SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1525

BAZA PODATAKA METALA U PROTEINIMA

Alan Tus

Zagreb, lipanj 2010.

SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 4. Ožujka 2010.

ZAVRŠNI ZADATAK br. 1525

Pristupnik: Alan Tus

Studij: Računarstvo

Modul: Računarska znanost

Zadatak: **Baza podataka metala u proteinima**

Opis zadatka:

Cilj rada je za sve proteine u PDB bazi podataka odabrati one koji sadrže metale pri čemu su metali vezani barem uz jedan proteinski lanac. Proteinski lanac se definira kao lanac aminokiselinskih ostataka koji je dulji od 50 ostataka. Isto tako treba istražiti i utjecaj vezanja metala za ligande. Ligandi su svi lanci kraći od 50 molekula koji ne sadrže nukleinske kiseline. Metal je vezan uz proteinski lanac ukoliko su barem tri elektron donora iz proteina udaljeni od njega manje od 3 Angstroma. Kao elektron donori se definiraju svi O, S, N i Cl atomi. Za svaki metal koji je vezan uz barem jedan proteinski lanac potrebno je odrediti sljedeće podatke: pdb datoteku u kojoj se nalazi, način na koji je određena struktura (NMR ili kristalografija X zrakama), rezoluciju, sve elektron donor atome od kojih je udaljen manje od 3 Angstrema, tip atoma, imena lanaca u kojima se nalaze vezani atomi, informaciju da li su atomi u proteinskom lancu ili se radi o ligandu, imena molekula kojima ti atomi pripadaju, kutove koje metal tvori sa svakom dvojkom atoma s kojom je vezan, te sve metale od kojih je udaljen manje od 7 Angstrema. Od dobivenih podataka potrebno je napraviti bazu podataka. Podatci se trebaju moći automatski osvježavati svakih mjesec dana, a stare podatke je pritom potrebno pohraniti.

Koristiti skriptni jezik python i biopython biblioteku. Za detaljnije informacije obratiti se mentoru.

Zadatak uručen pristupniku: 5. Ožujka 2010.

Rok za predaju rada: 18. Lipnja 2010.

Hvala roditeljima na razumijevanju i podršci tokom studija,

hvala prijateljima što su uvijek bili tu kad je zapelo,

hvala mentoru Mili Šikiću na stručnom vodstvu i zabavnom radnom okruženju

i hvala Marku Čupiću što nas je naučio Javu

Sadržaj

Baza podataka metala u proteinima

1 Uvod

Razvoj računala visokih performansi na razini sklopovlja i programske podrške uz eksponencijalni rast interneta otvorio je nove izvore informacija u svim područjima znanja. Time je omogućen pristup informacijama i bazama znanja kakav dotada nije bio ni zamisliv. Bioinformatika kao sinteza dviju naočigled nesrodnih grana znanosti omogućava znanstvenicima brži pristup većem broju kvalitetnijih informacija. Time se lakše dolazi do novih otkrića i brže ostvaruje napredak.

Gore navedene prednosti ćemo iskoristiti kako bismo stvorili novi alat koji će poslužiti u istraživanjima biologa i kemičara u području proteina. Ovaj rad se koncentrira na izradu baze podataka svih metala i njihovih donora u proteinima te pripadajućih udaljenosti i kutova. Znanje o ulozi pojedinog metala u proteinu čija je uloga poznata može izuzetno doprinijeti kvaliteti života u raznim područjima od zdravlja do ekologije.

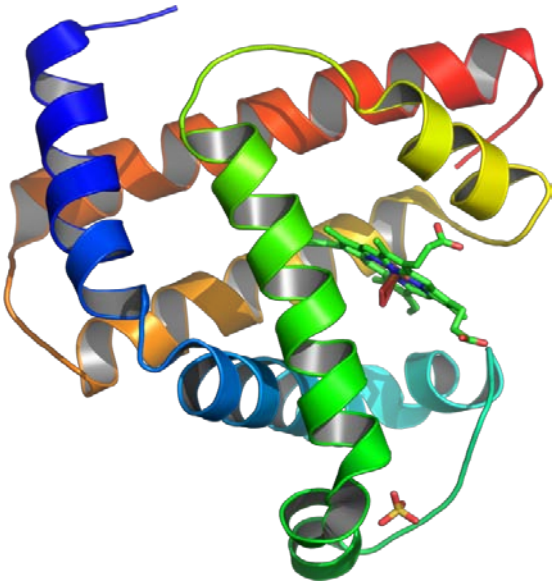
Datoteke s proteinskim strukturama se dobivaju iz baze proteina [1] te se potom parsiraju pomoću BioJava biblioteke, obrađuju u sustavu opisanom u ovom radu i pohranjuju u bazu podataka također opisanu ovim radom.

Ovaj rad je dio većeg projekta u koji je uključen diplomski rad „Baza zastupljenosti metala u proteinima“ [2] koji se bavi naknadom statističkom obradom podataka dobivenih ovim radom, te njihovom vizualizacijom pomoću tablica i grafova. Rezultati [2] su dostupni na http://valeria.zesoi.fer.hr:45288/metals/. Rezultati ovog projekta zasigurno će služiti biolozima i kemičarima u njihovom radu s proteinima.

2 Uloga metala u proteinima

Proteini imaju ključnu ulogu u gotovo svim biološkim procesima. Uloga proteina definirana je njegovom funkcijom koju određuje njegova struktura. *Proteomika* se bavi proučavanjem svojstava, interakcija, i funkcija proteina; to je znanstvena disciplina čiji je cilj opisati ukupnost proteina koji čine organizme (*proteome*).

Proteini su složene organske strukture koje se sastoje od aminokiselina povezanih peptidnim vezama, a čiji je slijed određen genima koji ih kodiraju. Linearan niz aminokiselina koje tvore protein uvija se u specifičnu trodimenzionalnu strukturu koja određuje njegovu funkciju. [3]



Slika 2.1 Primjer proteina – mioglobin [4]

Istraživanje genoma urodilo je spoznajom velikog broja aminokiselinskih sljedova koje kodiraju geni, međutim funkcija, struktura i interakcije proteina pripadnih sljedova uglavnom su nepoznate. Zato se ulažu veliki napori kako bi se strukture odredilo eksperimentalno ili računski. Primjerice, biotehnolozi računski određuju strukture proteina metodama rendgenske kristalografije i nuklearne magnetske rezonancije.

2.1 Metali u proteinima

Metali u proteinima imaju raznoliku ulogu. Od magnezija u klorofilu koji je važan za fotosintezu do željeza i bakra koji su važni za prijenos kisika u krvi. Znanje o broju i tipu aminokiselinskih ostataka koji koordiniraju s određenim metalom je važno

kako bismo znali koliko su specifični za pojavu određenih metala i time možda stekli uvid u funkciju proteina.

Trećina ili četvrtina svih proteina treba metal kako bi mogli obavljati svoju funkciju tako da možemo susresti velik broj različitih metala u proteinima. Mi ćemo se koncentrirati na sljedeće: FE, NI, MN, CA, CU, NA, MG, K, CO, ZN, CD, V, MO, W, PB, BR. [5]

2.2 Razdvajanje proteinskih kompleksa na lance

Proteinski lanci su građeni od aminokiselina, no često u prirodi dolaze u obliku proteinskih kompleksa koji osim proteinskih lanaca mogu sadržavati RNA ili DNA lance, ligande, metale i molekule vode. Osnovni gradivi elementi proteina prikazani su na slici 2.2. Proteinski kompleksi se sastoje od jednog ili više lanaca. U bazi proteinskih prostornih struktura [1] često se istim oznakama lanaca označavaju i proteinski lanci kao pripadni ligandi i vode. Za potrebe ovoga rada, za vrijeme obrade ovakvi lanci su rastavljeni na one jednostavnije lance. Lanci se sastoje od niza molekula i mogu biti različitih tipova, a mi prepoznajemo slijedeće: vodu, metal, lanac nukleinskih kiselina (RNA ili DNA), proteinske lance i ligande.

Voda je označena kao zasebni lanac iako se zapravo sastoji od niza ponekad i međusobno odvojenih molekula vode zaostalih u kristalografiji. Najlakše ju je prepoznati jer sadrži samo jednu grupu atoma i to HOH.

Metal zapravo nije lanac već samo jedan atom metala. Također ga je lako prepoznati jednostavnom usporedbom s popisom metala koje smo naveli u poglavlju 2.1.

