Univerzitet u Beogradu

Matematički Fakultet

Mirjana Milić

Statistička analiza povezanosti sekundarne strukture proteina i interfejsa prema drugim molekulima

Master Rad

Beograd

2015.

Univerzitet u Beogradu – Matematički fakultet

Master rad

Naslov:	**Statistička analiza povezanosti sekundarne strukture proteina i interfejsa prema drugim molekulima**
Student:	Mirjana Milić 1105/2010
Mentor:	dr Saša Malkov Univerzitet u Beogradu, Matematički fakultet
Članovi komisije:	dr Miodrag Živković Univerzitet u Beogradu, Matematički fakultet
	dr Nenad Mitić Univerzitet u Beogradu, Matematički fakultet
Datum:	

Sadržaj

Spisak slika i tabela

1. Uvod

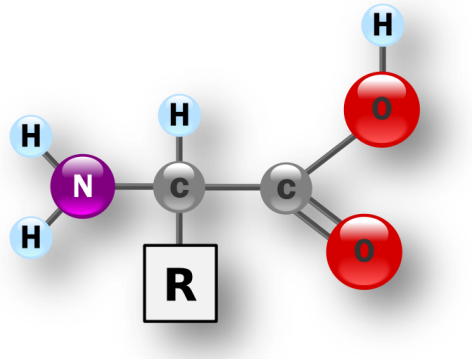
Proteini predstavljaju makromolekule koji imaju najvažniju ulogu u građi svih živih organizama i učestvuju u svim međućelijskim procesima [1]. Način delovanja proteina u organizmu je u velikoj meri određen njegovom strukturom. Raznovrsnost uloge proteina u organizmu je tesno vezana sa raznovrsnošću njihove strukture. Osim same strukture proteina, na njihovu ulogu utiče i njihova interakcija sa drugim molekulima ili proteinima.

Osnovnu jedinicu građe proteina čine **aminokiseline**, koje su međusobno povezane peptidnim vezama. Peptidna (amidna) veza je veza između amino grupe ($–NH_2$) jedne i karboksilne grupe (–COOH) druge aminokiseline, koja nastaje uz oslobađanje molekula vode (H_2O). Aminokiseline su mali molekuli, koji u svom sastavu imaju ugljenik C (α-carbon) za koga su vezane amino (NH_2) i karboksilna (COOH) grupa. Osim njih, ugljenik vezuje i jedan atom vodonika (H) i R-grupu. R-grupa se naziva bočni lanac aminokiseline. R-grupa je specifična za svaku aminokiselinu i određuje njene hemijske osobine. Uopštena struktura aminokiseline prikazana je na *Slici 1*.

Alanin	Ala (A)	Leucin	Leu (L)
Arginin	Arg (R)	Lizin	Lys (K)
Asparagin	Asn (N)	Metionin	Met (M)
Asparaginska kiselina	Asp (D)	Fenilalanin	Phe (F)
Cistein	Cys (C)	Prolin	Pro (P)
Glicin	Gly (G)	Serin	Ser (S)
Glutamin	Gln (Q)	Tirozin	Tyr (Y)
Glutaminska kiselina	Glu (E)	Triptofan	Trp (W)
Histidin	His (H)	Treonin	Thr (T)
Izoleucil	Ile (I)	Valin	Val (V)

Tabela 1. Aminokiseline i njihove troslovne (jednoslovne) oznake

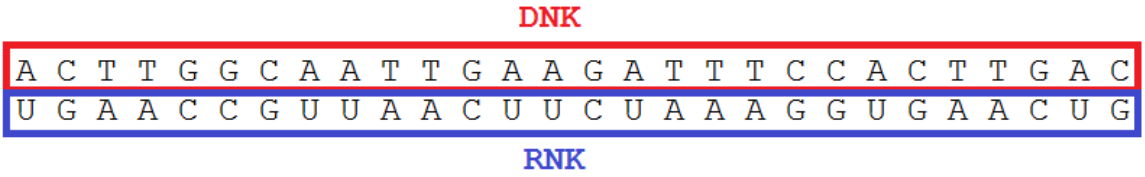
Primarna struktura proteina određena je nizom aminokiselina u polipeptidnom lancu (linearni polimer aminokiselina). Polipeptidni lanac nastaje kao posledica procesa koji se naziva sinteza proteina [2]. Početak polipeptidnog lanca se naziva *N-terminus*, dok se kraj naziva *C-terminus*. Glavni lanac, odnosno kičma proteina je izgrađena od α-ugljenika, kiseonika i azota svih aminokiselina koje učestvuju u građi proteina, dok bočne lance čine njihove R-grupe.



Slika 1. Uopštena struktura aminokiseline[1]

Raspored aminokiselina u proteinskim sekvencama je određen rasporedom nukleotida (purinskih i pirimidinskih baza) u delu lanca **dezoksiribonukleinske kiseline** (*DNK*) koji opisuje sastav proteina. DNK se sastoji od dva lanca nukleotida i ima ulogu nosioca sveukupne genetičke informacije jedne jedinke ili vrste. Nukleotidi koji učestvuju u izgradnji DNK su adenin (A), citozin (C), guanin (G) i timin (T). Nukleotidi grade dva komplementarno uparena lanca (A – T, C – G) koji su spiralno uvijeni.

Svaki od ova dva lanca sadrži informacije neophodne za nastanak proteina (sekvence baza). Da bi došlo do procesa sinteze proteina, neophodno je da se privremeno raskinu vodonične veze između nukleotida i da se DNK podeli na dva lanca. Nakon razdvajanja DNK lanca, *transkripcijom* se vrši sinteza odovarajućeg RNK[2] lanca. Transkripcija je postupak u kom se nukletotidi RNK uparuju sa komplementarnim nukleotidima jednog lanca DNK – adenin se uparuje sa uracilom, timin sa adeninom, citozin sa guaninom i guanin sa citozinom (*Slika 2).* Uracil je nukleotid koji se nalazi u RNK lancu umesto timina.



Slika 2. Transkripcija DNK lanca u RNK lanac

Nakon formiranja RNK lanca, procesom *translacije* nastaje lanac aminokiselina, odnosno proteinski lanac (ili proteinska sekvenca). Da bi se niz nukleotida "preslikao" u niz aminokiselina, neophodan je specijalan jezik, koji se naziva **genetički kod**. Naime, genetički kod je skup kodona – tripleta, koji su nastali kombinovanjem trojki nukleotida iz skupa adenin (A), uracil (U), citozin (C) i guanin (G). Broj kodona nastalih kombinovanjem četiri tipa nukleotida je 64 (4^3). Jedan kodon u lancu RNK određuje jednu aminokiselinu. Kako je broj mogućih kodona veći od broja aminokiselina koje

[1] Slika preuzeta sa http://en.wikipedia.org/wiki/Amino_acid
[2] RNK – ribonukleinska kiselina, posrednik u prenosu genetičkih informacija u procesu sinteze proteina.

grade protein, jednoj aminokiselini može odgovarati veći broj različitih kodona. Na primer, aminokiselina alanin je određena sa četiri kodona – GCU, GCC, GCA, GCG. Pored kodona koji određuju aminokiseline, postoje jos dve posebne vrste kodona – *start* i *stop* kodon. Start kodon je triplet nukleotida koji označava mesto u RNK lancu odakle treba da počne sinteza proteina, dok stop kodon označava kraj translacije. Start kodon je neophodan, jer translacija RNK potpuno zavisi od položaja sa kog počinje. Niz UCGACCUUGGUU može biti "pročitan" na više načina, u zavisnosti od toga sa kog mesta je translacija započeta:

– UCG → serin, ACC → treonin, UUG → leucin, GUU → valin

– CGA → arginin, CCU → prolin, UGG → triptofan, itd.

Start kodon predstavlja triplet AUG (koji je ujedno i kodon za metionin) i predstavlja mesto odakle translacija treba da počne. U *Tabeli 2* su prikazane aminokiseline sa kodonima koji ih određuju, kao i kodoni koji označavaju početak, odnosno kraj translacije:

Aminokiselina	Kodoni	Aminokiselina	Kodoni
Ala / A	GCU, GCC, GCA, GCG	Leu / L	UUA, UUG, CUU, CUC, CUA, CUG
Arg / R	CGU, CGC, CGA, CGG, AGA, AGG	Lys / K	AAA, AAG
Asn / N	AAU, AAC	Met / M	AUG
Asp / D	GAU, GAC	Phe / F	UUU, UUC
Cys /C	UGU, UGC	Pro / P	CCU, CCC, CCA, CCG
Gln / Q	CAA, CAG	Ser / S	UCU, UCC, UCA, UCG, AGU, AGC
Glu / E	GAA, GAG	Thr / T	ACU, ACC, ACA, ACG
Gly / G	GGU, GGC, GGA, GGG	Trp / W	UGG
His / H	CAU, CAC	Tyr / Y	UAU, UAC
Ile / I	AUU, AUC, AUA	Val / V	GUU, GUC, GUA, GUG
START	AUG	STOP	UAA, UGA, UAG

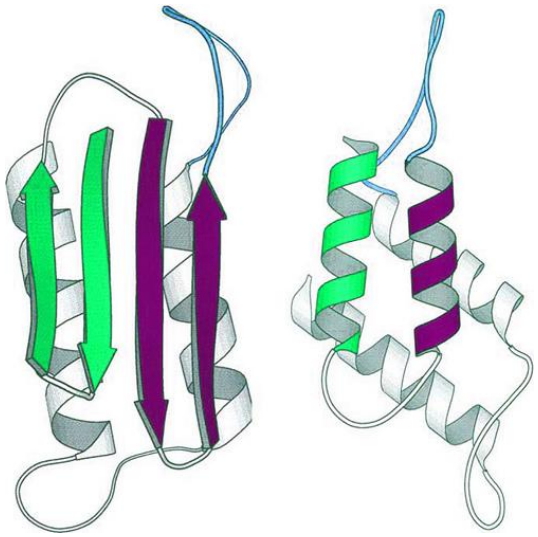
Tabela 2. Aminokiseline i kodoni koji ih određuju

Sekundarna struktura proteina određena je lokalnim konformacijama aminokiselina u proteinskim lancima. Obrasci koje grade vodonične veze u sklopu glavnog lanca definišu sekundarnu strukturu, pri čemu se zanemaruju veze između glavnog lanca i bočnih lanaca, kao i veze koje se javljaju između samih bočnih lanaca proteina. Sekundarna struktura proteina može imati različite oblike, koji se obično prepoznaju i klasifikuju kao sledeće vrste struktura:

1. G – spirala sa 3 zaokreta (3_{10}helix)
2. H – spirala sa 3,6 zaokreta (α–helix)
3. I – spirala sa 5 zaokreta (π–helix)
4. T – zavoj koji pravi vodonična veza

5. E – paralelne ili antiparalelne konformacije β–ravni, nazivaju se i “trake”
6. B – ostatak izolovan u β–mostu
7. S – krivina (jedina konformacija koja nije zasnovna na vodoničnoj vezi)
8. C – coil (R – ostaci koji ne pripadaju nijednoj od navedenih konformacija)[7]

α–heliks i β–ravni predstavljaju najprepoznatljivije i najpravilnije sekundarne strukture proteina, i pri tom su najzastupljenije u proteinima.



