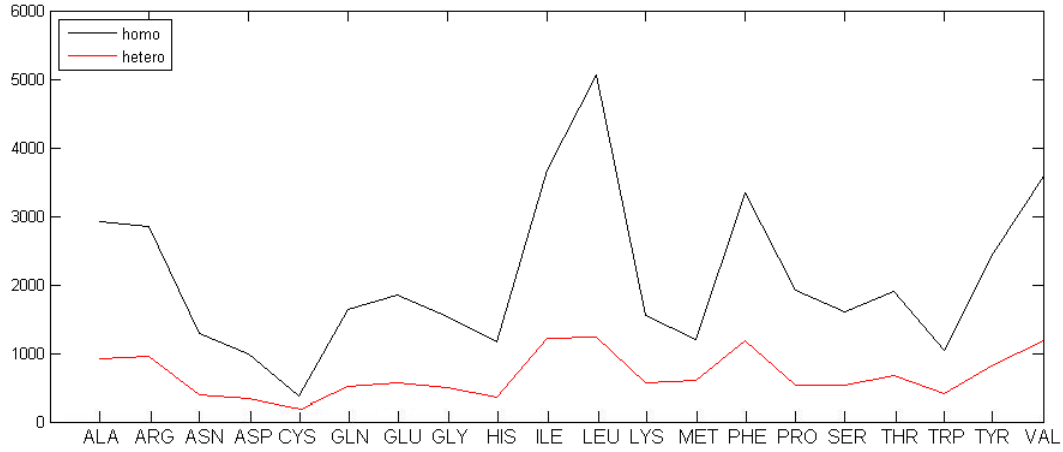
Slika 3.8: Interakcije s GLY



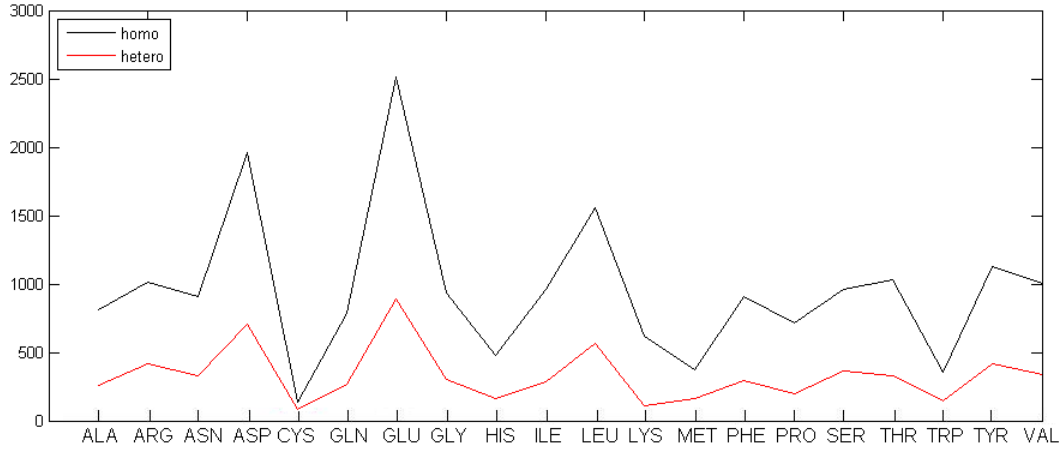
Slika 3.9: Interakcije s HIS



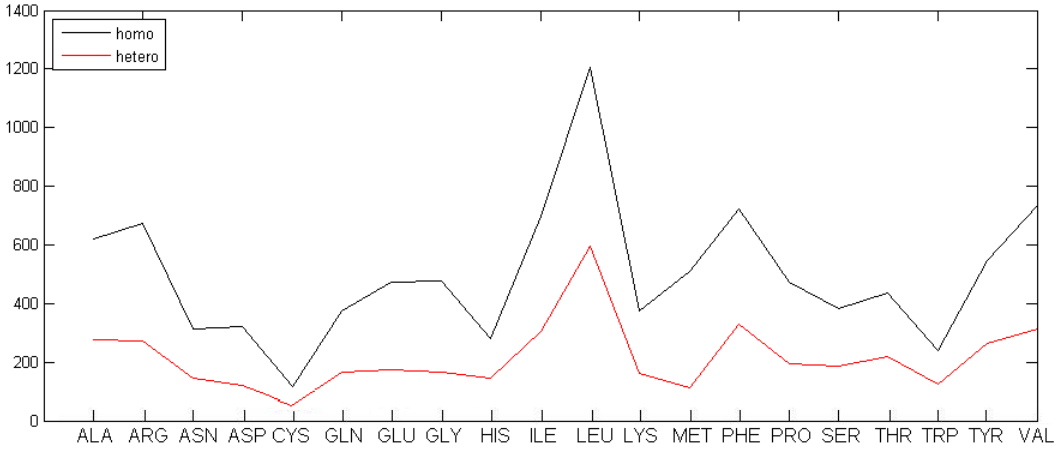
Slika 3.10: Interakcije s ILE



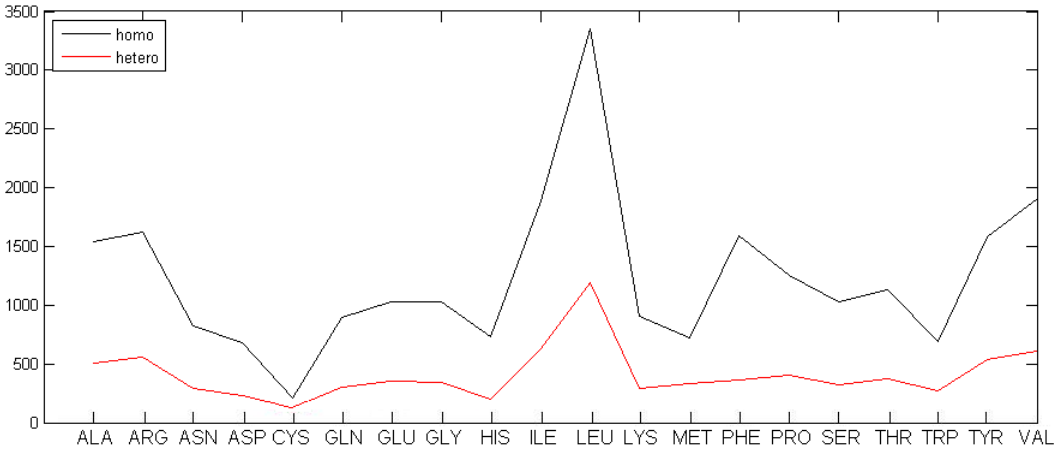
Slika 3.11: Interakcije LEU



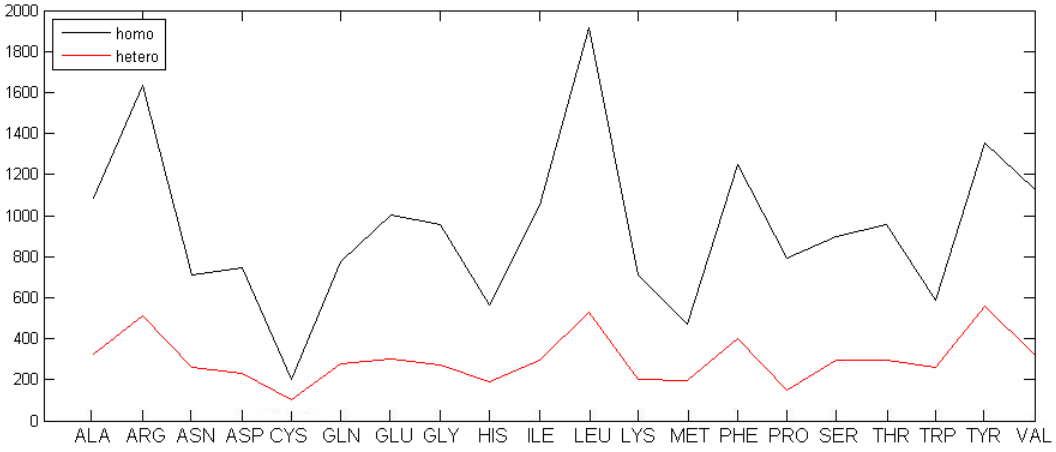
Slika 3.12: Interakcije s LYS



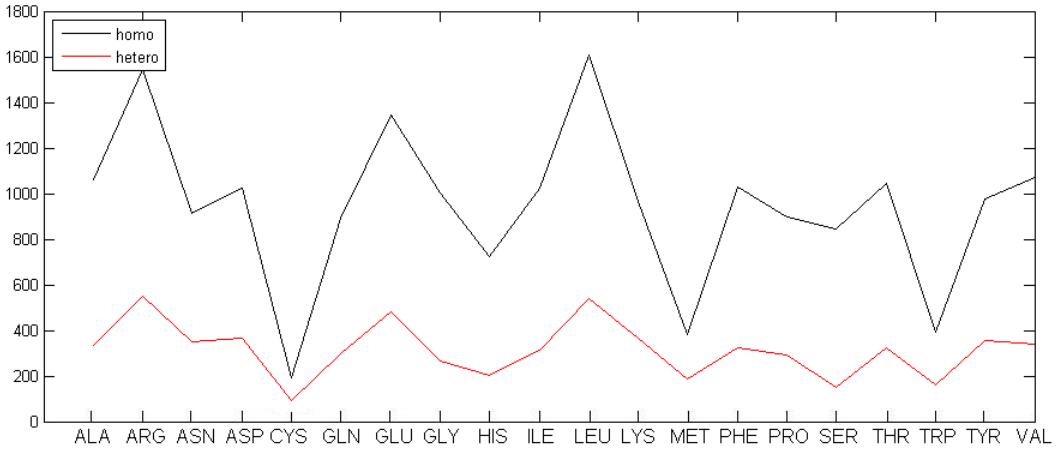
Slika 3.13: Interakcije s MET



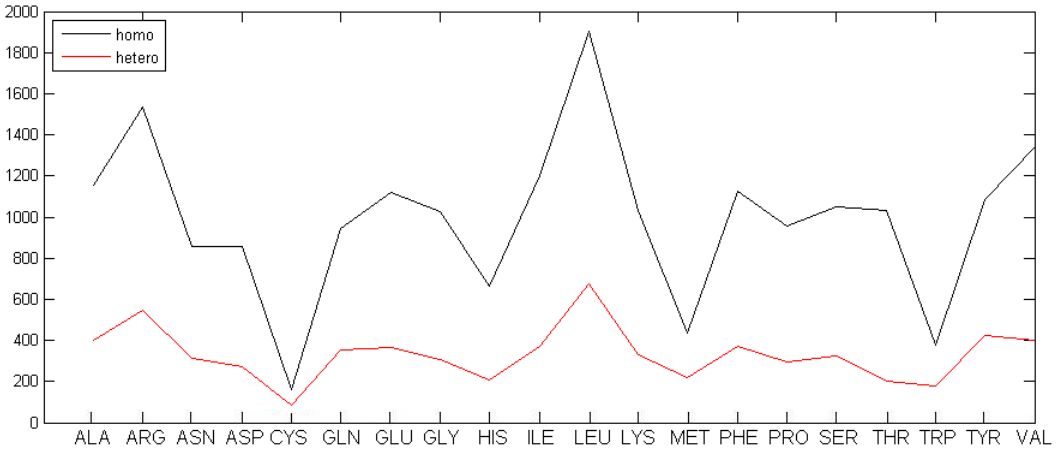
Slika 3.14: Interakcije s PHE



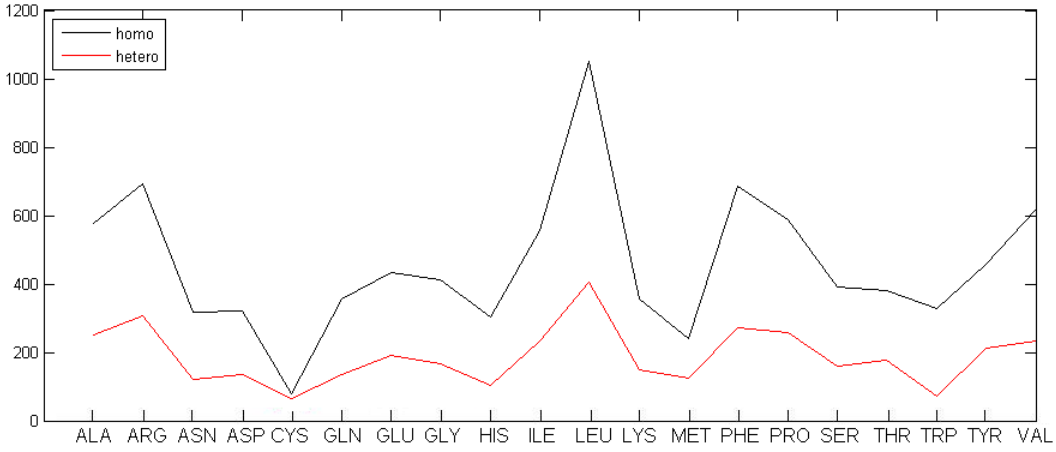
Slika 3.15: Interakcije s PRO



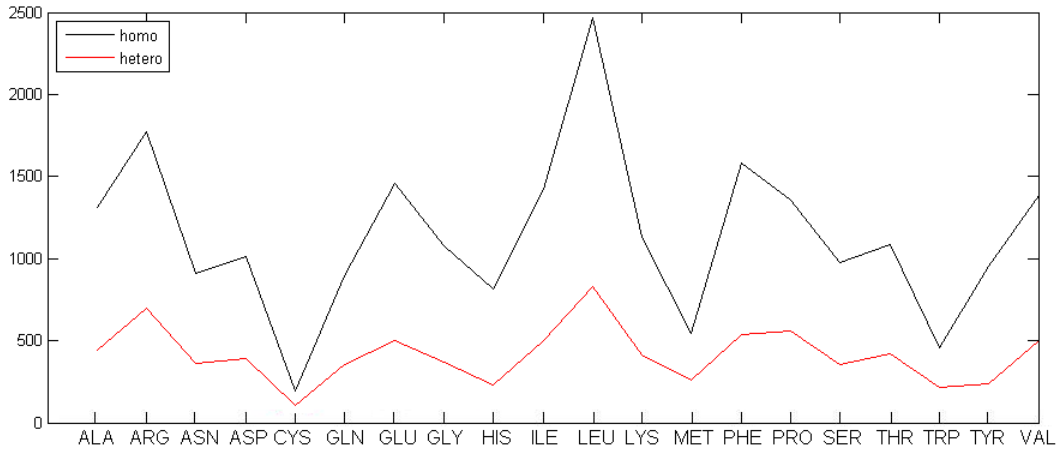
Slika 3.16: Interakcije s SER



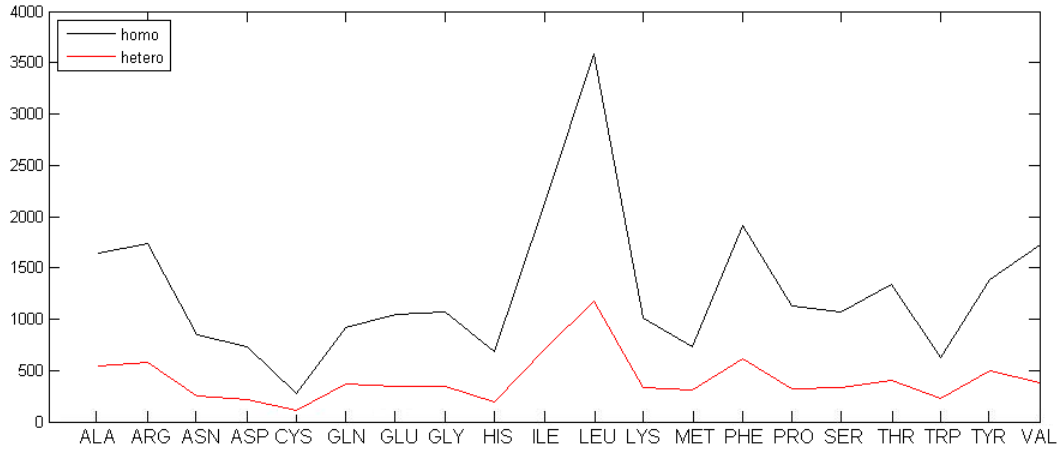
Slika 3.17: Interakcije s THR



Slika 3.18: Interakcije s TRP



Slika 3.19: Interakcije s TYR



Slika 3.20: Interakcije s VAL

Na slikama jasno primjećujemo da se određeni obrasci ponavljaju, a to su najčešće šiljak iznad LEU i ARG. Takve rezultate smo i očekivali budući da je iz prethodnih tablica očito da su