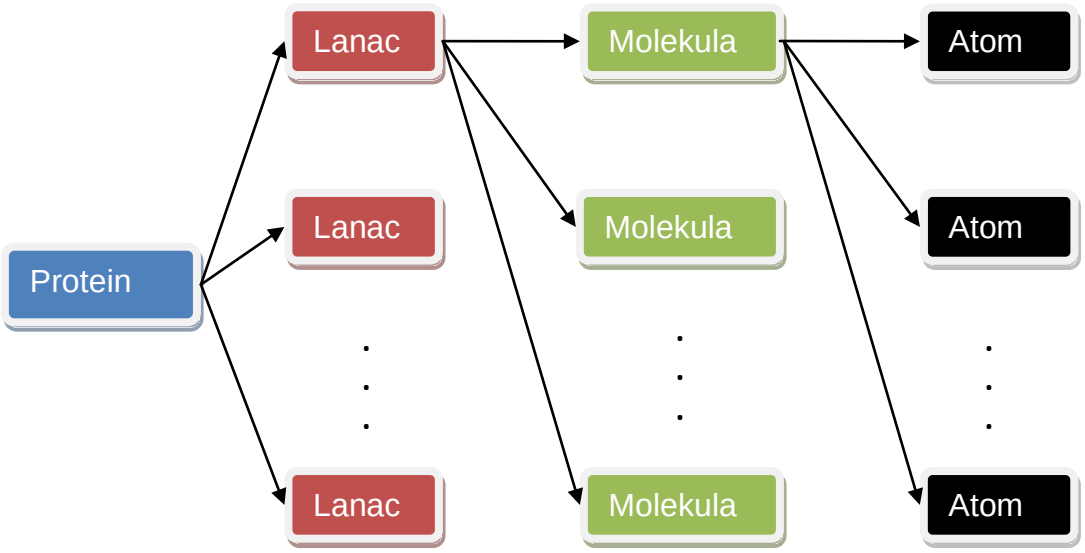
Nukleinske kiseline smo definirali kao one lance koji su sastavljeni od 90% molekula koje se nalaze u sljedećem skupu: DA, DT, DG, DC, DU, A, T, G, C, U.

Proteinski lance smo definirali kao lance dulje od 50 molekula i sastavljene od 90% molekula osnovnih 20 vrsta aminokiselina (ALA, ARG, ASN, ASP, CYS, GLU, GLN, GLY, HIS, ILE, LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR, VAL).

Ovaj uvjet za 90% je učinjen zbog toga što se često zna dogoditi da se u proteinskim lancima nalaze neuobičajene aminokiseline (npr. MSE, ACE i slično)

Ukoliko lanac ne pripada niti u jednu gore navedenu skupinu lanac je označen kao ***ligand***. Ligand je supstanca (mala molekula) koja ima sposobnost vezanja i tvorbe kompleksa sa biomolekulom u cilju obavljanja biološke funkcije.

Svaki lanac se sastoji od manjih gradivih molekula ili grupa atoma, koje se sastoje od atoma različitih elemenata. Atomi predstavljaju najmanju jedinicu strukture proteina. Među atomima, u ovome radu važni su jedino metali i njihovi elektron donori. Donori su svi atomi tipa O, N, CL, S koji se nalaze na udaljenosti manjoj od 3A (Angstrom – 10^{-10}) od nekog metala. Donori su negativno nabijeni za razliku od metala koji su pozitivno nabijeni te im 'doniraju' svoje elektrone kako bi cijeli sustav postao neutralan.



Slika 2.2 Prikaz strukture proteina

2.3 Biljna regeneracija tla

Primjer upotrebe proteina jest uklanjanje teških metala iz zagađenih područja. Biljna regeneracija tla (eng. *phytoremediation*) je proces u kojem se koriste žive zelene biljke za uklanjanje teških metala iz zagađenih područja. Posebno odabrane ili proizvedene biljke se sade na zagađenom području, a proteini u njima na sebe vežu određene metale i time čiste tlo. Ovaj proces je ekološki prihvatljiv i estetski ugodan način saniranja malo ili srednje zagađenih područja, a najveća prednost mu je što je gotovo dvostruko jeftiniji od uobičajenih metoda. Moguće ga

je koristiti u kombinaciji s uobičajenim metodama sanacije kao dodatnu pomoć ili završni korak. Biljna regeneracija tla ima i svojih mana. Postupak je ovisan o dubini do koje seže korijen korištene biljke, te o njenoj izdržljivosti obzirom na tip zagađenja. Veliki problem predstavlja i izlaganje biljojeda biljkama koje se koriste u procesu sanacije jer time teški metali dospijevaju u hranidbeni lanac i nastaje problem još većih razmjera.

Osim čišćenja tla moguće je čistiti i vodu i zrak i to ne samo od metala već i pesticida, otapala, eksploziva, nafte i njenih derivata te ostalih oblika zagađenja. Ovom metodom je moguće očistiti medij od arsena, kadmija, cinka, olova, soli, urana, žive, selenija i drugih tvari.

Biljke kroz svoj korijen izmjenom tvari iz tla uz vodu i minerale dobivaju i teške metale te ih vežu uz površinu korijena, pohranjuju u sam korijen ili prebacuju u biljna tkiva iznad površine. Neke biljke mogu apsorbirati više nego druge i otpornije su na veće koncentracije zagađenja, a druge biljke imaju mogućnost apsorpcije više od jednog metala.

Nakon što je gotov proces čišćenja biljke se sakupljaju i zbrinjavaju. Moguće je čak izvući metal iz biljaka ali se to rijetko radi jer je proces izuzetno skup. [6]

2.4 Postojeće baze metala u proteinskim kompleksima

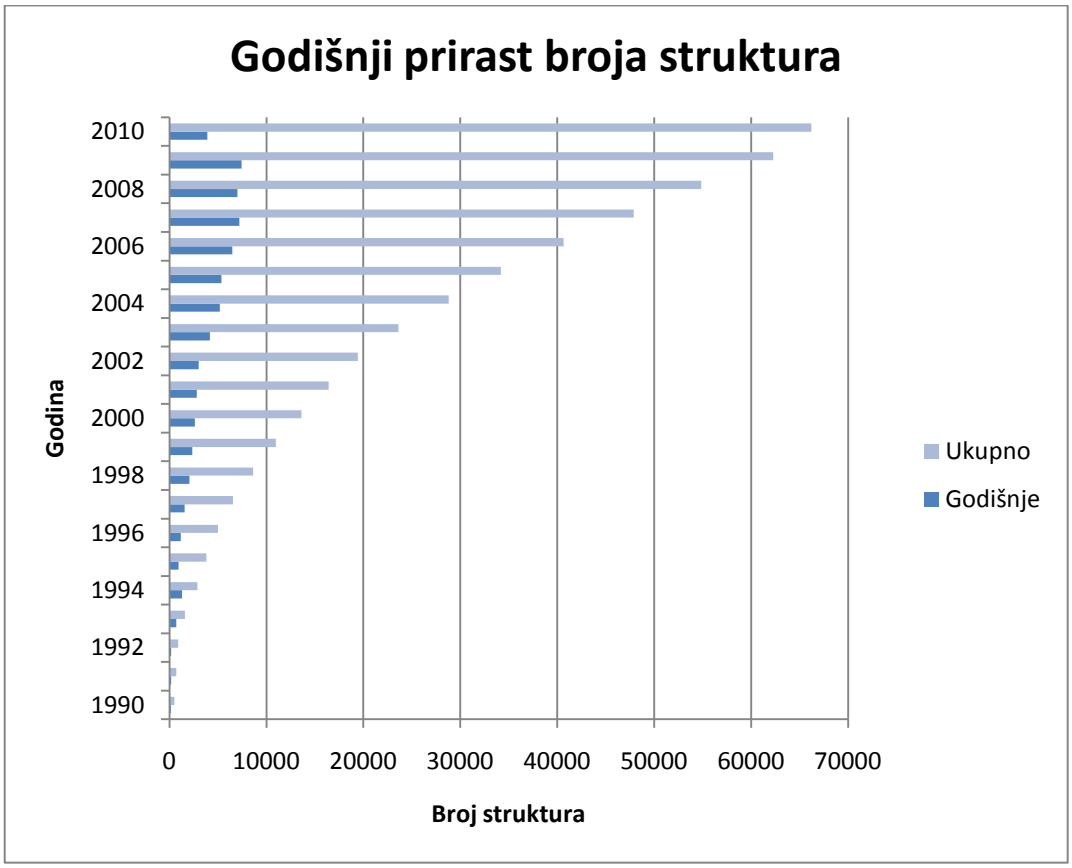
Od postojećih rješenja mogu spomenuti *Mespeus* [7] kao rješenje s najsličnijom funkcionalnosti, međutim trenutno nije funkcionalno jer je onemogućen unos svih parametara i nedostupni su mnogi rezultati. Nije nam poznato zašto je izvan funkcije. *Metalmine* [8] nudi popis svih struktura u kojima se nalazi neki od ponuđenih metala uz još neke dodatne informacije. *MIPS* [9] nudi pretragu baze podataka za svim strukturama koje sadrže određeni metal uz mogućnost definiranje dodatnih uvjeta. Sva ostala gotova rješenja na koja sam naišao u potrazi nisu obavljala ni približno sličan zadatak tako da mislim da je ovo trenutno jedinstveni alat i zasigurno najažurniji. Spomenuta postojeća rješenja će poslužiti prilikom kontrole točnosti rezultata ovog rada.