Slika 3. Sekundarne strukture proteina – β–ravni i α–helik – shematski prikaz [3]

Tercijarna struktura proteina predstavlja prostornu strukturu proteina, koja uključuje lokalne konformacije glavnog lanca proteina i sve interakcije koje nastaju između više različitih proteinskih lanaca koji čine jedan složen proteinski molekul.

Kvatenarna struktura proteina opisuje način na koji nekoliko polipeptidnih lanaca međusobno interaguje, i kao rezultat njihove interakcije nastaje funkcionalan protein.

[3] Slika preuzeta sa www.learner.org

2. Opis problema

Cilj ovog rada je analiza sastava delova proteinskih sekvenci koji neposredno utiču na građenje prostorne (tercijarne) strukture proteina. Predmet posmatranja su sekundarna struktura jedne sekvence i *interfejsi* koje ona gradi sa drugim sekvencama. Interfejse grade parovi aminokiselina koji su u direktnoj vezi u prostoru. Analiza obuhvata računanje koeficijenata korelacije između:

- aminokiselina i učešća u interfejsima,
- sekundarnih struktura i učešća u interfejsima.

Podaci, čiju analizu zahteva rešavanje ovog zadatka, podeljeni su u dve datoteke. Prva datoteka sadrži podatke o proteinskim sekvencama, dok druga sadrži podatke o interfejsima proteinskih sekvenci. Veliki obim podataka zahteva automatsku obradu, pa je u tu svrhu napisana konzolna aplikacija u programskom jeziku C#. U okviru aplikacije implementirani su priprema i statistička obrada podataka. Izlazni podaci su datoteke koje sadrže statistiku pojavljivanja aminokiselina (sekundarnih struktura) u interfejsima i van njih.

Početni podaci neophodni za ovaj rad su preuzeti iz baze podataka *PDB* (*Protein Data Bank*). PDB sadrži struktuirane podatke o makromolekulima kao što su proteini i nukleinske kiseline. Prema podacima iz avgusta 2015. godine (baza se ažurira na nedeljnom nivou), broj dostupnih podataka o proteinima je prešao sto hiljada [6]. Najčešće korišćena metoda za određivanje tercijarne strukture proteina koji se nalaze u ovoj bazi je metoda rendgenske kristalografije. Oko 90% podataka iz PDB baze je dobijeno ovom metodom. Ostali podaci su dobijeni primenom nuklearno – magnetno – rezonantne spektroskopije, elektronskom mikroskopijom itd.

Primenom algoritma *DSSP* (*Define Secondary Structure of Proteins*) nad PDB podacima, koji na osnovu vodoničnih veza i geometrijskih karakteristika ustanovljava sekundarnu strukturu proteina, dobijene su datoteke koje sadrže primarnu i sekundarnu strukturu proteinskih sekvenci [5]. Ovako dobijeni podaci predstavljaju osnovni materijal koji se koristi pri rešavanju problema koji će biti opisan u daljem tekstu.

2.1 ULAZNA DATOTEKA SA PROTEINSKIM SEKVENCAMA

Ulazna datoteka je zapisana u formatu DSV (*delimiter-separated values*) i čine je elementi PDBID, SEQ, SEQUENCE i SECONDARY STRUCTURE, pri čemu je karakter "|" separator. Struktura datoteke sa proteinskim sekvencama je prikazana na *Slici 4*.

```
PDBID SEQ        SEQUENCE                  SECONDARY STRUCTURE
1LOI| A | MPLVDFFCETCSKPWLVGWWDQFKR |  HHHHHHHTSS TTGGGGHHHHT
```

Slika 4. Struktura datoteke sa proteinskim sekvencama sa primerom proteinske sekvence

– ***PDBID***

Prva četiri karaktera (*PDBID*) predstavljaju jedinstveni identifikator proteina u *PDB*–u. Identifikator je kombinacija slova i brojeva i kao takav se koristi i u naučnoj literaturi za označavanje podataka koji su preuzeti iz *PDB*–a.

– ***SEQ***

Ovaj podatak može biti slovo ili broj i označava sekvencu proteina. Uređeni par (*PDBID, SEQ*) predstavlja identifikator proteinske sekvence u ovoj datoteci.

– ***SEQUENCE***

Niz karaktera koji predstavlja lanac aminokiselina koje grade proteinsku sekvencu (primarna struktura proteina). Aminokiseline su označene jednoslovnim oznakama po *IUPAC* nomenklaturi (*International Union Of Pure And Applied Chemistry Nomenclature*). Mala slova koja se pojavljuju u sekvencama predstvljaju retke modifikacije cisteina. Osim standardnih oznaka aminokiselina i malih slova, u zapisima proteinskih sekvenci se mogu javiti slova X i Z, i za takve aminokiseline su prepoznate sekundarne strukture koje te aminokiseline grade. Znak "!" koji se može naći u zapisima proteinskih sekvenci zamenjuje jednu ili više modifikovanih ili nepoznatih aminokiselina. Primer takve sekvence se nalazi na *Slici 5*.

```
2K6R|A|GQQYTA!IKGRTFRNEKELRDFIEKFXGR
      |      !  S   SSHHHHHHHHHHS
```

Slika 5. Proteinska sekvenca koja sadrži modifikovane aminokiseline

– ***SECONDARY STRUCTURE***

Niz karaktera koji predstavlja sekundarnu strukturu koju grade aminokiseline. Sekundarnu strukturu na položaju *i* gradi aminokiselina koja se u niski *SEQUENCE* nalazi na *i* – tom položaju. Ukoliko aminokiselina ne pripada nijednoj strukturi, na odgovarajućem položaju se nalazi blanko karakter.

2.2 ULAZNA DATOTEKA SA INTERFEJSIMA KOJE GRADE PROTEINSKE SEKVENCE

Za razliku od prethodne datoteke, gde su podaci za svaku proteinsku sekvencu smešteni u jedan red, u ovoj datoteci jedan red sadrži informacije o pojedinačnoj aminokiselini koja učestvuje u interfejsu. Struktura datoteke je prikazana na *Slici 6*.

– ***PDBID, SEQ***

Uređeni par koji predstavlja jedinstveni identifikator sekvence proteina kojoj pripadaju aminokiseline koje grade interfejs. Aminokiseline, koje pripadaju istoj sekvenci, u datoteci ne moraju biti zapisane jedna ispod druge.

```
PBDID SEQ RES_POS  AA
1A2P , C ,   32  ,ALA
1A2P , C ,   45  ,VAL
```

Slika 6. Sruktra datoteke koja sadrži interfejse proteinskih sekvenci

– ***RES_POS, AA***

Aminokiseline (*AA*) su u ovoj datoteci predstavljene troslovnom oznakom (*ALA – alanin, VAL – valin*). Uređeni par (*RES_POS, AA*) predstavlja aminokiselinu (*AA*) i položaj koji ona ima u sekvenci na osnovu podataka iz baze *PDB*. Položaji aminokiselina iz interfejsa u najvećem broju slučajeva odstupaju od njihovih stvarnih položaja u proteinskim sekvencama iz prve datoteke (1). Različita numeracija položaja aminokiselina u sekvenci i položaja atoma u proteinu u okviru datoteka iz PDB-a dovode do ovog problema. Položaj aminokiseline u proteinskoj sekvenci određuje njen položaj u niski koja predstavlja lanac aminokiselina te sekvence (*SEQUENCE*). Odstupanje položaja aminokiselina može da se javi iz nekoliko razloga.

Prvi razlog je taj što položaji aminokiselina u sekvencama u *PDB* bazi ne moraju da počinju od 0, tj. prva aminokiselina u sekvenci može da ima položaj čija je vrednost manja ili veća od 0.

Drugi razlog je izostavljanje modifikovanih amniokiselina iz sekvenci, koje se javlja u nekim podacima iz prve datoteke. To utiče na položaje aminokiselina koje se u lancu nalaze nakon izostavljenih aminokiselina, ali njihovi položaji u interfejsima nisu izmenjeni u skladu sa tim.

Treći razlog je to što u *PDB* bazi mogu da postoje aminokiseline čiji su položaji obeleženi istim brojem, ali je jedna od njih uz taj broj dodatno označena slovom (npr. 9 i 9A). U podacima iz ove datoteke postojanje slova je zanemareno, pa se u tom slučaju mogu pojaviti aminokiseline sa istim *RES_POS* brojem. Redosled kojim su takve aminokiseline navedene u datoteci nije od značaja, jer su poređane leksikografski.

Četvrti, i poslednji, razlog za odstupanje od stvarnih položaja u sekvencama jeste numerisanje dve susedne aminokiseline u sekvenci nesusednim brojevima.

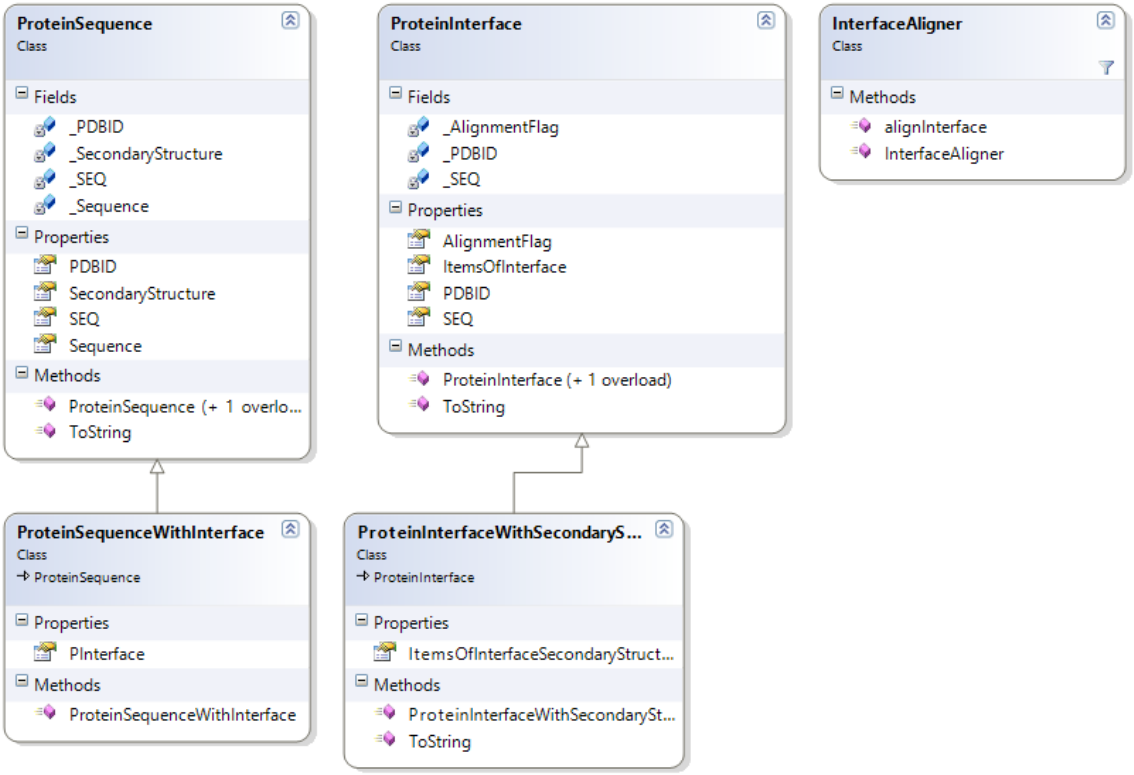
Sve ove činjenice otežavaju pronalaženje sekundarnih struktura aminokiselina koje učestvuju u interfejsima. Neke od navedenih problema je moguće rešiti, dok će neki podaci u analizi biti zanemareni zbog nedostataka koji ne mogu biti uklonjeni na jednoznačan način.