to dva najzastupljenija tipa aminokiselinskih ostataka. Drugi rezultat koji možemo primijetiti je sličnost statističkih vrijednosti dobivenih za homo i hetero lance. Crni dio grafa, koji predstavlja interakcije kod homo lanaca što se rastova i padova tiče gotovo uvijek vjerno prati crveni dio grafa koji predstavlja interakcije hetero lanaca. Ovo znači da uočene pravilnosti što se frekvencija pojavljivanja tiče statistički vrijede i za homo i hetero lance, ali je razlika u kvantiteti odnosno broju interakcija, koji je za homo lance uvijek veći.

Rezultati konkretne numeričke zastupljenosti interakcija aminokiselinskih ostataka pojedine aminokiseline s drugim tipovima aminokiselinskih ostataka navedeni su u tablici 3.4 za hetero i 3.5 za homo lance. Ovo je tablični prikaz prethodnih vrijednosti grafova sa slika 3.1 – 3.20. Obije tablice su simetrične jer je, primjerice, interakcija ALA – ARG ista kao i interakcija ARG – ALA. Iz odgovarajućeg polja u tablici možemo očitati da je u skupu pronađeno ukupno 542 takvih interakcija. Zbog simetričnosti tablice, ako zbrojimo elemente gornjeg ili donjeg trokuta tablice s elementima na dijagonali dobit ćemo ukupan broj parova aminokiselinskih ostataka koji su u interakciji. Za hetero lance taj broj iznosi ukupno 69916 parova, a za homo lance ukupno 211128 parova u interakciji.

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	269	542	295	247	80	288	295	333	187	484	913	263	276	506	325	333	403	250	445	543
ARG	542	407	495	1152	141	559	1372	613	288	493	955	420	274	556	511	549	544	307	695	578
ASN	295	495	163	296	50	343	316	277	146	246	398	332	145	296	261	351	312	120	359	255
ASP	247	1152	296	66	43	253	173	220	253	175	349	709	120	229	233	368	273	136	393	213
CYS	80	141	50	43	59	82	43	111	54	127	176	83	54	131	101	93	82	64	103	106
GLN	288	559	343	253	82	141	314	291	156	296	524	266	164	305	275	298	354	134	353	374
GLU	295	1372	316	173	43	314	130	238	286	317	565	891	174	350	304	484	366	191	499	348
GLY	333	613	277	220	111	291	238	170	177	320	507	303	166	340	272	264	307	165	369	346