3 Podatci

Ulazni skup podataka predstavlja preslika svih dostupnih proteinskih struktura iz PDB baze [1] u mmCIF formatu (objašnjeno u poglavlju 3.2). Podatke dohvaćamo s udaljenog poslužitelja putem `ftp` protokola uz pomoć `rsync` naredbe o kojoj će biti više riječi u poglavlju 4.4.2.

3.1 Protein data bank

Baza proteinskih struktura [1] (engl. Protein data bank, PDB) je središnji svjetski repozitorij informacija o 3D strukturama velikih bioloških molekula. Spomenute molekule pronalazimo u svim organizmima, od bakterija, plijesni i biljaka do životinja i ljudi. Prikupljeno znanje se koristi kako bi se pokušala otkriti uloga pojedine strukture u ljudskom tijelu i potom primijeniti u razvoju lijekova. Pohranjeni su svi tipovi struktura od najmanjih proteina i fragmenata strukture DNA do kompleksnih molekularnih tvorevina poput ribosoma. Baza proteinskih struktura osnovana je 1971. godine, kada je u njoj bilo pohranjeno 7 struktura.



Slika 3.1 Godišnji prirast broja proteinskih struktura u [1]

3.2 mmCIF

mmCIF (*engl. macromolecular Crystallographic Information File*) je prilagodljiv i proširiv oznaka-vrijednost (*engl. tag-value*) oblik za zapisivanje makromolekularnih struktura podataka. Isprva je razvijen za opisivanje malih organskih molekularnih struktura, ali je prihvaćen 1990. godine na kongresu međunarodnog udruženja za kristalografiju (*engl. International Union of Cristallography – IUCr*). Osnovana je radna skupina koja proširila dotada stvoreni tip podatka kako bi mogao prihvatiti opis makromolekularnih kristalografskih struktura. [10]

Više o mmCIF strukturi podataka je moguće saznati na [11], a detalje o postupku kristalografije proteinskih struktura i načinima prikupljanju podataka na [12].

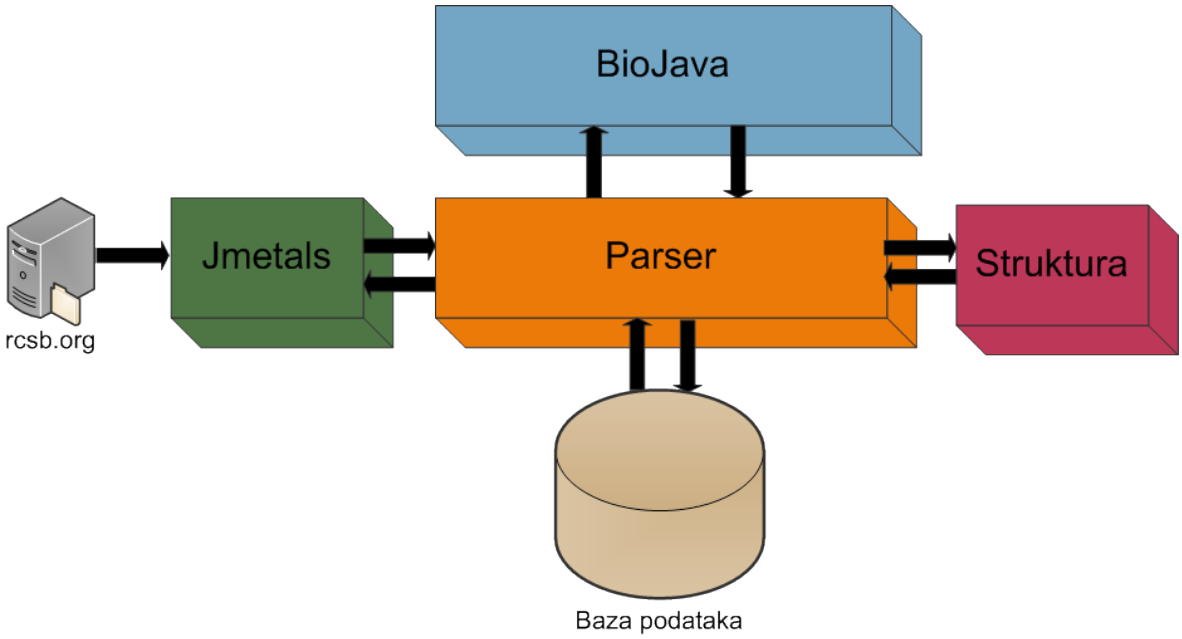
4 Implementacija

Cijeli sustav je pisan u Javi, a za pohranu podataka se koristi mySQL baza podataka. Sustav je vrlo jednostavno osmišljen tako da ga možemo ga podijeliti u pet osnovnih cjelina kao što je prikazano na slici 4.1.

JMetals predstavlja kontrolni dio sustava koji pokreće cijeli sustav, popunjava listu zadataka, inicijalizira unaprijed zadani broj dretvi i pokreće ih, a nakon završetka obrade pohranjuje statistike i gasi sustav. ***BioJava*** predstavlja BioJava biblioteku.

Svaku dretvu predstavlja jedan ***Parser*** koji vrši cijelu obradu podataka i njihovu pohranu u bazu podataka. *Parser* na početku svog rada dohvaća zadatak iz popisa zadataka, šalje datoteku u *BioJava* na parsiranje otkuda se vraća struktura proteina, analizira strukturu te ukoliko zadovoljava obrađuje ju i pohranjuje u *Bazu podataka*.

Baza podataka predstavlja mySQL bazu podataka koja čuva sve podatke nastale kao rezultat obrade proteinskih struktura.



Slika 4.1 Arhitektura sustava

U tekstu zadatka je navedeno kako implementaciju treba izvesti u *Pythonu* koristeći biblioteku *BioPython* [17], te je sukladno tome prva inačica bila implementirana u *Pythonu*. Nakon nekoliko probnih pokretanja prve inačice na manjem uzorku ulaznih datoteka i analize rezultata pokazalo se da je izvršavanje

vrlo sporo, te da bi obrada cijelog skupa ulaznih datoteka trajala oko pet dana. Kao glavne razlog za prebacivanje projekta u *Javu* osim predugog trajanja obrade podataka naveo bih lošu izvedbu biblioteke *BioPython* i nedostatak dokumentacije za istu. U dogovoru s mentorom odlučio sam odbaciti *Python* i koristiti *Javu* i ekvivalentnu biblioteku *BioJava*. Druga inačica sustava je bila pisana u *Javi* te su se performanse drastično poboljšale. Parsiranje cijelog skupa ulaznih datoteka trajalo je prosječno devet sati. U trećoj inačici je iskorištena višedretvenost i sada obrada cijelog skupa ulaznih datoteka traje samo tri sata.

Također smo na osnovu dobivenih rezultata i problema na koje smo naišli prilikom prvog testiranja odlučili promijeniti tip ulazne datoteke s .pdb na .cif. Datoteke s nastavkom .pdb su se pokazale iznimno nepouzdanima po pitanju preciznosti i dostupnosti nekih podataka, te smo naišli na mnoge nekonzistentnosti u strukturi. Isto tako .pdb oblik je već pomalo zastario i nudi manje informacija o strukturama od .cif oblika.

Na DVD-u uz kod je priložena pripadajuća dokumentacija koja se nalazi u Javadoc-u. Zato u ovom radu neću previše u ulaziti u detalje implementacije pojedinih metoda već ću ukratko opisati postupke koji se primjenjuju u poglavlju 4.5.