3. Implementacija

3.1 KLASE PODATAKA

Kao što se vidi iz opisa struktura datoteka i njihovog sadržaja, podaci najpre moraju biti pripremljeni i prilagođeni automatskoj obradi. Pre samog procesa pripreme podataka nameće se pitanje čitanja tih podataka i smeštanja u memoriju. Jedan od načina jeste definisanje klasa podataka, na osnovu formata podataka iz ulaznih datoteka.

Osnovna klasa u programu je klasa *ProteinSequence*. Objekti ove klase biće podaci pročitani iz ulazne datoteke sa proteinskim sekvencama. Kako se na osnovu podataka iz obe datoteke dolazi do zaključka da ne postoje podaci o interfejsima za sve proteinske sekvence, klasu *ProteinSequence* nasleđuje klasa *ProteinSequenceWithInterface*. Razlika u odnosu na baznu (roditeljsku) klasu jesu podaci o interfejsu koji grade neke aminokiseline iz proteinske sekvence. To su podaci koji se čitaju iz druge datoteke i koji se smeštaju u objekat klase *ProteinInterface*. Da bi koeficijenti korelacije mogli biti izračunati, neophodni su podaci o sekundarnim strukturama koje grade aminokiseline koje su deo interfejsa. Međutim, u datoteci sa podacima o interfejsima proteinskih sekvenci ti podaci nedostaju. Pronalaženje poravnanja interfejsa u odnosu na proteinsku sekvencu omogućava pronalaženje tih sekundarnih struktura. Način za pronalaženje sekundarih struktura aminokiselina iz interfejsa, tj.pronalaženje poravnanja, biće opisano u okviru odeljka o pripremi podataka. Interfejsi za koje su pronađene sekundarne strukture se predstavljaju objekte klase *ProteinInterfaceWithSecondaryStructure*, koja nasleđuje klasu *ProteinInterface*. Dijagram klasa je prikazan na *Slici 7.*



Slika 7. Dijagram klasa podataka

3.1.1 Klasa ProteinSequence

Klasa ima četiri atributa:

- *PDBID (private String _PDBID)* – identifikator proteina (kolona *PDBID* iz ulazne datoteke);
- *SEQ (private char _SEQ)* – oznaka sekvence proteina (kolona *SEQ*);
- *Sequence (private String _Sequence)* – niz karaktera koji predstavlja lanac aminokiselina koje grade proteinsku sekvencu (*SEQUENCE*);
- *SecondaryStructure (private String _SecondaryStructure)* – niz karaktera koji predstavlja sekundarnu strukturu proteinske sekvence (*SECONDARY STRUCTURE*).

Pored atributa, klasa ima dva konstruktora:

- *ProteinSequence()* – Podrazumevani konstruktor bez parametara;
- *ProteinSequence(ProteinSequence p)* – Konstruktor koji pravi objekat identičan objektu *p*.

Jedini metod implementiran u ovoj klasi je metod *ToString()*, koji vraća zapis proteinske sekvence u obliku kao u ulaznoj datoteci:

PDBID|SEQ|SEQUENCE|SECONDARY STRUCTURE

3.1.2 Klasa ProteinSequenceWithInterface

Klasa *ProteinSequenceWithInterface* nasleđuje klasu *ProteinSequence*. U odnosu na baznu klasu sadrži svojstvo *PInterface* koje predstavlja objekat klase *ProteinInterface*. Klasa ima jedan konstruktor sa parametrima *ProteinSequence* i *ProteinInterface*. Za objekat tipa *ProteinSequence* se poziva bazni konstruktor, dok se svojstvo *PInterface* instancira prosleđivanjem objekta tipa *ProteinInterface* konstruktoru klase.

3.1.3 Klasa ProteinInterface

Objekti ove klase se predstavljaju podatake iz datoteke o intefrejsima koje grade proteinske sekvence. Klasa sadrži četiri atributa:

- *PDBID (private String _PDBID)* – identifikator proteina kome pripada interfejs;
- *SEQ (private char _SEQ)* – oznaka sekvence proteina;
- *AlignmentFlag (int _AlignmentFlag)* – Podrazumevana vrednost koja se dodeljuje ovom svojstvu je 0. Ova vrednost se postavlja na 1 ukoliko je za interfejs iz ovog objekta pronađeno poravnanje, odnosno pronađene su tačni položaji aminokiselina u odgovarajućoj sekvenci.

Četvrti atribut *ItemsOfInterface* je tipa *SortedDictionary<int,char>* i svaki član ove strukture je uređeni par položaj – aminokiselina (*<POS_RES, AA>)*.

Klasa sadrži dva konstruktora:

- *ProteinInterface()* – Konstruktor bez parametara u kome se instancira *ItemsOfInterface*;
- *ProteinInterface(ProteinInterface pi)* – Konstruktor koji pravi objekat identičan objektu *pi*.

Metod *ToString()* je implementiran tako da se podaci, koje sadrži objekat ove klase, ispisuju u formatu koji je istovetan zapisu iz ulazne datoteke – svaka aminokiselina koja učestvuje u interfejsu se zapisuje kao poseban red oblika:

PDBID, SEQ, RES_POS, AA

3.1.4 Klasa ProteinInterfaceWithSecondaryStructure

Klasa koja nasleđuje klasu *ProteinInterface*. Razliku u odnosu na baznu klasu čini svojstvo *ItemsOfInterfaceSecondaryStructure*. Kao i *ItemsOfInterface*, ovo svojstvo je tipa *SortedDictionary<int,char>*, ali njegove elemente čine uređeni parovi *<RES_POS, SS>*, gde je *SS* oznaka sekundarne strukture koju gradi aminokiselina koja se nalazi na položaju *ItemsOfInterface[RES_POS]*. Klasa sadrži konstrukor koji poziva konstruktor bazne klase i instancira svojstvo *ItemsOfInterfaceSecondaryStructure*. Pored ovog konstruktora, postoji i konstruktor koji kao argument ima objekat ove klase i pravi objekat identičan prosleđenom. Metod *ToString()* svaku aminokiselinu koja učestvuje u interfejsu zapisuje kao poseban red oblika:

PDBID, SEQ, RES_POS, AA,SS.

3.2 Pripremanje podataka

Priprema podataka obuhvata proveru ulaznih podataka i pripremu podataka za statističku obradu. Priprema se odvija u 3 koraka:

1. Obrada datoteke sa proteinskim sekvencama
2. Obrada datoteke sa interfejsima proteinskih sekvenci
3. Pronalaženje poravnanja interfejsa

3.2.1 Obrada datoteke sa proteinskim sekvencama

Prilikom opisa strukture datoteke rečeno je da u niski, koja predstavlja lanac aminokiselina koje grade proteinsku sekvencu, postoje modifikacije cisteina čije oznake odstupaju od *IUPAC* nomenklature. *Cistein* je u tim slučajevima označen malim latiničnim slovom i ove karaktere treba zameniti velikim slovom *C*, da bi svi *cisteini* imali jedinstvenu oznaku. Ovaj korak ne utiče na ostale nestandardne znakove (X, Z, !) koji se mogu javiti u sekvenci.

Metod koji je implementiran u svrhu obrade ove datoteke zove se *CheckProteinSequenceFile()* i kao parametre prihvata dve niske, od kojih je prva putanja do datoteke koja se proverava, a druga ime te datoteke. Metod se zasniva na učitavanju podataka u listu objekata klase *ProteinSequence* i proveri da li proteinska sekvenca sadrži aminokiselinu *cistein* zapisanu kao malo latinično slovo. Ukoliko to jeste slučaj, malo latinično slovo se zamenjuje slovom *C*. Osim izmena u sekvencama, blanko karakteri u sekundarnim strukturama će biti zamenjeni karakterom *C*. Izvorna datoteka

se ne menja, već se pravi nova datoteka *ProteinFile_Checked.dat* sa istovetnom strukturom.

Prethodni metod zahteva implementaciju metoda *ReadProteinSequenceFile()* za učitavanje proteinskih sekvenci u listu objekata klase *ProteinSequence*. Parametar koji se prosleđuje ovom metodu je tok za čitanje podataka iz ulazne datoteke (*StreamReader*), dok je povratna vrednost metoda lista objekata klase *ProteinSequence* (*List<ProteinSequence>*). Metod učitava tekst i svaku liniju teksta deli na četiri niske, koje su međusobno odvojene znakom „|". Svaka od niski se smešta u odgovarajuće polje objekta klase *ProteinSequence.*

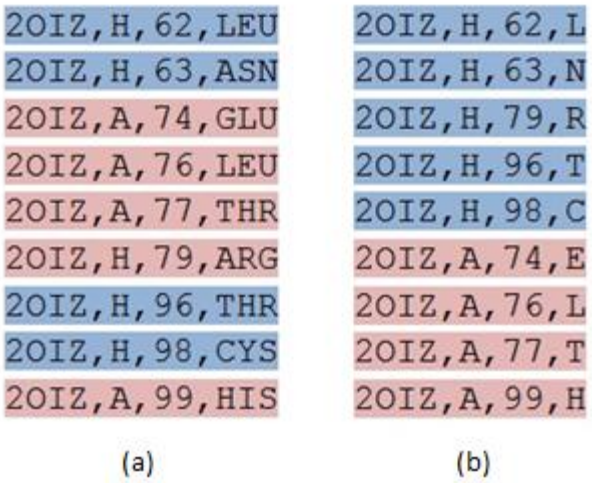
3.2.2 Obrada datoteke sa interfejsima proteinskih sekvenci

Ovaj korak obuhvata rešavanje tri problema koja se javljaju u datoteci sa interfejsima. Prvi problem je označavanje aminokiselina troslovnim oznakama po *IUPAC* nomenklaturi. Zbog usaglašavanja oznaka aminokiselina iz interfejsa sa oznakama aminokiselina iz proteinskih sekvenci, troslovne oznake biće zamenjene odgovarajućim jednoslovnim. Izuzev standardnih aminokiselina, kojih ima 20, u podacima mogu da se jave aminokiseline kao npr. *GLX* (*GLN/GLU* – Glutamin/Glutaminska kiselina). Sve aminokiseline koje odstupaju od *UIPAC* nomenklature, ili iz nekog razloga nisu mogle biti prepoznate biće ignorisane, jer njihov mali broj ne utiče na statistiku, pa nije od značaja za ovaj rad.