HIS	187	288	146	253	54	156	286	177	66	174	354	164	146	200	190	204	207	105	233	198
ILE	484	493	246	175	127	296	317	320	174	405	1219	288	307	633	296	311	373	234	498	713
LEU	913	955	398	349	176	524	565	507	354	1219	1233	569	597	1189	531	540	676	406	829	1181
LYS	263	420	332	709	83	266	891	303	164	288	569	113	161	292	200	366	333	149	416	335
MET	276	274	145	120	54	164	174	166	146	307	597	161	112	330	196	187	219	123	263	315
PHE	506	556	296	229	131	305	350	340	200	633	1189	292	330	363	400	322	374	272	534	613
PRO	325	511	261	233	101	275	304	272	190	296	531	200	196	400	151	293	295	258	556	320
SER	333	549	351	368	93	298	484	264	204	311	540	366	187	322	293	152	326	160	354	339
THR	403	544	312	273	82	354	366	307	207	373	676	333	219	374	295	326	200	176	422	401
TRP	250	307	120	136	64	134	191	165	105	234	406	149	123	272	258	160	176	73	214	232
TYR	445	695	359	393	103	353	499	369	233	498	829	416	263	534	556	354	422	214	238	499
VAL	543	578	255	213	106	374	348	346	198	713	1181	335	315	613	320	339	401	232	499	379