4.1 Java

Java je programski jezik razvijen u tvrtki *Sun Microsystems*, a glavna mu je karakteristika prenosivost na različite platforme. Sintaksa jezika je većinom prenesena iz *C++*, ali za razliku od njega kombinira sintaksu strukturiranog, generičkog i objektno orijentiranog programiranja. Sav kod je pisan u klasama i sve se sastoji od objekata, tako da je Java zapravo isključivo objektno orijentirani jezik. U usporedbi sa srodnim *C++*, *Java* ne podržava neke funkcionalnosti poput pokazivača, preopterećivanja operatora i višestrukog nasljeđivanja, ali isključivo u svrhu pojednostavljivanja i izbjegavanja nastanka mogućih pogrešaka. [13]

4.2 BioJava

U implementaciji je korišten paket BioJava. Projekt otvorenog koda namijenjen da pruži razvojni okvir u Javi za obradu bioloških podataka. Pruža analitičke i statističke metode, parsere za česte oblike datoteka i dopušta manipuliranje

sekvencama i 3D strukturama. Cilj projekta jest omogućiti brz razvoj aplikacija iz područja bioinformatike. BioJava [14] je dio Open Bio udruženja, neprofitne organizacije sačinjene od volontera usredotočenih podržavanju razvoja programske podrške u bioinformatici. Iz iste organizacije su potekli i drugi slični projekti kao što su *BioPython, BioPerl, BioRuby, BioPerl* i drugi. [15]

Korištena inačica *BioJave* 1.7.1, objavljena 15. siječnja 2008. godine. U međuvremenu je pokrenut projekt BioJava2 koji je rađen kao nadogradnja na BioJava1, međutim od njega se odustalo jer se pokazalo se da će konačni projekt biti prevelik i da je potreban novi pristup. Sredinom 2009. godine je pokrenut BioJava3 projekt koji je trenutno u alpha fazi razvoja tako da nije bio pogodan za korištenje u ovom radu.

Za potrebe ovog rada su nedostajale neke metode i članske varijable unutar struktura te parser nije vraćao potrebne podatke o strukturama. Zato sam preuzeo izvorni kod BioJava projekta i nadogradio ga kako bi zadovoljavo potrebe ovog rada. Izmijenjena inačica je dostupna na DVD-u priloženom uz ovaj rad.

4.3 Baza podataka

Baza podataka je implementirana u *mySQL*-u. *MySQL* je trenutno najpopularnija besplatna relacijska baza podataka. Više o *mySQL*-u možete saznati na [16].

Prilikom izrade baze podataka nisam se u potpunosti pridržavao pravila o izradi relacijskih baza podataka i zato njene relacije nisu normalizirane. Trenutni oblik baze podataka sadrži puno redundantnih podataka, ali to je svjesno i ciljano napravljeno jer se radi s velikim količinama podataka i upiti u bazu podataka bi bili vrlo složeni, a time i dugotrajni.

Baza podataka zapravo prati strukturu proteina. Vizualni prikaz odnosa relacija baze podataka je vidljiv na slici 4.2. U bazi podataka se nalazi ukupno šest relacija.

U relaciji ***PDB*** se nalaze podatci o svim strukturama kao što su jedinstvena oznaka strukture, naziv, način i rezolucija na kojoj su dobiveni podatci te datum objave. Na relaciju *PDB* se vežu relacije *chain*, *residue* i *atom*. Konkretno vežu se na jedinstvenu oznaku strukture.

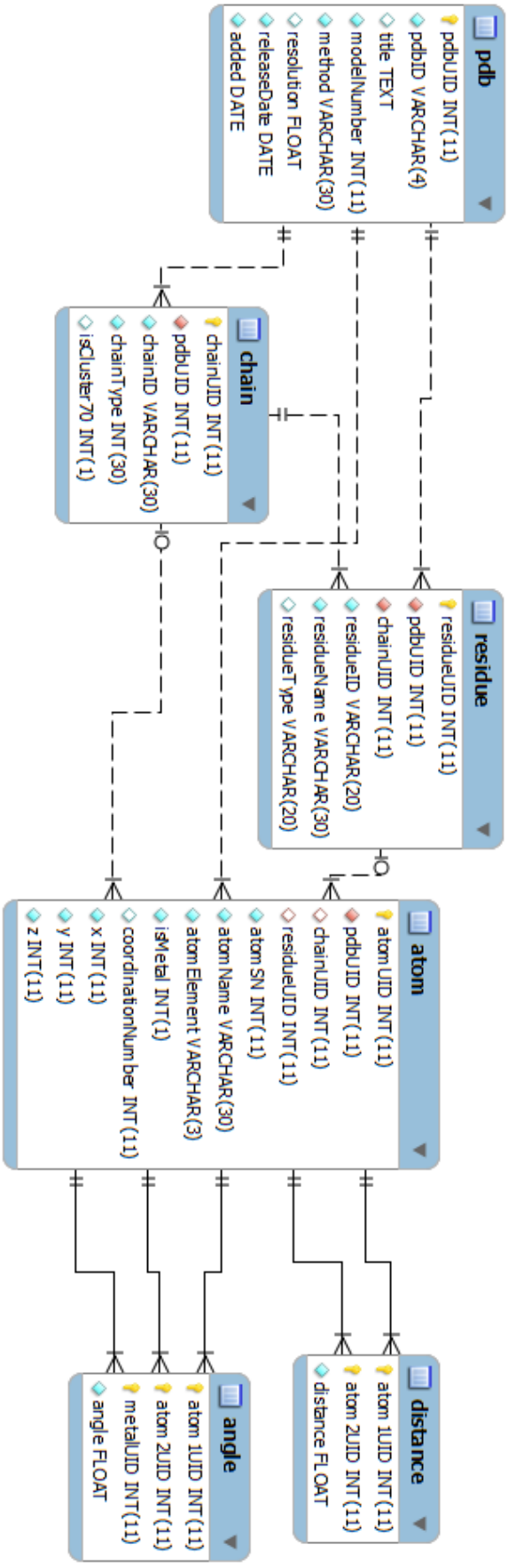
Relacija ***chain*** sadrži pojedinosti o svakom korištenom lancu svih unesenih struktura. Sadrži jedinstvenu oznaku, oznaku strukture kojoj pripada, naziv i tip lanca (*water, metal, nucleicAcid, aminoAcid, ligand*). Na jedinstvenu oznaku lanca u relaciji *chain* se vežu relacije *residue* i *atom*.

Relacija ***residue*** u sebi sadrži podatke o korištenim grupama atoma koje sačinjavaju korištene lance. Za svaku grupu atoma postoji jedinstvena oznaka, oznaka strukture i lanca kojem pripada, naziv i tip (*amino, hetatm, nucleotide*). Na jedinstvenu oznaku lanca u relaciji *residue* se veže relacija *atom*.

Relacija ***atom*** čuva podatke o svim korištenim atomima. Točnije, sadrži jedinstvenu oznaku atoma, oznaku strukture, lanca i grupe atoma kojoj pripada, serijski broj atoma unutar strukture kojoj pripada, naziv, element, zastavicu koja označava radi li se o metalu, koordinacijski broj i koordinate izražene u Angstromima. Na jedinstvenu oznaku atoma unutar relacije *atom* se vežu relacije *distance* i *angle*.

Relacija ***angle*** sadrži podatke o svim mogućim kombinacijama kutova donora pojedinog metala. Sadrži jedinstvene oznake dvaju atoma koji čine krakove kuta i jedinstvenu oznaku atoma metala koji čini vrh kuta te vrijednost kuta izraženu u radijanima.

Relacija ***distance*** u sebi sadrži podatke o udaljenosti svih donora metala koji su udaljeni manje od 3A od metala. Sadrži jedinstvene oznake dvaju atoma između kojih se traži udaljenost i vrijednost udaljenosti osim udaljenosti donora sadrži i udaljenost metala koji su udaljeni manje od 7A.



Slika 4.2 Nacrt baze podataka

4.4 Arhitektura sustava

Sustav je organiziran u tri paketa:

- *hr.fer.zesoi.metals.main* – sadrži klase *Main*, *Database* i *Parser*.
- *hr.fer.zesoi.metals.objects* – sadrži korištene strukture podataka.
- *hr.fer.zesoi.metals.util* – sadrži klase *Data* i *Tools*

Main klasa zapravo samo inicijalizira cijeli sustav, pokreće obradu i na kraju ispisuje statistike. Podijeljena je na *main* metodu te na nekoliko pomoćnih metoda koje inicijaliziraju popis zadataka, ispisuju završnu statistiku i zapisuju popis obrađenih struktura u datoteku.