Drugi problem predstavljaju aminokiseline sa istim *RES_POS* brojem, koje pripadaju istoj sekvenci. Kao što je već rečeno, poredak aminokiselina koje spadaju u ovu grupu je leksikografski i na osnovu raspoloživih informacija se ne može utvrditi njihov tačan redosled u sekvenci. Za interfejse kod kojih se javlja ovaj problem (takvih podataka nema mnogo) poravnanje ne može biti pronađeno, pa se oni izostavljaju iz statistike.

Treći problem su aminokiseline koje pripadaju jednoj sekvenci, a nisu zapisane u datoteci jedna ispod druge (*Slika 8 (a)*). Ideja je da se čitanjem ove datoteke za svaku proteinsku sekvencu, koja ima interfejs, instancira objekat tipa *ProtienInterface()* i da taj objekat sadrži podatke o svim aminokiselinama iz interfejsa . Ako bi se podaci čitali red po red, i ako bi svako neslaganje uređenog para *(PDBID, SEQ)* sa uređenim parom iz prethodnog reda značilo pravljenje novog objekta, u toku izvršavanja programa bi se instanciralo više objekta sa istim *(PDBID, SEQ)* vrednostima. Ovo bi značilo da se prilikom pronalaženja poravnanja više puta pretražuje ista proteinska sekvenca, što potencijalno pravi i problem sa pronalaženjem više mogućih poravnanja[4]. Da bi se izbeglo bespotrebno instanciranje više objekata sa istim uređenim parom *(PDBID, SEQ)* i da bi se omogućila samo jedna pretraga proteinske sekvence neophodno je grupisati podatke tako da se aminokiseline jedne sekvence nalaze jedna ispod druge (*Slika 8 (b)*).

[4] Detaljnije o ovom problemu biće reči u odeljku o pronalaženju poravnanja interfejsa



Slika 8. Datoteka sa interfejsima pre (a) i posle obrade(b) – primer

Metod *CheckProteinInterfaceFile()* ima ulogu da podatke o interfejsima obradi na prethodno opisan način. Parametar koji se prosleđuje ovom metodu jeste tok za čitanje podataka iz datoteke. Metod ne menja ulaznu datoteku, već se pomoću toka za pisanje pravi nova datoteka (*iface_sorted.csv*) iz koje će biti učitani podaci za pronalaženje poravnanja u koraku 3.

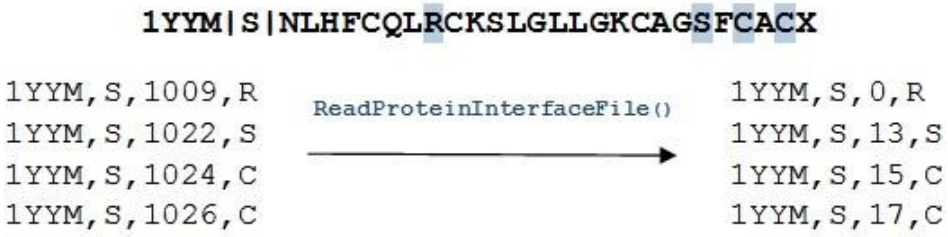
Pored metoda za obradu datoteke implementiran je i metod za učitavanje interfejsa iz datoteke. Metod *ReadProteinInterfaceFile()* se koristi za čitanje datoteke koja je rezultat izvršavanja prethodno opisanog metoda, a kao povratnu vrednost ima listu objekata klase *ProteinInterface*. Parametar koji se prosleđuje metodu je tok za čitanje podataka. Metod je implementiran tako da čita red po red iz datoteke. Učitavanjem prvog reda pravi se objekat klase *ProteinInterface*, dodaje se u listu, a vrednosti *PDBID* i *SEQ* postavljaju na vrednosti učitane iz tekućeg reda. Prilikom ubacivanja aminokiselina u atribut *ItemsOfInterface*, njihove vrednosti će biti promenjene na sledeći način:

- prva pročitana aminokiselina imaće položaj 0;
- ostalim aminokiselinama dodeljuju vrednosti njihove položaje umanjene za *RES_POS* vrednost prve pročitane aminokiseline.

ItemsOfInterface se popunjava učitanim podacima sve dok se ne pročita red u kome se vrednost PDBID i SEQ razlikuju od odgovarajućih polja u trenutnom *ProteinInterface* objektu. U tom trenutku se pravi novi *ProteinInterface* objekati i na prethodno opisan način nastavlja dalje čitanje datoteke.

Razlog za korigovanje vrednosti položaja aminokiselina nalazi se u tome što *RES_POS* vrednosti u ulaznoj datoteci često odstupaju od stvarnih položaja aminokiselina u proteinskim sekvencama, kao što je već rečeno u opisu strukture ove datoteke (2.1). Označavanje položaja aminokiselina u proteinskim sekvencama počinje

od nule (sekvence se čuvaju u objektu tipa *String*, pa je položaj prve aminokiseline *Sequence[0]*), pa je, zbog pretrage sekvence kod pronalaženja poravnanja, najmanja moguća vrednost položaja prve aminokiseline iz interfejsa 0. Položaji ostalih aminokiselina se menjaju u skladu sa promenom položaja prve aminokiseline i na taj način se čuvaju rastojanja između aminokiselina. Na *Slici 9* je prikazan primer odstupanja položaja u podacima o interfejsima u odnosu na podatke o proteinskoj sekvenci i vrednosti koje podaci dobijaju nakon korekcije primenom metoda *ReadProteinInterfaceFile()*.



Slika 9. Odstupanje položaja aminokiselina iz interfejsa od stvarnih položaja u sekvenci

3.3 PRONALAŽENJE PORAVNANJA INTERFEJSA

Poslednji korak u pripremi podataka jeste pronalaženje poravnanja za aminokiseline iz interfejsa u odnosu na proteinske sekvence. Ovaj korak je neophodan, jer se samo nakon pronalaženja poravnanja mogu ispravno odrediti sekundarne strukture koje grade aminokiseline iz interfejsa.

Postojanje oznaka X i Z u sekvencama nema uticaj na pronalaženje poravnanja, dok pojavljivanje znaka "!" ima bitnu ulogu u ovom postupku. Bitno je naglasiti da se pretpostavlja da znak „!" ne može da zameni deo interfejsa, jer u interfejsima mogu da se nađu samo aminokiseline koje poznate.

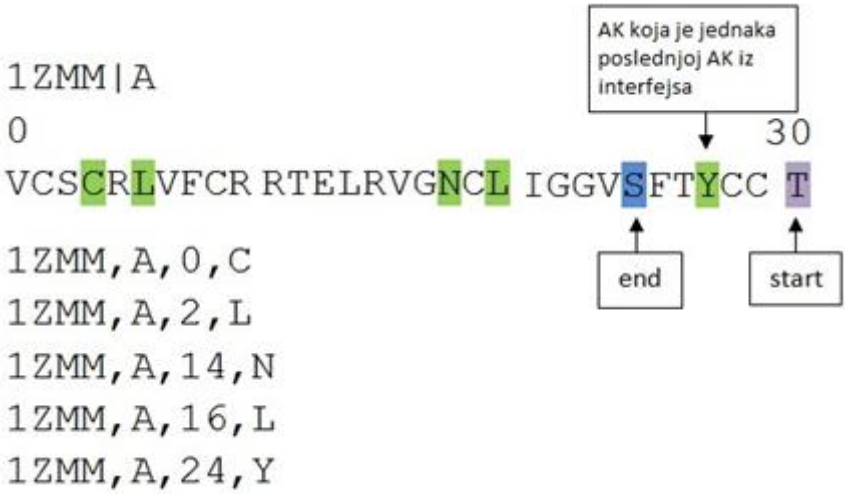
Sekvence su podeljene u dve grupe, jer postoje razlike u načinu pronalaženja poravnanja za interfejse:

- sekvence koje ne sadrže znak „!" u lancu aminokiselina – to su sekvence u kojima ne postoje modifikovane aminokiseline za koje nije pronađena sekundarna struktura
- sekvence koje sadrže znak „!", odnosno u lancu aminokiselina postoji jedna ili više modifikovanih aminokiselina koje nemaju sekundarnu strukturu.

Za pronalaženje poravnanja kod proteinskih sekvenci kod kojih ne postoje modifikovane aminokiseline obeležene znakom "!", koristi se algoritam koji je vrlo jednostavan. Da bi se izbegla pretraga cele proteinske sekvence, potrebno je pronaći položaj na kojome pretraga počinje i položaj na kojme se završava.

Početno mesto pretrage (*start*) je položaj poslednje aminokiseline u sekvenci (*sequence.length – 1*). Kako je u interfejsu položaj prve aminokiseline postavljen na 0, krajnje mesto pretrage sekvence (*end*) će biti na položaju koji ima vrednost položaja poslednje aminokiseline iz interfejsa. Pretragom sekvence u segmentu (*end, start*) traži se položaj *j* na kome se nalazi aminokiselina jednaka poslednjoj aminokiselini u interfejsu. Kada se pronađe takav položaj *j*, neophodno je proveriti da li se u sekvenci mogu pronaći preostale aminokiseline iz interfejsa.

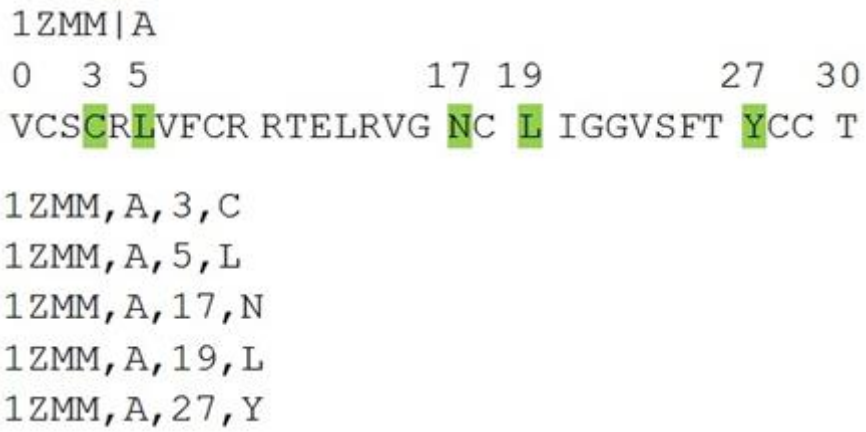
Smatra se da je poravnanje pronađeno ako u sekvenci postoje sve aminokiseline iz interfejsa pronađene na osnovu podataka o njihovim položajima. Sve aminokiseline iz interfejsa će biti uključene u statističku obradu samo ako je pronađeno poravnanje jedinstveno. Nastavak pretrage od prvog položaja koja prethodi onoj na kojoj je pronađeno prvo poravnanje (*j–1*), omogućava proveru jedinstvenosti pronađenog poravnanja. Na isti način se nastavlja pretraga ukoliko se ne pronađe poklapanje sa svim aminokiselinama iz interfejsa. Pretraga se zaustavlja na položaju *end*, i ako nijedno poravnanje do tog trenutka nije pronađeno, pretraga je neuspešna i podaci iz interfejsa ne ulaze u statistiku.