Tablica 3.4: Brojevi interakcija pojedinih parova aminokiselinskih ostataka - hetero

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	1220	1629	774	784	205	886	956	1189	715	1540	2920	811	620	1539	1087	1061	1154	577	1313	1647
ARG	1629	1846	1329	3345	243	1348	4188	1882	827	1476	2846	1016	673	1617	1637	1547	1538	692	1772	1732
ASN	774	1329	842	770	153	823	873	888	544	763	1282	911	315	827	710	914	855	316	910	850
ASP	784	3345	770	489	97	670	566	686	821	667	996	1957	320	681	748	1022	856	322	1014	734
CYS	205	243	153	97	216	107	139	156	117	220	380	134	116	210	203	190	161	78	191	271
GLN	886	1348	823	670	107	855	871	843	544	909	1640	783	373	895	776	890	947	357	897	918
GLU	956	4188	873	566	139	871	784	971	1062	982	1843	2509	473	1032	1002	1347	1122	433	1460	1046
GLY	1189	1882	888	686	156	843	971	753	592	913	1527	932	477	1033	957	1002	1027	413	1075	1075
HIS	715	827	544	821	117	544	1062	592	615	631	1169	480	279	732	567	726	661	305	816	687
ILE	1540	1476	763	667	220	909	982	913	631	1870	3684	961	697	1884	1047	1025	1204	556	1422	2137
LEU	2920	2846	1282	996	380	1640	1843	1527	1169	3684	5055	1558	1207	3350	1914	1606	1901	1050	2467	3579
LYS	811	1016	911	1957	134	783	2509	932	480	961	1558	619	376	907	713	964	1031	357	1130	1008
MET	620	673	315	320	116	373	473	477	279	697	1207	376	511	724	471	384	436	239	546	733
PHE	1539	1617	827	681	210	895	1032	1033	732	1884	3350	907	724	1595	1251	1027	1128	686	1585	1908
PRO	1087	1637	710	748	203	776	1002	957	567	1047	1914	713	471	1251	795	898	958	588	1353	1124
SER	1061	1547	914	1022	190	890	1347	1002	726	1025	1606	964	384	1027	898	844	1047	391	974	1071
THR	1154	1538	855	856	161	947	1122	1027	661	1204	1901	1031	436	1128	958	1047	1031	380	1085	1340
TRP	577	692	316	322	78	357	433	413	305	556	1050	357	239	686	588	391	380	327	459	620
TYR	1313	1772	910	1014	191	897	1460	1075	816	1422	2467	1130	546	1585	1353	974	1085	459	954	1387
VAL	1647	1732	850	734	271	918	1046	1075	687	2137	3579	1008	733	1908	1124	1071	1340	620	1387	1729

Tablica 3.5: Brojevi interakcija pojedinih parova aminokiselinskih ostataka - homo

Korisno je, da bi se izbjegle eventualne nedoumice, primijetiti još i sljedeće: ako zbrojimo sve vrijednosti u jednom retku tablice 3.4 ili 3.5, dobivena vrijednost ne mora odgovarati nekoj od vrijednosti iz tablica 3.1 ili 3.2. Iako se može učiniti da su tako pobrojane sve interakcije u kojima sudjeluje pojedini tip aminokiselinskog ostatka pa njihov broj mora biti jednak ukupnom broju tog tipa aminokiselinskog ostataka koji sudjeluje u interakcijama, prva počinjena greška je da element koji se u tom retku nalazi na glavnoj dijagonali mora biti pribrojen još jednom, jer su u njemu oba aminokiselinska ostatka zadanog tipa, a drugi je da jedan jedini aminokiselinski ostatak može tvoriti više parova interakcija, iz čega proizlazi razlika u dobivenim brojevima.

U nastavku su u tablicama 3.6 i 3.7 prikazane prethodne tablice sa slika 3.4 i 3.5 u postocima, gdje je svako polje tablice podijeljeno s ukupnim brojem parova u interakciji koji je naveden ranije.