Database klasa obavlja sve poslove vezane uz bazu podataka, kao što su uspostavljanje veze, izvršavanje upita, potvrđivanje transakcija i zatvaranje veze.

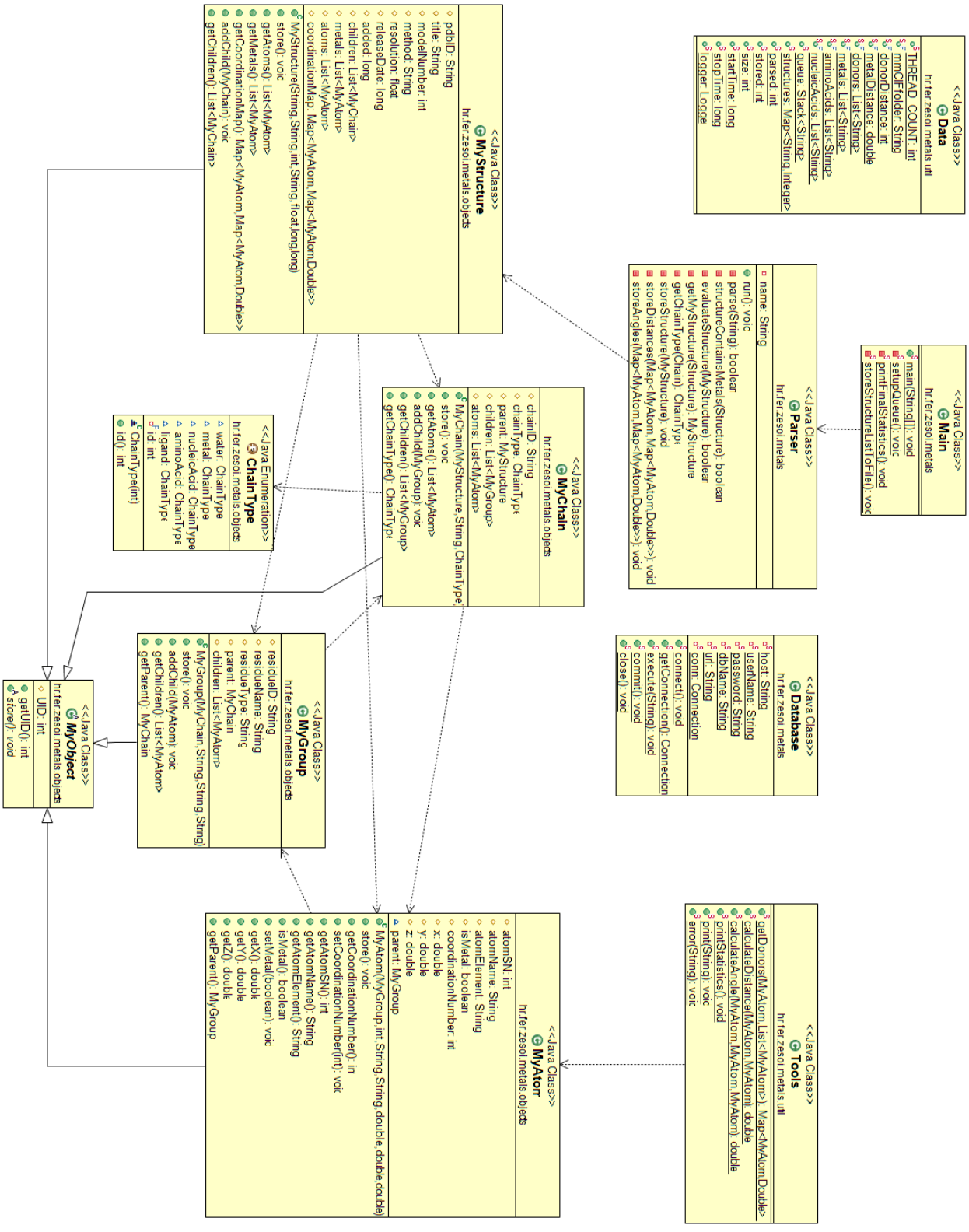
Parser klasa je ona koja obavlja sav posao obrade podataka. Učitava strukturu pomoću metode iz BioJava biblioteke, provjerava zadovoljava li zadane kriterije te ju potom obrađuje.

Data klasa sadrži globalne varijable (popis metala, aminokiselina, nukleinskih kiselina…) i metodu za zapisivanje pogrešaka u log datoteku.

Tools klasa sadrži često korištene metode pri obradi podataka poput izračuna udaljenosti i kutova te pronalaženje donora oko nekog metala.

U *hr.fer.zesoi.metals.objects* paketu se nalaze objekti ***MyStructure, MyChain, MyGroup i MyAtom*** koji predstavljaju elemente strukture proteina. Slične objekte nalazimo u BioJava paketu međutim odustao sam od njihove nadogradnje i odlučio se za vlastitu implementaciju.

Cjelovitu sliku sustava moguće je vidjeti na slici 4.3. gdje je prikazan dijagram razreda.



Slika 4.3 Dijagram razreda

4.4.1 Višedretvenost

U svrhu ubrzanja procesa obrade podataka upotrijebljena je višedretvenost. Na taj način smo postigli paralelizaciju procesa obrade ulaznih podataka, maksimalno iskoristili procesorske resurse te vrijeme koje smo inače izgubili čekajući da se struktura učita iz datoteke i da se podatci upišu u bazu podataka sada koristimo za daljnju obradu.

Za implementaciju višedretvenosti korišteni su gotovi alati koje nudi Java. Razred `Parser` implementira sučelje `Runnable` koje zahtijeva metodu `run()`. Implementacija sučelja `Runnable` nam omogućava da pretvorimo ovaj razred u samostalnu dretvu koja obavlja neki posao i može se pokrenut i zaustaviti. Potrebno je samo prilikom poziva `new Thread` kojim se stvara nova dretva predati razred `Parser` kao argument. Time će se prilikom pokretanja dretve pokrenuti metoda `run()` koju smo implementirali u razredu `Parser` i pokrenuti obrada podataka.

Baza podataka nam ovdje predstavlja usko grlo iz više razloga. To je najsporiji proces uz onaj čitanja iz datoteke. Odjednom joj može pristupiti samo jedna dretva. Proces stvaranja višestrukih novih veza s bazom podataka je izuzetno zahtjevan što se tiče memorijskih i procesorskih resursa, a i nema smisla jer bi se ionako u samoj bazi podataka ponovno morala provoditi neka vrsta sinkronizacije. Zahvaljujući višedretvenosti jedna dretva može pisati u bazu, još jedna ili više njih čekati na upis u bazu podataka dok ostale dretve dalje nesmetano obavljaju svoj posao.

Korišteni model višedretvenosti je zapravo vrlo jednostavan. Pri inicijalizaciji programa se stvara lista zadataka u obliku popisa putanja do datoteka sa strukturama koje je potrebno obraditi. Nakon toga se stvara unaprijed definirani broj dretvi. Preporučeni broj dretvi za optimalan rad je dvostruki broj od broja procesorskih jezgri. Ukoliko broj dretvi prelazi dvostruki broj procesorskih jezgri kod obrade većih struktura dolazi do zagušenja i rad se toliko usporava da se više ne isplati koristit višedretvenost. Potom se stvorene dretve pokreću i počinje se s radom. Svaka dretva skida iz liste zadataka po jedan zadatak, obrađuje ga i potom skida novi zadatak. Pristup listi zadataka je sinkroniziran tako da dvije dretve ne mogu istovremeno skinuti isti zadatak.

Zahvaljujući višedretvenosti i malo optimizacije korištenih metoda vrijeme izvršavanja je smanjeno s devet sati na samo tri sata. U odnosu na početnih pet dana koliko je trebalo prvoj inačici pisanoj u Pythonu ovo je veliki uspjeh. Daljnja optimizacija je zasigurno moguća no dobitci su zanemarivi jer smo dostigli ograničenja raspoloživih memorijskih i procesorskih resursa.

4.4.2 Cron

Cron je GNU/Linux alat koji omogućuje precizno definiranje periodičnog pokretanje neke naredbe ili skupa naredbi. Pošto se baza proteina [1] tjedno osvježava i nadopunjuje novim strukturama potrebno će biti vršiti tjedna osvježavanja naše baze podataka. Ovaj alat će nam omogućiti automatizaciju cjelokupnog procesa.