Slika 10. Pretraga proteinske sekvence koja ne sadrži modifikovane aminokiseline označene znakom "!"

Na *Slici 10* je prikazan primer pronalaženja poravnanja za proteinsku sekvencu A proteina *1ZMM*. Pretraga počinje od aminokiseline koja se u sekvenci nalazi na položaju 30. Sve dok se ne pronađe aminokiselina jednaka poslednjoj aminokiselini u interfejsu

(*tirozin – Y*) pretraga se pomera za jedno mesto u levo. Prvi položaj na kome se nalazi *tirozin* jeste položaj 27. Sledeća aminokiselina koja se proverava je *leucin* (*L*), koji se, prema podacima iz interfejsa, nalazi osam mesta levo od *tirozina*. Aminokiselina koja se u sekvenci nalazi na osam mesta levo od tirozina jednaka je aminokiselini iz interfejsa, pa se na isti način nastavlja provera svih ostalih aminokiselina. Poravnanje pronađeno za ovu proteinsku sekvencu je prikazano na *Slici 11*.



Slika 11. Poravnanje pronađeno pretragom sekvence A proteina 1ZMM

Kako položaj 24 predstavlja kraj pretrage za potencijalno pronalaženje poravnanja, proverom aminokiselina na segmentu (24, 27) se ne nalazi nijedan *tirozin*, pa je pronađeno poravnanje jedinstveno.

Pronalaženje poravnanja kod proteinskih sekvenci koje sadrže modifikovane aminokiseline obeležene znakom „!“ u unutrašnjosti lanca je nešto komplikovanije od prethodno opisanog. U slučaju sekvenci koje ne sadrže modifikovane aminokiseline obeležene znakom „!“ pretraga se zaustavlja na položaju koji je jednak vrednosti položaja poslednje aminokiseline iz interfejsa. Kod sekvenci sa modifikovanim aminokiselinama svako pojavljivanje znaka „!“ u sekvenci između aminokiselina koje pripadaju interfejsu otvara prostor za novu pretragu, jer aminokiseline koje se nalaze levo od znaka „!“ nemaju jedan mogući položaj, već njihov položaj zavisi od broja aminokiselina koje su zamenjene znakom „!“. Zbog toga je neophodno da se za svaku aminokiselinu iz interfejsa odredi vrednost minimalnog (krajnjeg) položaja na kojoj se ta aminokiselina može nalaziti, ali tako da sve aminokiseline koje joj prethode mogu biti pronađene u sekvenci. Osim minimalnih položaja aminokiselina, neophodno je odrediti početno i krajnje mesto pretrage. Implementacija metoda koji se koristi pronalaženje liste minimalnih položaja, kao i početnog i krajnjeg mesta pretrage:

```
private void setStartAndEndSearchIndexes()
{
    minimumPositionInSequence = proteinWithInterface.PInterface.ItemsOfInterface.Keys.ToList();
    startSearch = proteinWithInterface.Sequence.Length - 1;
    endSearch = minimumPositionInSequence.Last();

    if (proteinWithInterface.Sequence.Contains('!'))
    {
        exclamationMarkPositionsList = exclamationMarkPositions(proteinWithInterface);
        if (endSearch > exclamationMarkPositionsList[0])
           foreach (int i in exclamationMarkPositionsList)
           {
               if (minimumPositionInSequence.Exists(x => x >= i))
               {
                   int nextAA = minimumPositionInSequence.First(x => x >= i);
                   int j = minimumPositionInSequence.IndexOf(nextAA);
                   int shift = i - nextAA + 1;
                   while (j < minimumPositionInSequence.Count())
                   {
                       minimumPositionInSequence[j] +=shift;
                       j++;
                   }
               }
           }
         endSearch = minimumPositionInSequence.Last();
    }
}
```

Početak pretrage (*start*) je i u ovom slučaju vrednost položaja poslednje aminokiseline u sekvenci. Inicijalno, kraj pretrage je poslednji element liste minimalnih položaja.

Na početku algoritma lista minimalnih položaja (*minimumPositionInSequence*) jednaka je listi korigovanih položaja aminokiselina iz interfejsa[5]. Ukoliko su sve te vrednosti manje od položaja prvog znaka „!“ u sekvenci, onda su to ujedno i minimalni (tj. krajnji levi) položaji na kojima se te aminokiseline mogu nalaziti. Ako to nije slučaj, svako pojavljivanje znaka „!“ u sekvenci će uticati na korigovanje minimalnih položaja svih aminokiselina koje se nalaze desno od njega. Prva sledeća aminokiselina će imati krajnji levi položaj jednak položaju znaka „!“ uvećanog za 1, a svi ostali položaji će biti korigovani u skladu sa tim. Nakon pronalaženja minimalnih položaja aminokiselina, novi kraj pretrage će biti poslednja vrednost iz korigovane liste *minimumPositionInSequence.*

Pretragom segmenta (*end, start*) u sekvenci traži se položaj *j* na kome se nalazi poslednja aminokiselina iz interfejsa. Kada se takav položaj pronađe, proverava se da li u sekvenci mogu biti pronađene sve preostale aminokiseline iz interfejsa ili bar deo njih. Ako se pronađe poravnanje za sve aminokiseline iz interfejsa, proverava se jedinstvenost tog poravnanja. Ako je pronađeno poravnanje za deo interefejsa, neophodno je proveriti da li se između poslednje aminokiseline za koju je pronađeno poravnanje (*AK[i+1]*) i prve sledeće za koju je pronalazak poravnanja bio neuspešan (*AK[i]*) nalazi znak „!“. Ukoliko to nije slučaj, pronalazak poravnanja na položaju *j* je neuspešan, i nastavlja se dalja pretraga sekvence na segmentu (*end, j-1*). Ako se između *AK[i] i AK[i+1]* nalazi znak „!“, trenutne vrednosti (*end, j*) se čuvaju, jer će one biti neophodne u slučaju provere jedinstvenosti poravnanja (ako poravnanje bude pronađeno), ili u slučaju neuspeha za nastavak dalje pretrage. Na prethodno opisan način počinje nova pretraga

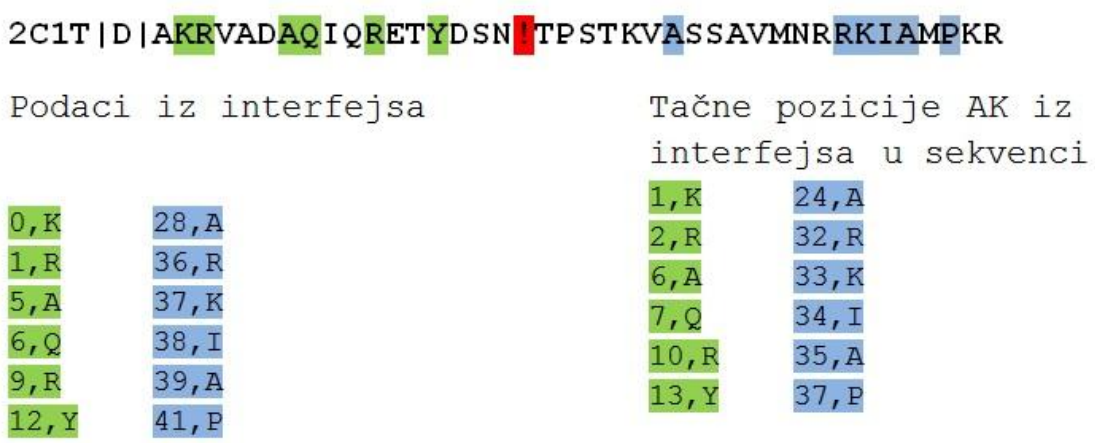
[5] Razlozi za korekciju položaja aminokiselina u interfejsima su objašnjeni u odeljku 3.2.2

sekvence za aminokiseline iz interfejsa za koje poravnanje nije pronađeno (AK*[0..i]*). Novo početno mesto pretrage (*start*) je položaj *j-1*, a krajnje mesto pretrage (*end*) je *max*(*minimumPositionsInSequence[i], očekivani položaj AK[i] u sekvenci*). Očekivani položaj *AK[i]* u sekvenci je položaj na kome se očekuje pronalazak aminokiseline *AK[i]* na osnovu podataka iz interfejsa, ali se to ne dešava, zbog pojave znaka „!“. Ovaj postupak se ponavlja sve dok se ne pronađe poravnanje, ili dok se ne dođe do krajnjeg mesta pretrage bez uspešno pronađenog poravnanja.

U oba slučaja, pretraga se vraća unazad, dok je to moguće, odnosno dok postoje sačuvane granice prethodnih pretraga. Sva uspešno pronađena poravnanja se čuvaju. Ukoliko je pronađeno samo jedno poravnanje, sve aminokiseline iz interfejsa postaju deo podataka nad kojima se vrši statistička analiza.

Kod višestrukih poravnanja pronađenih u ovom tipu proteinskih sekvenci, mogu da postoje delovi interfejsa za koje je pronađeno jedinstveno poravnanje. Takvi delovi interfejsa će biće uključeni u obradu, dok će one aminokiseline koje nemaju jedinstveno određen položaj biti odbačene.

Na *Slici 12* je prikazan primer pretrage sekvence D proteina 2C1T sa modifikovanim aminokiselinama obeleženim znakom „!“ i interfejsa. Pretraga počinje od poslednje aminokiseline u sekvenci koja se nalazi na položaju 39, dok je kraj pretrage poslednja vrednost u listi *minimumPositionsInSequence* – 31. Prvi položaj na kome se nalazi aminokiselina koja je jednaka poslednjoj aminokiselini iz interfejsa (*prolin*) je 37. Daljom pretragom sekvence pronalazi se poravnanje za aminokiseline koje prema rastojanjima iz interfejsa imaju položaje koji imaju veću vrednost od položaja na kome se nalazi znak „!“ (na *Slici 10* su označeni plavom bojom). Poslednja takva aminokiselina je *alanin* (24). Sledeća aminokiselina u interfejsu je *tirozin* (*Y*), koji prema podacima iz interfejsa mora da se nalazi 16 mesta levo od *alanina (24)*. Na tom položaju se ne nalazi *tirozin*, već *izoleucin* (*I*). Do ovakvog neslaganja došlo je zbog znaka „!“ koji je se nalazi na segmentu sekvence od položaja 8 do položaja 24. Znak „!“ označava da se na tom položaju nalazi jedna ili više modifikovanih aminokiselina, pa se položaji aminokiselina iz interfejsa levo od ovog znaka moraju pomeriti određeni broj mesta udesno (pomeranje se vrši za slučaj da je broj aminokiselina koje nedostaju veći od 1 – znak „!“ menja jednu aminokiselinu, pa pomeranje u tom slučaju nije potrebno). Kako ne postoji način da se na osnovu podataka unapred utvrdi broj aminokiselina koje su izostavljene, pamte se vrednosti granica tekuće pretrage (31, 37) i započinje se nova pretraga za preostale aminokiseline iz interfejsa. Nove vrednosti početka i kraja pretrage su 17 i 12. Prvo, i jedino, pojavljivanje *tirozina* na ovom segmentu je na položaju 13. Poravnanje za ostale aminokiseline se pronalazi u skladu sa rastojanjima iz interfejsa (aminokiseline označene zelenom bojom *Slika 12*).