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.38	0.78	0.42	0.35	0.11	0.41	0.42	0.48	0.27	0.69	1.31	0.38	0.39	0.72	0.46	0.48	0.58	0.36	0.64	0.78
ARG	0.78	0.58	0.71	1.65	0.2	0.8	1.96	0.88	0.41	0.71	1.37	0.6	0.39	0.8	0.73	0.79	0.78	0.44	0.99	0.83
ASN	0.42	0.71	0.23	0.42	0.07	0.49	0.45	0.4	0.21	0.35	0.57	0.47	0.21	0.42	0.37	0.5	0.45	0.17	0.51	0.36
ASP	0.35	1.65	0.42	0.09	0.06	0.36	0.25	0.31	0.36	0.25	0.5	1.01	0.17	0.33	0.33	0.53	0.39	0.19	0.56	0.3
CYS	0.11	0.2	0.07	0.06	0.08	0.12	0.06	0.16	0.08	0.18	0.25	0.12	0.08	0.19	0.14	0.13	0.12	0.09	0.15	0.15
GLN	0.41	0.8	0.49	0.36	0.12	0.2	0.45	0.42	0.22	0.42	0.75	0.38	0.23	0.44	0.39	0.43	0.51	0.19	0.5	0.53
GLU	0.42	1.96	0.45	0.25	0.06	0.45	0.19	0.34	0.41	0.45	0.81	1.27	0.25	0.5	0.43	0.69	0.52	0.27	0.71	0.5
GLY	0.48	0.88	0.4	0.31	0.16	0.42	0.34	0.24	0.25	0.46	0.73	0.43	0.24	0.49	0.39	0.38	0.44	0.24	0.53	0.49
HIS	0.27	0.41	0.21	0.36	0.08	0.22	0.41	0.25	0.09	0.25	0.51	0.23	0.21	0.29	0.27	0.29	0.3	0.15	0.33	0.28
ILE	0.69	0.71	0.35	0.25	0.18	0.42	0.45	0.46	0.25	0.58	1.74	0.41	0.44	0.91	0.42	0.44	0.53	0.33	0.71	1.02
LEU	1.31	1.37	0.57	0.5	0.25	0.75	0.81	0.73	0.51	1.74	1.76	0.81	0.85	1.7	0.76	0.77	0.97	0.58	1.19	1.69
LYS	0.38	0.6	0.47	1.01	0.12	0.38	1.27	0.43	0.23	0.41	0.81	0.16	0.23	0.42	0.29	0.52	0.48	0.21	0.59	0.48
MET	0.39	0.39	0.21	0.17	0.08	0.23	0.25	0.24	0.21	0.44	0.85	0.23	0.16	0.47	0.28	0.27	0.31	0.18	0.38	0.45
PHE	0.72	0.8	0.42	0.33	0.19	0.44	0.5	0.49	0.29	0.91	1.7	0.42	0.47	0.52	0.57	0.46	0.53	0.39	0.76	0.88
PRO	0.46	0.73	0.37	0.33	0.14	0.39	0.43	0.39	0.27	0.42	0.76	0.29	0.28	0.57	0.22	0.42	0.42	0.37	0.8	0.46
SER	0.48	0.79	0.5	0.53	0.13	0.43	0.69	0.38	0.29	0.44	0.77	0.52	0.27	0.46	0.42	0.22	0.47	0.23	0.51	0.48
THR	0.58	0.78	0.45	0.39	0.12	0.51	0.52	0.44	0.3	0.53	0.97	0.48	0.31	0.53	0.42	0.47	0.29	0.25	0.6	0.57
TRP	0.36	0.44	0.17	0.19	0.09	0.19	0.27	0.24	0.15	0.33	0.58	0.21	0.18	0.39	0.37	0.23	0.25	0.1	0.31	0.33
TYR	0.64	0.99	0.51	0.56	0.15	0.5	0.71	0.53	0.33	0.71	1.19	0.59	0.38	0.76	0.8	0.51	0.6	0.31	0.34	0.71
VAL	0.78	0.83	0.36	0.3	0.15	0.53	0.5	0.49	0.28	1.02	1.69	0.48	0.45	0.88	0.46	0.48	0.57	0.33	0.71	0.54

Tablica 3.6: Relativna učestalost (%) pojavljivanja interakcija pojedinih tipova aminokiselinskih ostataka - hetero