1. Sinkroniziraj s [1] poslužiteljem.

```
rsync -rlpt -v -z --delete --port=33444
rsync.wwpdb.org::ftp_data/structures/divided/mmCIF/
./mmCIF
```

2. Arhiviraj bazu podataka

```
mysqldump -p metali > metals.sql
```

3. Isprazni bazu podataka

```
TRUNCATE TABLE angle;

TRUNCATE TABLE distance;

TRUNCATE TABLE atom;

TRUNCATE TABLE residue;

TRUNCATE TABLE chain;

TRUNCATE TABLE pdb;
```

4. Pokreni parsiranje

```
Java –jar JMetals.jar 8
```

4.4.3 Valeria

Obrada podataka se vrši na računalu pokretanom s Intel® Core™2 Quad CPU Q6600 @ 2.40GHz procesorom i 3,5 GB RAMa. Za pohranu podataka na raspolaganju nam je 550GB diskovnog prostora.

4.5 Metode

Redoslijed aktivnosti obrade podataka je grafički prikazan kao dijagramu aktivnosti na slici 4.4., te je u nastavku detaljnije opisan.

4.5.1 Pokretanje sustava

Sustav na početku rada pregleda direktorij u kojem se nalaze datoteke s proteinskim strukturama i sprema putanje svih datoteka u listu zadataka. Lista zadataka se nalazi u razredu *Data* i predstavljena je strukturom stoga čiju implementaciju nudi Java. Pristup popisu zadataka je sinkroniziran kako bi se spriječio istovremeni pristup dvije dretve i dohvat istog zadatka.

Sustav potom prelazi na stvaranje i pokretanje dretvi. Prilikom pokretanja sustava iz komandne linije putem parametara je predan broj dretvi koje je potrebno inicijalizirati i pokrenuti. Ukoliko parametar nije postavljen automatski se stvara osam dretvi.

4.5.2 Obrada podataka

Svaka dretva zasebno vrši obradu podataka. Na početku obrade dohvaća se putanja datoteke iz popisa zadataka i šalje se BioJavi na parsiranje. Biojava vraća strukturu koja se provjerava odgovara li sljedećim uvjetima:

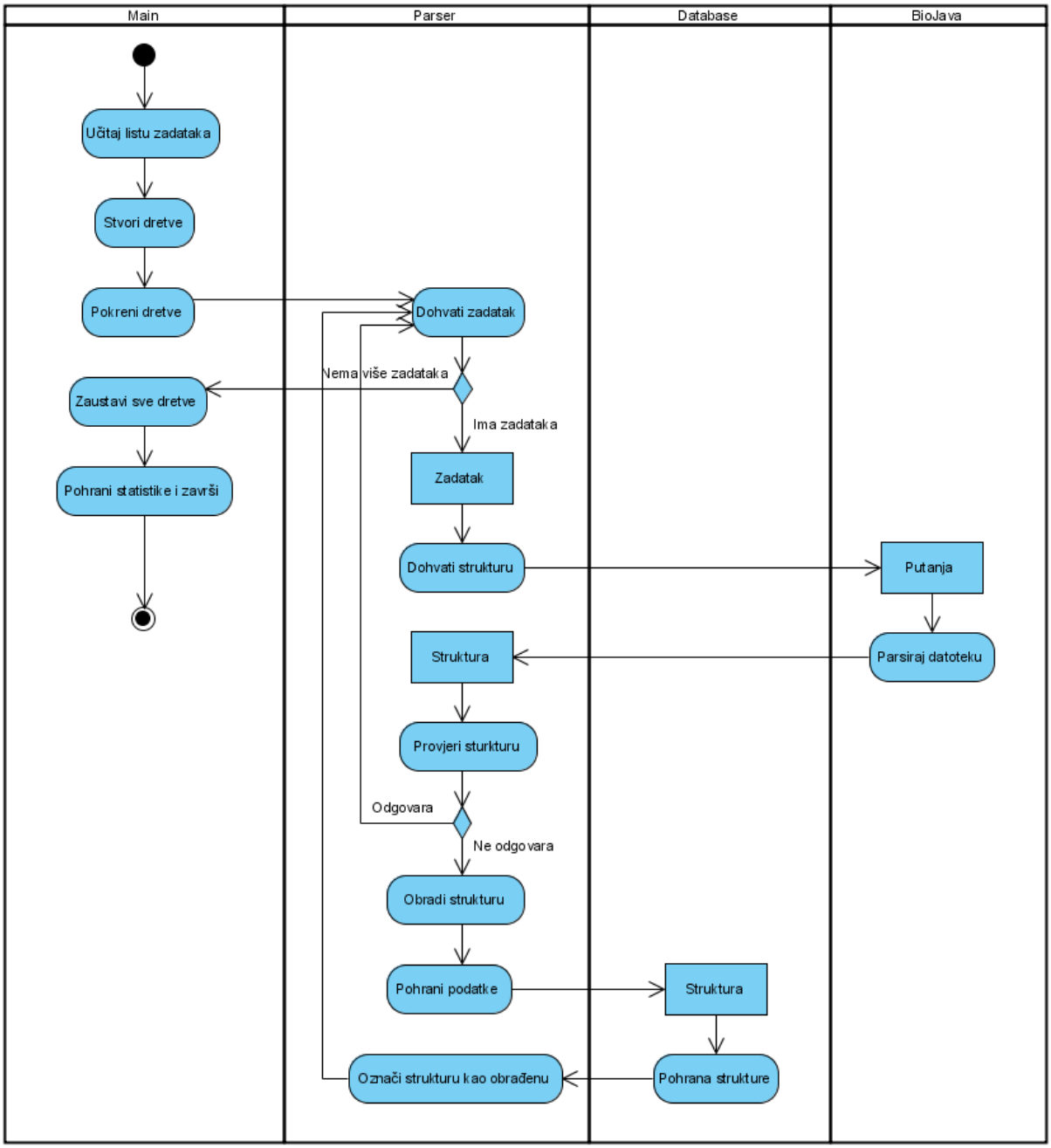
1. Sadrži li barem jedan metal? – Iteracijom po svim atomima strukture traži se barem jedan metal iz skupa metala koje smo naveli u poglavlju 2.1., a čiji se popis nalazi u razredu *Data*.
2. Sadrži li barem jedan proteinski lanac? – (definirano u poglavlju 2.2).
3. Sadrži li neki proteinski lanac barem dva donora povezana s metalom? – donore smo opisali u poglavlju 2.2. Kažemo da je donor povezan s metalom ukoliko je na manjoj udaljenosti od 3A.

Ukoliko proteinska struktura zadovolji sva tri uvjeta prelazi se na obradu podataka. Prvo se cijela struktura prevede iz BioJava strukture u našu implementaciju strukture. Pritom se zadržavaju samo korisni podatci i definira se vrsta svakog lanca. Izdvajaju se svi metali u poseban popis i za svaki metal se određuje donore i njihovu udaljenost od metala te susjedne metale i njihove udaljenosti (dva metala susjedni ako se nalaze na udaljenosti manjoj od 7A i ako su povezani s istim lancem). Sada je potrebno još samo odrediti kutove između svih donora pojedinog

metala. Kutovi su definirani tako da se metal nalazi u vrhu kuta, a donori čine krakove. Potrebno je izračunati sve vrijednosti kutova za sve permutacije parova njegovih donora. Na kraju se određuju korisni atomi tako što se stvara popis jedinstvenih vrijednosti svih atoma koji se nalaze na popisu metala ili donora. Izdvajaju se samo korisni atomi kako bismo ubrzali proces obrade, smanjili veličinu baze podataka i izbjegli unos nepotrebnih podataka koji bi nam samo usporavali buduće analize. Korisni atomi i svi viši dijelovi strukture (grupe atoma, lanci i struktura) koji se vežu na njih se spremaju u bazu podataka. U bazu podataka se također spremaju sve izračunate udaljenosti i kutovi. Ovime je završila obrada jedne strukture i dohvaća se novi zadatak.

4.5.3 Završetak procesa

Nakon što se isprazni popis zadataka završena je obrada ulaznog skupa datoteka i sustav gasi prethodno pokrenute dretve. Ispisuje se završna statistika koja govori koliko je struktura pregledano, koliko ih je spremljeno i koliko je cijeli proces trajao. Na kraju se sprema popis svih obrađenih struktura u jednu datoteku pod nazivom „*structures.txt*“ i pored svake strukture se upisuje broj 1 ukoliko je struktura spremljena u bazu podataka. Ovo će poslužiti kod budućih nadogradnji sustava kada će se umjesto cijelog skupa ulaznih datoteka svaki puta pregledavati samo nove datoteke koje se dotada još nisu pojavile.