Slika 12. Primer pronalaženja poravnanja za proteinske sekvence sa modiikovanim aminokiselinama obeleženim znakom "!"

Provera jedinstvenosti poravnanja u ovom primeru ima dva koraka „unazad". Prvi je provera da li je aminokiselina na položaju 12 *tirozin*. Kako se na položaju 12 ne nalazi *tirozin* i ne postoji više nijedan položaj za proveru, zaključuje se da u ovom slučaju više nema poravnanja za ovaj interfejs. Nakon ovoga pretrga se vraća još jedan korak unazad, na segment [31, 37] i proverava se da li na tom segmenti postoji aminokiselina *prolin*. Kako na ovom segmentu ne postoji nijedna takva aminokiselina, na osnovu toga se zaključuje da je prvobitno pronađeno poravnanje jedinstveno.

Implementacija klase *InterfaceAligner* je zasnovana na dva, prethodno opisana, načina za pronalaženje poravnanja. Parametar koji se prosleđuje konstruktoru klase je lista objekata klase *ProteinWithInterface* (proteini koji imaju interfejs), lista objekata *ProteinInterfaceWithSecondaryStructure*. Osim ovih parametara, konstruktoru se prosleđuju i tri toka za pisanje podataka. Glavni metod klase, *alignInterface* obrađuje prosleđene podatke i rezultate smešta u objekte klase *ProteinWithInterface*. Pomoću prosleđenih tokova rezultati se upisuju u formatu:

PDBID,SEQ,RES_POS,AA,SS

i to:

- u datoteku sa podacima o interfejsima proteina za koje je metod pronalaženja poravnanja bio uspešan (potpuno ili delimično jedinstveno poravnanje) (*alignedInterfaces.csv*),
- u datoteku sa podacima o interfejsima ili delovima interfejsa proteina za koje je pronađeno više mogućih poravnanja (*multipleInterfaceAlignments.csv*).

Proteinske sekvence za koje je postupak pronalaženja poravnanja bio neuspešan zapisane su u datoteci *interfaceWithNoAlignmentsOrWithMultipleAlignments.csv*.

Kako se pronalaženje poravnanja zasniva isključivo na rastojanjima između aminokiselina koje učestvuju u interfejsu, očekivano je da interfejsi sa malim brojem aminokiselina (jedna ili dve aminokiseline) imaju više mogućih poravnanja. Na *Slici 13* prikazana je sekvenca A proteina 2ZI0, u kojoj samo *leucin* (*L*) učestvuje u interfejsu.

Slika13. Višestruko poravnanje za interfejs sekvence koja ne sadrži modifikovane aminokiseline obeležene znakom "!"

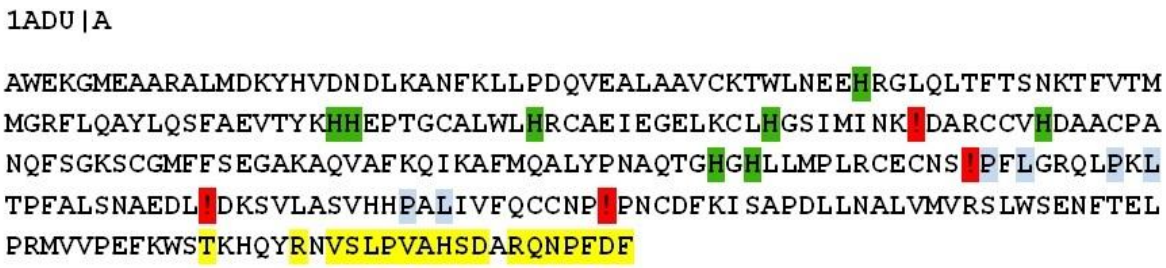
U datoteci sa podacima o interfejsima položaj na kome se nalazi *leucin* u sekvenci ima vrednost 15. Na osnovu podataka koje imamo, tačan položaj ove aminokiseline nije moguće utvrditi i zbog toga će ovi podaci biti izostavljeni iz satatistike.

Višestruka poravnanja se mogu naći i kod interfejsa u kojima učestvuje veći broj aminokiselina. Takav interfejs se javlja u slučaju sekvence A proteina 1ADU. Interfejs se sastoji od 21 aminokiseline:

1ADU,A,262,H	1ADU,A,515,L	1ADU,A,523,R
1ADU,A,406,P	1ADU,A,516,P	1ADU,A,524,Q
1ADU,A,408,L	1ADU,A,517,V	1ADU,A,525,N
1ADU,A,506,T	1ADU,A,518,A	1ADU,A,526,P
1ADU,A,511,R	1ADU,A,519,H	1ADU,A,527,F
1ADU,A,513,V	1ADU,A,520,S	1ADU,A,528,D
1ADU,A,514,S	1ADU,A,521,D	1ADU,A,529,F

Na *Slici 14* prikazana je proteinska sekvenca na kojoj su obeleženi svi mogući položaji aminokiselina iz interfejsa.

Aminokiseline *prolin* (*P*) i *leucin* (*L*) imaju tri položaja na kojima mogu da se nalaze, u zavisnosti od broja aminokiselina koje su zamenjene znakom „!". Mogući položaji (označeni plavom bojom na *Slici 14*) su (182, 184), (189, 191) i (214, 216). Mogući položaji *histidina* (*H*) u odnosu na položaje *prolina* i *leucina* su: 82, 83, 93, 106, 121, 167 i 169. *Histidin*, osim na ovim položajima, može da se nađe i na položaju 47, ali samo u slučaju da su *prolin* i *leucin* na položajima (182, 184) ili (189, 191). Kao i u prethodnom slučaju, položaj *histidina* zavisi od broja aminokiselina koje su zamenjene znakom „!". Ukupan broj mogućih poravnanja je 23.



Slika 14. Višestruko poravnanje za interfejs sekvence koja sadrži modifikovane aminokiseline obeležene znakom "!"

Na osnovu svih pronađenih poravnanja zaključuje se da su položaji prve tri aminokiseline su problematični, dok se za preostalih osamnaest pronalaze jedinstveni položaji (aminokiseline označene žutom bojom) kao takvi mogu biti deo statističke obrade.

Pored interfejsa sa više mogućih poravnanja, postoje i interfejsi kod kojih nije moguće pronaći poravnanje. To su interfejsi kod kojih su susedne aminokiseline iz sekvence označene nesusednim brojevima.

3.4 IZLAZNE DATOTEKE

Do sada je bilo reči o pripremi podataka za statističku obradu. Da bi koeficijenti korelacije bili izračunati, za prethodno obrađene podatke biće izračunate vrednosti:

1. Za svaku pojedinačnu aminokiselinu A - izlazna datoteka *AA_Interface_statisticMatrix*:
   - Broj aminokiselina A koje grade interfejs – a
   - Broj aminokiselina A koje ne grade interfejs – b
   - Broj aminokiselina različitih od A (u oznaci A*) koje grade interfejs – c
   - Broj aminokiselina različitih od A koje ne grade interfejs – d

	In Interface	Out Of Interface
A	a	b
A*	c	d

Tabela 3. Struktura izlazne datoteke za aminokiseline

2. Za svaku pojedinačnu sekundarnu strukturu S - izlazna datoteka *SS_Interface_statisticMatrix*:
   - Broj sekundarnih struktura S koje grade interfejs – a
   - Broj sekundarnih struktura S koje ne grade interfejs – b
   - Broj sekundarnih struktura različitih od S (u oznaci S*) koje grade interfejs – c
   - Broj sekundarnih struktura različitih od S koje ne grade interfejs – d

	In Interface	Out Of Interface
S	a	b
S*	c	d

Tabela 4. Struktura izlazne datoteke za sekundarne strukture

Suma vrednosti (a + b + c + d) jednaka je ukupnom broju aminokiselina, odnosno sekundarnih struktura u uzorku.

Izlazne datoteke, čije su strukture prikazane u *Tabeli 3* i *Tabeli 4* su u *.csv* (*comma – separated values*) formatu, ali će zbog jednostavnijeg računanja koeficijenta korelacije biti prebačene u *Excel* format.

3.5 KOEFICIJENT KORELACIJE

Koeficijent korelacije predstavlja meru povezanosti (zavisnosti) dve promenljive. Vrednosti koeficijenta korelacije pripadaju opsegu [-1, 1]. Ako je vrednost koeficijenta -1, odnosno 1, tada su podaci linearno zavisni. Kada je vrednost koeficijenta 0, tada među podacima ne postoji nikakva povezanost. U ovom radu podaci koji se koriste za računanje koeficijenta korelacije su binarni podaci. Formula za koeficijent korelacije kada su obe slučajne promenljive binarne [9]:

$$\rho(X,Y) = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sqrt{\bar{X}(1-\bar{X})\bar{Y}(1-\bar{Y})}} \quad (1)$$

Kako je formulu (1) neophodno primeniti na podatke čije su strukture prikazane u *Tabeli 3 i Tabeli 4*, formula se svodi na računanje koeficijenta φ [8]:

$$\varphi = \frac{(a*d) - (b*c)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (2)$$

4. Rezultati

Ukupan broj proteinskih sekvenci u ulaznoj datoteci je 5288, dok je ukupan broj proteinskih sekvenci koje imaju interfejse: 2982. Nakon obrade datoteke sa interfejsima, broj ispravnih interfejsa koji predstavljaju ulazne podatke za algoritam za pronalaženje poravnanja je 2874. Interfejsi proteina koji sadrže više aminokiselina sa istim RES_POS brojem su 1IAK – sekvenca A, 1TOC – sekvenca U, 1YM0 – sekvenca B, 2AA3 – sekvenca A, 2UUK – sekvenca A, 3B7E – sekvnca A, i neće biti deo podataka koji će biti analiziran. Osim ovih šest, izostavljene su i sekvenca B proteina 1SGH i sekvenca I proteina 3J0F. Za ove sekvence u ulaznim podacima ne postoji primarna i sekundarna struktura, pa će i one biti isključene iz dalje obrade.

Primenom algoritma za pronalaženje poravnanja i sekundarnih struktura koje grade aminokiseline iz interfejsa, broj interfejsa koji mogu da se koriste u daljem radu je smanjen na 2895. Za 25 interfejsa proteinskih sekvenci nije bilo moguće odrediti jedinstveno potpuno ili delimično poravnanje, dok za 54 proteinske sekvence nije bilo moguće pronaći nijedno poravnanje. Broj interfejsa kod kojih je pronađeno delimično poravnanje je 61.