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0.58	0.77	0.37	0.37	0.1	0.42	0.45	0.56	0.34	0.73	1.38	0.38	0.29	0.73	0.51	0.5	0.55	0.27	0.62	0.78
ARG	0.77	0.87	0.63	1.58	0.12	0.64	1.98	0.89	0.39	0.7	1.35	0.48	0.32	0.77	0.78	0.73	0.73	0.33	0.84	0.82
ASN	0.37	0.63	0.4	0.36	0.07	0.39	0.41	0.42	0.26	0.36	0.61	0.43	0.15	0.39	0.34	0.43	0.4	0.15	0.43	0.4
ASP	0.37	1.58	0.36	0.23	0.05	0.32	0.27	0.32	0.39	0.32	0.47	0.93	0.15	0.32	0.35	0.48	0.41	0.15	0.48	0.35
CYS	0.1	0.12	0.07	0.05	0.1	0.05	0.07	0.07	0.06	0.1	0.18	0.06	0.05	0.1	0.1	0.09	0.08	0.04	0.09	0.13
GLN	0.42	0.64	0.39	0.32	0.05	0.4	0.41	0.4	0.26	0.43	0.78	0.37	0.18	0.42	0.37	0.42	0.45	0.17	0.42	0.43
GLU	0.45	1.98	0.41	0.27	0.07	0.41	0.37	0.46	0.5	0.47	0.87	1.19	0.22	0.49	0.47	0.64	0.53	0.21	0.69	0.5
GLY	0.56	0.89	0.42	0.32	0.07	0.4	0.46	0.36	0.28	0.43	0.72	0.44	0.23	0.49	0.45	0.47	0.49	0.2	0.51	0.51
HIS	0.34	0.39	0.26	0.39	0.06	0.26	0.5	0.28	0.29	0.3	0.55	0.23	0.13	0.35	0.27	0.34	0.31	0.14	0.39	0.33
ILE	0.73	0.7	0.36	0.32	0.1	0.43	0.47	0.43	0.3	0.89	1.74	0.46	0.33	0.89	0.5	0.49	0.57	0.26	0.67	1.01
LEU	1.38	1.35	0.61	0.47	0.18	0.78	0.87	0.72	0.55	1.74	2.39	0.74	0.57	1.59	0.91	0.76	0.9	0.5	1.17	1.7
LYS	0.38	0.48	0.43	0.93	0.06	0.37	1.19	0.44	0.23	0.46	0.74	0.29	0.18	0.43	0.34	0.46	0.49	0.17	0.54	0.48
MET	0.29	0.32	0.15	0.15	0.05	0.18	0.22	0.23	0.13	0.33	0.57	0.18	0.24	0.34	0.22	0.18	0.21	0.11	0.26	0.35
PHE	0.73	0.77	0.39	0.32	0.1	0.42	0.49	0.49	0.35	0.89	1.59	0.43	0.34	0.76	0.59	0.49	0.53	0.32	0.75	0.9
PRO	0.51	0.78	0.34	0.35	0.1	0.37	0.47	0.45	0.27	0.5	0.91	0.34	0.22	0.59	0.38	0.43	0.45	0.28	0.64	0.53
SER	0.5	0.73	0.43	0.48	0.09	0.42	0.64	0.47	0.34	0.49	0.76	0.46	0.18	0.49	0.43	0.4	0.5	0.19	0.46	0.51
THR	0.55	0.73	0.4	0.41	0.08	0.45	0.53	0.49	0.31	0.57	0.9	0.49	0.21	0.53	0.45	0.5	0.49	0.18	0.51	0.63
TRP	0.27	0.33	0.15	0.15	0.04	0.17	0.21	0.2	0.14	0.26	0.5	0.17	0.11	0.32	0.28	0.19	0.18	0.15	0.22	0.29
TYR	0.62	0.84	0.43	0.48	0.09	0.42	0.69	0.51	0.39	0.67	1.17	0.54	0.26	0.75	0.64	0.46	0.51	0.22	0.45	0.66
VAL	0.78	0.82	0.4	0.35	0.13	0.43	0.5	0.51	0.33	1.01	1.7	0.48	0.35	0.9	0.53	0.51	0.63	0.29	0.66	0.82

Tablica 3.7 : Relativna učestalost (%) pojavljivanja interakcija pojedinih tipova aminokiselinskih ostataka - homo

Iduća statistika predstavlja analizu sličnu prethodnoj, samo ne promatra pojedine interakcije nego dva susjedna aminokiselinska ostatka s jednog lanca koji su u interakciji s dva susjedna aminokiselinska ostatka s drugog lanca. Kombinacije s najviše ponavljanja prikazane su u tablici 3.8.

Homo strukture					Hetero strukture				
LEU	LEU	LEU	LEU	61	ARG	ALA	ASP	PHE	6
ALA	LEU	LEU	ALA	58	ARG	LEU	GLU	LEU	6
LEU	ALA	ALA	LEU	58	ASP	ASP	LEU	ARG	6
LEU	VAL	VAL	LEU	50	ASP	PHE	ARG	ALA	6
VAL	LEU	LEU	VAL	50	GLU	LEU	ARG	LEU	6
ARG	LEU	LEU	ARG	46	GLY	GLY	LEU	LEU	6
LEU	ARG	ARG	LEU	46	LEU	ARG	ASP	ASP	6
GLU	LEU	LEU	GLU	41	LEU	LEU	GLY	GLY	6
LEU	GLU	GLU	LEU	41					