Slika 4.4 Dijagram aktivnosti

5 Rezultati

Ulazni skup podataka bio je veličine 14GB i sadržavao 65 968 datoteka. Nakon parsiranja i obrade podataka koji su trajali tri sata dobiveni su slijedeći rezultati.

Spremljeno je 13 849 struktura u bazu podataka i ona je veličine 70MB. Rezultati obrade podataka su provjereni na dva načina. Ručno, tako što je odabrano nekoliko slučajnih struktura i provjereno je podudaraju li se dobiveni podatci s onima u bazi proteina i samoj datoteci. Drugi način provjere je bio usporedbom rezultata iz sličnih servisa poput [9] i [8]. Rezultati su se pokazali pouzdanima.

Za sada je u planu samo jedna nadogradnja. Omogućiti pokretanje obrade podataka uz uvjet da pregledava samo svježe dodane strukture i osvježi postojeću bazu podataka. Sustav je moguće vrlo lako nadograditi tako da je otvoren za daljnji razvoj. Poboljšanja i optimizacije su uvijek moguće no smatram da je za dani zadatak koji će se izvršavati maksimalno jednom tjedno postignuta brzina obrade zadovoljavajuća.

5.1 Cluster 70 skup

Cluster 70 predstavlja skup proteinskih lanaca grupiranih u grozdove (*engl. cluster*) unutar kojih je međusobna sličnost pojedinih lanaca najmanje 70%. Datoteka koja sadrži taj skup se generira tjedno te se može preuzeti s RCSB poslužitelja [1].

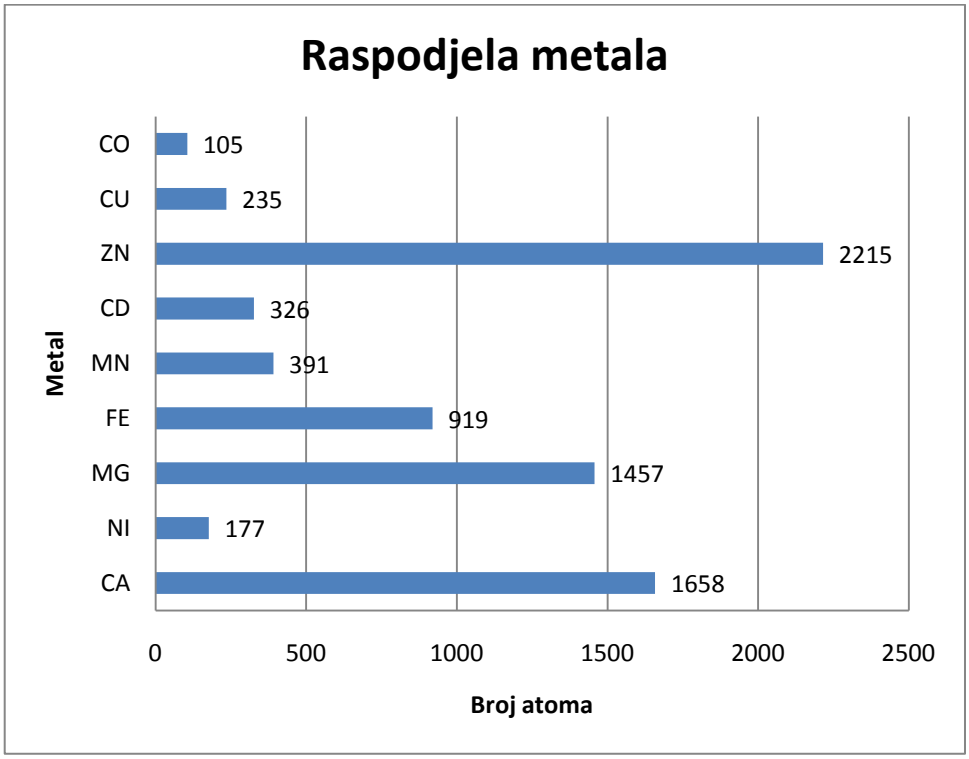
Budući da se popis *Cluster 70* osvježava svakog tjedna potrebno je koristiti najsvježije dostupne podatke. Kako bi se automatizirao ovaj proces razvijen je program koji obavlja ažuriranje te je opisan u [2].

5.2 Diskusija

Važno je napomenuti da su sve ovdje prikazane analize rađene na skupu atoma koji pripadaju lancima iz *Cluster* 70 skupa kako bi spriječili redundanciju i dobili što preciznije rezultate. Rezultate dobivene ovim radom ćemo usporediti s rezultatima dobivenim u [5]. U radovima su korištene različite metode označavanja pripadnosti pojedinih lanaca *Cluster* 70 skupu. U [5] je uvijek odabran prvi lanac iz skupa i on je korišten kao predstavnik, a u ovom radu je odabran prvi lanac iz skupa koji

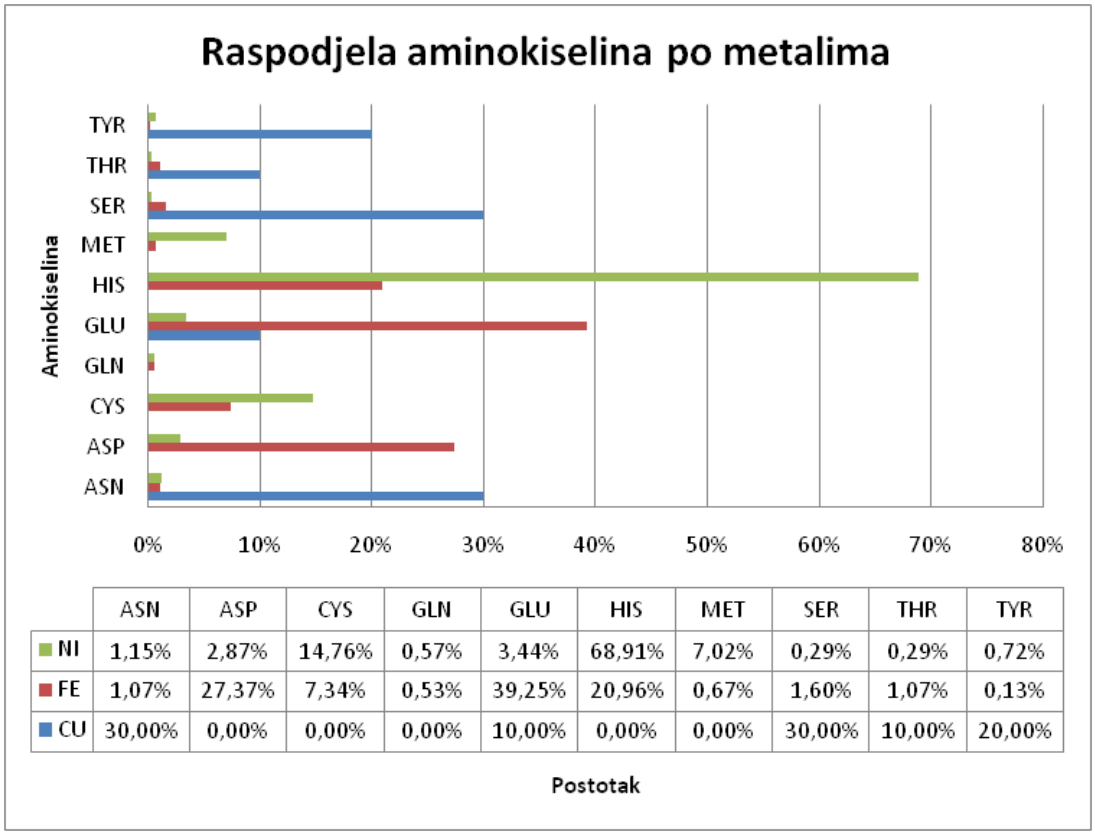
sadržava metal te je on korišten u daljnjem radu. Vjerujem kako je metoda korištena u ovom radu povoljnija i da bi trebala pružiti preciznije rezultate.

Prvi grafikon na slici 5.1. nam govori o količini pojedinog metala u ukupnom pronađenom broju atoma svih traženih metala iz skupa opisanog u poglavlju 2.1. Prikazano je samo prvih 9 metala s najviše atoma. Primijetimo kako su ZN, CA i MG redom tri najčešća metala u obrađenim strukturama. U [5] su rezultati prilično slični osim razlike u poretku prva tri metala, gdje su redom MG, ZN i CA. Zanimljivo je kako su se pojavile drastične razlike u broju atoma tri vodeća metala primjenom drugačije metode opisane u ovom poglavlju. Ostali metali imaju sličan broj atoma koji lagano varira, ali ni približno drastično poput prva tri.