Broj aminokiselina, odnosno sekundarnih struktura, u ukupnom uzorku proteinskih sekvenci iznosi 783981, dok je broj amiknokiselina (sekundarnih struktura) u poravnatim interfejsima 84301. Modifikovane aminokiseline sa oznakama X (ukupno 4130) i Z (ukupno 1), kao i sekundarne strukture koje one grade su izostavljene iz analize. Modifikovane aminokiseline označene malim slovima su u ovom radu tumačene kao cistein, i kao takve su deo statističke analize. Kompletna statistika pojavljivanja aminokiselina u sekundarnim strukturama i učešće sekundarnih struktura u interfejsima se nalazi u *Dodatku 1*. Na osnovu ovih podataka, koeficijenti korelacije su izračunati u *Microsoft Excel* – u 2007.

4.1 KOEFICIJENT KORELACIJE IZMEĐU AMINOKISELINA I UČEŠĆA U INTERFEJSIMA

	r
A	-16.01
C	-9.39
D	-9.80
E	-2.94
F	13.49
G	-24.76
H	7.46
I	3.07
K	-8.27
L	14.55
M	13.17
N	-2.66
P	-0.49
Q	5.34
R	27.49
S	-9.74
T	-3.72
V	-5.24
W	10.28
Y	16.84

Tabela 5. Koeficijenti korelacije između aminokiselina i učešća u interfejsima

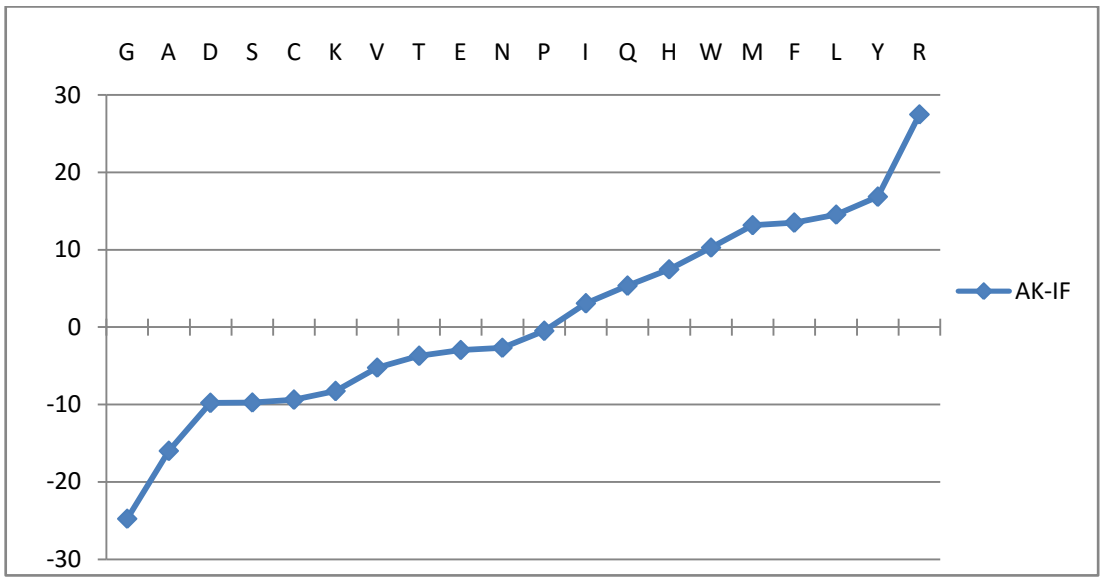
Tabela 5 sadrži vrednosti koeficijenata korelacije izračunate na osnovu formule (2), dok je dijagram prikazan na *Slici 15*. Rezultati su vrlo bliski 0, pa su zbog lakšeg prikaza pomnoženi sa 10^3 i zaokruženi na 2 decimale. Korelacije, čije su vrednosti podebljane, su korelacije čija je apsolutna vrednost veća od praga značajnosti. Značajnost korelacije se utvrđuje po formuli [10]:

$$|\rho| \geq \rho_{lim} = \frac{t_{lim}}{\sqrt{t_{lim}^2 + n - 2}} = \frac{1.96}{\sqrt{3.84 + n - 2}}$$

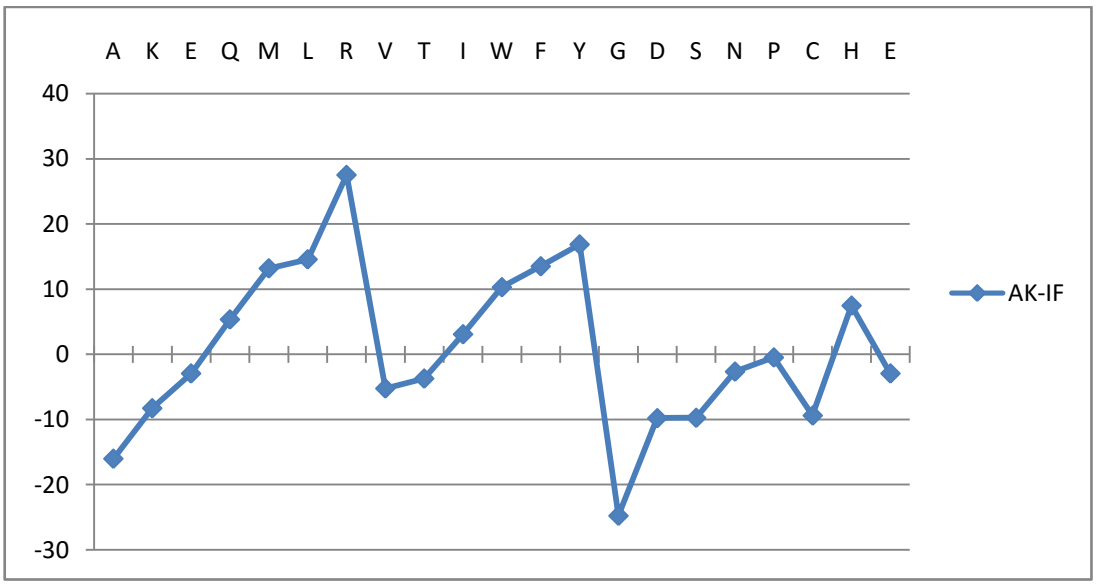
Za vrednost n = 783981 (što je ukupan broj aminokiselina/sekundarnih struktura u uzorku) prag značajnosti iznosi $2{,}21*10^{-3}$. Aminokiselina sa najmanjom apsolutnom vrednošću koeficijenta korelacije je prolin (P). Takođe, to je jedina aminokiselina čiji je koeficijent korelacije po apsolutnoj vrednosti manji od praga značajnosti.

Osim upoređivanja sa pragom značajnosti, dobijene koeficijente možemo uporediti i sa srednjom vrednošću, koja, izračunata na osnovu apsolutnih vrednosti koeficijenata korelacije, iznosi $10{,}24*10^{-3}$. Upoređivanjem ove vrednosti sa pojedinačnim

koeficijentima korelacije možemo da zaključimo da aminokiseline Alanin (A), Fenilalanin (F), Glicin (G), Leucil (L), Metionin (M), Arginin (R), Triptofan (W) i Tirozin (Y) imaju najznačajnije koeficijente korelacije. Alanin i Glicin su negativno korelirani, dok Fenilalanin, Leucil, Arginin, Triptofan i Tirozin imaju pozitivne vrednosti koeficijenta korelacije.



Slika 15. Dijagram koeficijenata korelacije između aminokiselina i interfejsa u rastućem poretku



Slika 16. Dijagram koeficijenata korelacije aminokiselina grupisanih prema sklonosti ka građenju određenih sekundarnih struktura i interfejsa

Na *Slici 16* prikazan je dijagram koeficijenata korelacije pri čemu su aminokiseline grupisane na osnovu sklonosti aminokiseline ka građenju određene sekundarne strukture

i zatim uređene po rastućoj vrednosti koeficijenta korelacije[9]. Po ovom kriterijumu, aminokiseline možemo podeliti u četiri grupe:

- aminokiseline sa sklonostima ka građenju α-heliksa (Ala, Leu, Glu, Gln, Arg, Met, Lys)
- aminokiseline sa sklonostima ka građenju β-ravni (Val, Ile, Tyr, Phe, Thr, Trp)
- aminokiseline sa sklonostima ka građenju ostalih sekundarnih struktura – 3-heliks, klupko, zaokret koji pravi vodonična veza, krivina (Gly, Asn, Pro, Asp, Ser)
- aminokiseline koje nemaju izražene sklonosti ka građenju neke konkretne sekundarnoj strukturi (Cys, His)

Na dijagramu se vidi da četiri aminokiseline sa najznačajnijim koeficijentima korelacije (Alanin (A), Leucil (L), Metionin (M) i Arginin (R)), pripadaju grupi aminokiselina koje teže ka građenju α-heliksa. Tri aminokiseline sa značajnim korelacijama (Fenilalanin (F), Triptofan (W) i Tirozin (Y), pripadaju grupi aminokiselina koje imaju sklonosti ka građenju β-ravni.

Sve aminokiseline koje spadaju u treću grupu, imaju negativne vrednosti koeficijenata korelacije, a među njima su i aminokiseline Asparagin (N) i Prolin (P) – aminokiseline sa najmanjim koeficijentima korelacije. Osim njih, u ovu grupu aminokiselina spada i Glicin (G) – aminokiselina sa koeficijentom korelacije koji ima najveću apsolutnu vrednost.

4.2 KOEFICIJENT KORELACIJE IZMEĐU SEKUNDARNIH STRUKTURA I UČEŠĆA U INTERFEJSIMA

U *Tabeli 6* prikazani su koeficijenti korelacije između sekundarnih struktura i interfejsa izračunati pomoću formule (2).