Tablica 3.8: Parovi aminokiselinskih ostataka koji su u interakciji

Rezultati ukazuju da iako se brojčanošću ne ističu u većoj mjeri, devet od devet najčešćih parova aminokiselinskih ostataka koji integriraju u sebi sadrže bar dva ostatka tipa LEU za homo lance, odnosno šest od osam najčešćih parova aminokiselinskih ostataka koji integriraju u sebi sadrže bar jedan ostatak tipa LEU kod hetero lanaca. Vodeća kombinacija iz hetero lanaca samo za jedan nadmašuje kombinaciju LEU LEU LEU LEU koja je vodeća po brojnosti kod homo lanaca. Valja još samo primijetiti kako se pojedina kombinacija četiri aminokiselinska ostatka može pojaviti u četiri različita oblika koji su zapravo isti. Ako par ALA – LEU integrira s parom GLU – GLY, tada tu istu kombinaciju možemo zapisati i kao GLU – GLY : ALA – LEU, LEU – ALA : GLY – GLU i GLY – GLU : LEU – ALA. Ovo ne vrijedi samo u slučaju kada su sva četiri aminokiselinska ostatka istog tipa.

Posljednja statistika provedena je prebrojavanjem različitih tipova interakcija između prozora koji su definirani kao neprekinuti slijed od devet aminokiselinskih ostataka dvaju lanaca, takvih da su središnji aminokiselinski ostaci u interakciji. Za takve prozore definira se *niže područje* kao četiri aminokiselinska ostatka koji prethode središnjem i *više područje* kao četiri aminokiselinska ostatka koji dolaze nakon središnjeg. Prebrojavanjem broja interakcija iz nižeg područja u više, iz višeg područja u više, iz višeg područja u niže i iz nižeg područja u niže obuhvaćenim dvama prozorima kojima su središnji aminokiselinski ostaci u interakciji dobiveni su rezultati prikazani u tablicama 3.7 za hetero i homo lance. Navedeni tipovi interakcija označeni su redom s *NV*, *VV*, *VN*, *NN*. Također prebrojane su i sve interakcije u više područje (*V*), sve interakcije u niže područje (*N*) i sve interakcije prema središnjem aminokiselinskom ostatku (*S*).

	NN	NV	VN	VV	N	V	S
Hetero	37410	46507	46541	37620	102198	102090	165560
Homo	101657	156268	156078	102201	309241	309287	321600

Tablica 3.7: Statistika interakcija u prozorima

4. Zaključak

Proteinske interakcije u osnovi su odgovorne za najveći dio složenih bioloških procesa koje čovjek nastoji razumjeti i time znati kako utjecati na njih. Ispitivanje i analiza ovakvih interakcija, dobivanje podataka, postavljanje hipoteza te provođenje ispitivanja nisu nimalo jednostavni. Početni korak u ovakvim ispitivanjima je izbor inicijalnog skupa podataka. Brojne datoteke s opisima proteinskih struktura dostupne su na [3], ali njihova brojnost i raznolikost zahtijevaju da se ipak pažljivije posvetimo problemu izbora početnog skupa.

U ovom sam se seminarskom radu stoga posvetio analizi postojećih neredundantnih skupova iz istog područja. Pojam *neredundantan* opisuje svojstvo koje pri kreiranju skupa želimo postići i znači da skup sadrži raznolike strukture koje u sebi sadrže bar dio velike raznolikosti proteina i njihovih svojstava, bez suvišnog ponavljanja koje bi moglo izazvati i pristranost rezultata ispitivanja. Osim analize postojećih skupova, također sam predstavio automatizirani postupak izrade skupa na temelju nekih dostupnih resursa, te usporedio dobivene rezultate odnosno svojstva dobivenih proteinskih lanaca.

Dobiveni rezultati ukazuju na veliku sličnost u ponašanju homo i hetero lanaca, kao i na statistički češće sudjelovanje određenih aminokiselinskih ostataka u interakcijama. Sam skup predstavlja rezultat koji svoju primjenjivost može pokazati tek u nekom daljnjem ispitivanju. Očekujem i nadam se da će opisani postupak odnosno rezultati koje je polučio pronaći primjenu kao podloga za daljnja ispitivanja vezana za proteinske interakcije.

5. Literatura

[1] Što je bioinformatika, http://www.irb.hr/hr/research/initiatives/bioinf/korisno/bioinfodef/, 20. travanj 2009.

[2] What Is Bioinformatics?, http://www.ncbi.nlm.nih.gov/, 29. ožujak 2002.

[3] RSCB PDB, http://www.rcsb.org/pdb/home/home.do, svibanj 2009.

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

[5] Rost B., Sander C., Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216 –226.

[6] Naderi-Manesh H, Sadeghi M, Araf S, Movahedi AAM. Predicting of protein surface accessibility with information theory. Proteins 2001;42:452– 459.

[7] Carugo O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. Protein Eng 2000;13:607– 609.

[8] Cuff J. A., Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.

[9] EBI PISA, Protein Interfaces, Surfaces and Assemblies, http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html, svibanj 2009.