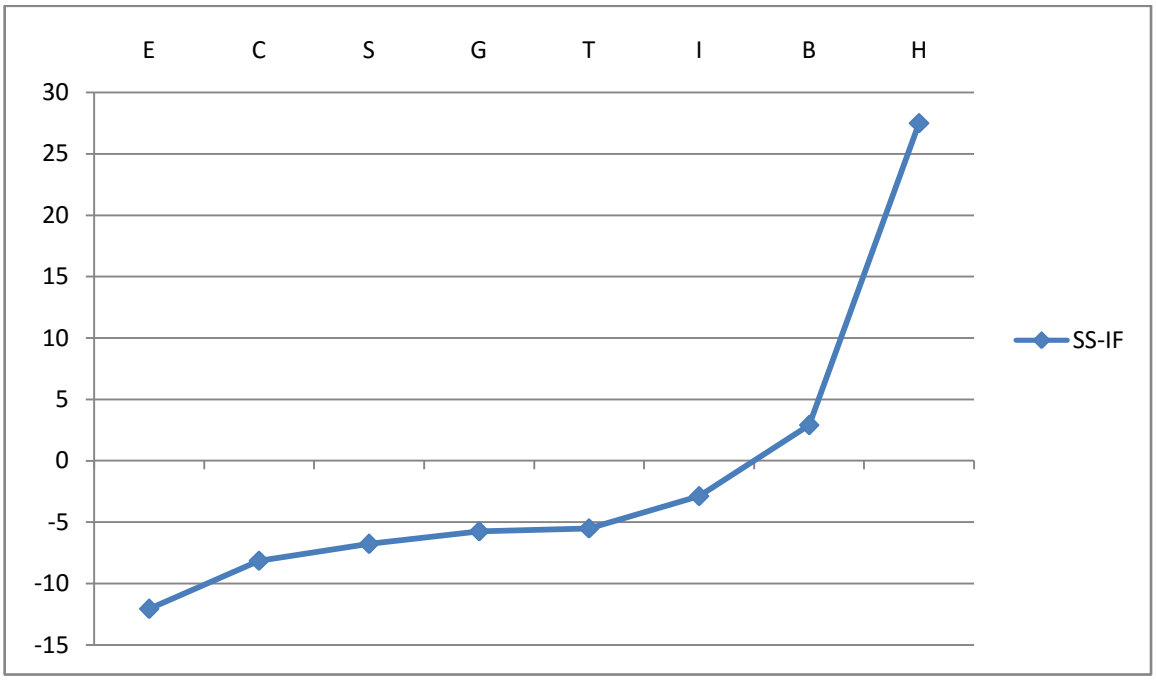
	r
B	2.89
C	-8.15
E	-12.07
G	-5.76
H	27.49
I	-2.89
S	-6.78
T	-5.51

Tabela 6. Koeficijenti korelacije između sekundarnih struktura i učešća u interfejsima

Poređenjem koeficijenata korelacije sa pragom značajnosti $2{,}21*10^{-3}$ može se ustanoviti da su nijedan koeficijent ne može biti odbačen kao beznačajan.

Srednja vrednost koeficijenata korelacije za sekundarne strukture i interfejse iznosi $8{,}94*10^{-3}$. Sekundarne strukture α–heliks (H) (pozitivna koreliranost) i β–ravni (E) (negativna koreliranost) su jedine sekundarne strukture čije su apsolutne vrednosti koeficijenata korelacije veće od $8{,}94*10^{-3}$. Na dijagramu na *Slici 17* prikazani su koeficijenti korelacije za sekundarne strukture.

Od svih sekundarnih struktura, aminokiseline najčešće grade α–heliks i i β–ravni. U uzorku od 783981 aminokiseline, α–heliks grade 250607 aminokiselina, od kojih 30031 učestvuje u izgradnji interfejsa. β–ravni grade 179590 aminokiselina, od čega 18079 aminokiseline grade interfejs.



Slika 17. Dijagram koeficijenata korelacije između sekundarnih struktura i interfejsa u rastućem poretku

5. Diskusija i zaključak

U radu su analizirani podaci o prostornoj građi proteina, kako bi se uočile eventualne zavisnosti između tipa aminokiselina i njihovog učešća u interfejsima, kao i između vrste sekundarne strukture koju grade aminokiseline i interfejsa. Najsloženiji deo je bilo pronalaženje sekundarnih struktura za aminokiseline koje su deo interfejsa. Razlog je u nekonzistentnosti podataka o proteinima u PDB-u.

Početna pretpostavka je bila da je pomeraj za sve pozicije aminokiselina u interfejsu isti. Kako algoritam zasnovan na prethodnoj pretpostavci nije dao željene rezultate (veliki broj interfejsa bez poravnanja), detaljnom analizom odgovarajućih PDB datoteka i ulaznih podataka o proteinskim sekvencama i interfejsima, napravljen je novi algoritam za pronalaženje poravnanja interfejsa sa proteinskom sekvencom. Algoritam posmatra interfejs proteina kao celinu i poravnanje je uspešno pronađeno samo ako je pronađeno poravnanje za sve aminokiseline koje su deo interfejsa. Kod višestrukih poravnanja, izvaja se deo interfejsa koji je presek svih pronađenih poravnanja. Ovakav algoritam je jedinstveno odredio položaje za 97,49% aminokiselina iz interfejsa. Za interfejse za koje poravnanje nije uspešno pronađeno, alternativno rešenje bi bilo da se izdvoji maksimalan podskup za koje je moguće pronaći jedinstveno poravnanje.

Rezultati pokazuju da postoji izrazita korelacija izmedju α–heliksa i interfejsa. Specifična struktura heliksa je najverovatnije razlog za to. Zbog položaja aminokiselina u heliksima, one su znatno izloženije spoljnim interakcijama sa drugim molekulima.

Za razliku od α–heliksa, ostali heliksi su sekundarne strukture koje se ređe javljaju u proteinima. Ovo može da ukazuje na to da su α–heliksi stukture koje “prirodne” za peptidne nizove, dok se ostali heliksi grade u delovima proteina koji su već u specifičnom odnosu sa drugim delovima proteina. Zbog postojećih veza smanjena je mogućnost učešća u interfejsima.

Interesantno je da β–ravni ne stupaju u više interakcija sa drugim molekulima. Jedan od razloga za to može da bude položaj rezidualnih delova aminokiselina u β–ravni. Rezidualni delovi aminokiselina su postavljeni približno normalno u odnosu na β–ravan i međusobno su relativno “gusti”, što im ostavlja malo prostora da stupe u interakcije sa drugim većim molekulima.

Dodatak 1

	B	B*	C	C*	E	E*	G	G*	H	H*	I	I*	S	S*	T	T*
A	42	361	865	9018	839	9237	178	2084	2466	24303	0	2	361	3431	526	4872
C	15	219	322	3354	304	3765	29	385	304	2985	0	2	93	1017	131	1019
D	54	338	1114	11169	567	5111	177	2143	1312	11087	0	8	502	5090	648	6555
E	49	321	863	7136	937	7876	230	2329	2478	20778	1	7	388	3725	594	5479
F	67	444	682	4640	1142	9761	161	1072	1490	8865	0	12	277	1706	334	1892
G	37	384	946	11242	621	7366	133	1420	663	6727	0	19	705	8842	1165	13583
H	26	196	498	4025	587	3922	95	692	686	4633	0	2	191	1699	302	2000
I	75	575	742	6262	1636	16039	107	769	2068	13935	0	11	285	1880	225	1541
K	56	450	882	8235	902	8057	151	1701	1818	15961	0	9	384	3889	524	5343
L	91	662	1253	9670	1713	16243	254	2143	4206	25972	1	11	458	3226	581	3714
M	15	134	378	2546	370	2808	67	400	847	4802	0	1	117	724	153	786
N	42	318	873	8010	515	4380	122	1228	1032	7496	0	5	436	3869	637	6290
P	39	326	1628	13502	401	2967	153	1659	519	3926	0	2	380	3257	615	5589
Q	31	221	588	4954	697	4767	117	1132	1598	11582	0	3	249	2088	337	2912
R	54	383	1039	6248	1156	7226	185	1219	2318	12975	0	5	426	2709	570	3310
S	52	519	1200	12424	969	8423	175	2050	1355	10713	1	4	450	4954	507	5330
T	79	579	1020	9739	1185	10815	126	939	1207	9331	0	5	425	3898	383	3335
V	90	649	903	7807	1979	21252	92	798	1887	13855	1	11	291	2451	278	2007
W	22	147	246	1613	424	3069	70	422	530	3166	0	4	83	571	132	754
Y	58	346	604	4021	1135	8427	163	956	1277	7454	1	3	287	1605	301	1827

Statistika pojavljivanja aminokiselina u sekundarnim strukturama i učešće u interfejsima

*Zvezdicom su označene sekundarne strukture koje gradi aminokiselina, ali one nisu deo interfejsa

	Pripada interfejsu	Ne pripada interfejsu
A	5277	53308
A*	79024	646372
C	1198	12746
C*	83103	686934
D	4374	41501
D*	79927	658179
E	5540	47651
E*	78761	652029
F	4153	28392
F*	80148	671288
G	4270	49583
G*	80031	650097
H	2385	17169
H*	81916	682511
I	5138	41012
I*	79163	658668
K	4717	43645
K*	79584	656035
L	8557	61641
L*	75744	638039
M	1947	12201
M*	82354	687479
N	3657	31596
N*	80644	668084
P	3735	31228
P*	80566	668452
Q	3617	27659
Q*	80684	672021
R	5748	34075
R*	78553	665605
S	4709	44417
S*	79592	655263
T	4425	38641
T*	79876	661039
V	5521	48830
V*	78780	650850
W	1507	9746
W*	82794	689934
Y	3826	24639
Y*	80475	675041

	Pripada interfejsu	Ne pripada interfejsu
B	994	7572
B*	83307	692108
C	16646	145615
C*	67655	554065
E	18079	161511
E*	66222	538169
G	2785	25541
G*	81516	674139
H	30061	220546
H*	54240	479134
I	5	126
I*	84296	699554
S	6788	60631
S*	77513	639049
T	8943	78138
T*	75358	621542

Ukupan broj aminokiselina u proteinskim sekvencama: **783931**

Ukupan broj aminokiselina koje grade interfejse: **84301**

Tabele sa podacima koji su korišćeni za izračunavanje koeficijenta korelacije.

Reference

[1] **Protein**, *Wikipedia*, https://en.wikipedia.org/wiki/Protein (13. jul 2015. godine).

[2] **Protein Biosynthesis**, *Wikipedia*, https://en.wikipedia.org/wiki/Protein_biosynthesis (27. maj 2015. godine).

[3] Artur M. Lesk, **Introduction to Protein Science: Architecture, Function, and Genomics** (Second Edition), *Oxford Univeristy Press*, 2010, ISBN 978-0-19-954130-0.

[4] Engelbert Buxbaum, **Fundamentals of Protein Structure and Function**, *Springer* 2007, ISBN 978-0-387-26352-6.

[5] Gert Vriend, **DSSP,** http://swift.cmbi.ru.nl/gv/dssp/ (24. avgust 2015. godine).

[6] **PDB Current Holdings Breakdown**, http://www.rcsb.org/pdb/statistics/holdings.do (24. avgust 2015. godine).

[7] **Protein secondary structure,** https://en.wikipedia.org/wiki/Protein_secondary_structure (2. avgust 2015. godine).

[8] **Phi coefficient**, https://en.wikipedia.org/wiki/Phi_coefficient, (24. avgust 2015. godine).

[9] S.Malkov, M.Živkovic, M.Beljanski, M.Hall, S.Zarić, **A reexamination of the propensities of amino acids towards a particular secondary structure: classification of amino acids based on their chemical structure**, *Journal of Molecular Modeling*, 2008, 14(8):769-775

[10] Samuels, Myra L., Witmer, Jeffrey A., Schaffner, Andrew A., **Statistics for the life sciences** (4th Edition), *Pearson Education Inc.*, 2012, ISBN 978-0-321-35280-5.