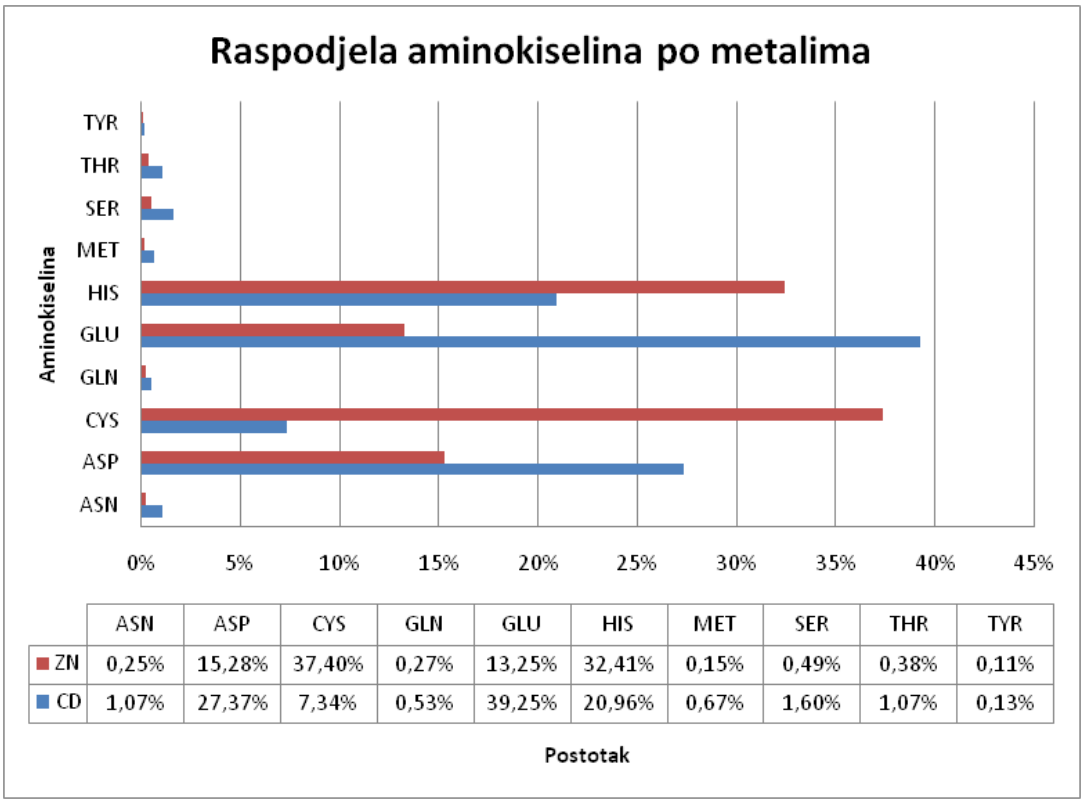
Slika 5.1 Raspodjela metala

Nadalje na slikama 5.2, 5.3 i 5.4 su prikazani postotci prisutnosti donora iz pojedinih aminokiselina za najčešće aminokiselina i metala. Metali su grupirani kao i u [5] prema sličnosti njihovih koordinacijskih sfera: Cu^{2+}, Fe^{2+} i Ni^{2+}, zatim Zn^{2+} i Cd^{2+} i naposljetku Ca^{2+}, MN^{2+}, Mg^{2+} i Co^{2+}. Prikazane su samo slijedeće aminokiseline: Asn, Asp, Cys, Gln, Glu, His, Met, Ser, Thr i Tyr.



	ASN	ASP	CYS	GLN	GLU	HIS	MET	SER	THR	TYR
NI	1,15%	2,87%	14,76%	0,57%	3,44%	68,91%	7,02%	0,29%	0,29%	0,72%
FE	1,07%	27,37%	7,34%	0,53%	39,25%	20,96%	0,67%	1,60%	1,07%	0,13%
CU	30,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	30,00%	10,00%	20,00%

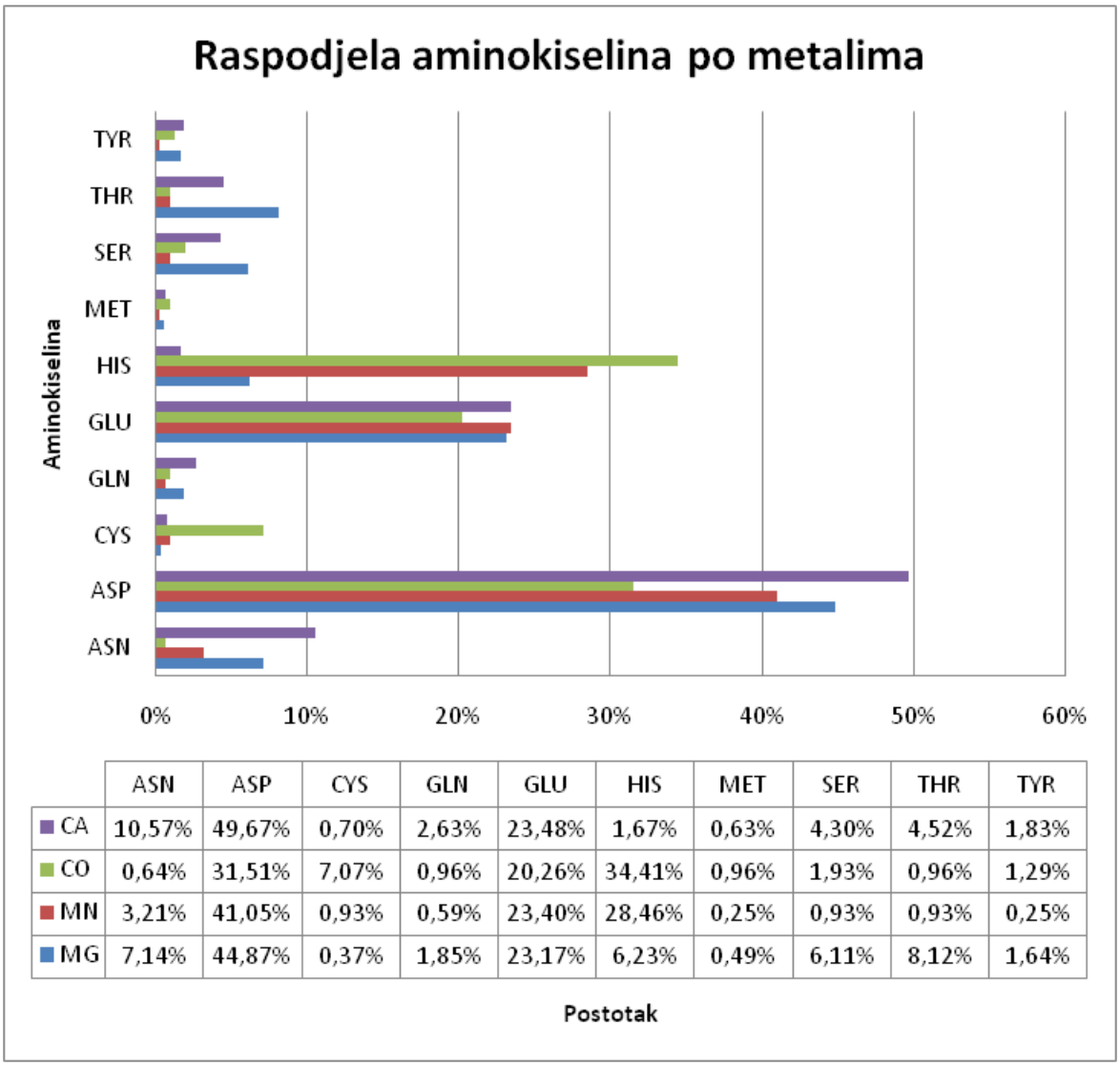
Slika 5.2 Raspodjela aminokiselina po metalima za CU, FE i NI



	ASN	ASP	CYS	GLN	GLU	HIS	MET	SER	THR	TYR
ZN	0,25%	15,28%	37,40%	0,27%	13,25%	32,41%	0,15%	0,49%	0,38%	0,11%
CD	1,07%	27,37%	7,34%	0,53%	39,25%	20,96%	0,67%	1,60%	1,07%	0,13%

Slika 5.3 Raspodjela aminokiselina po metalima za ZN i CD

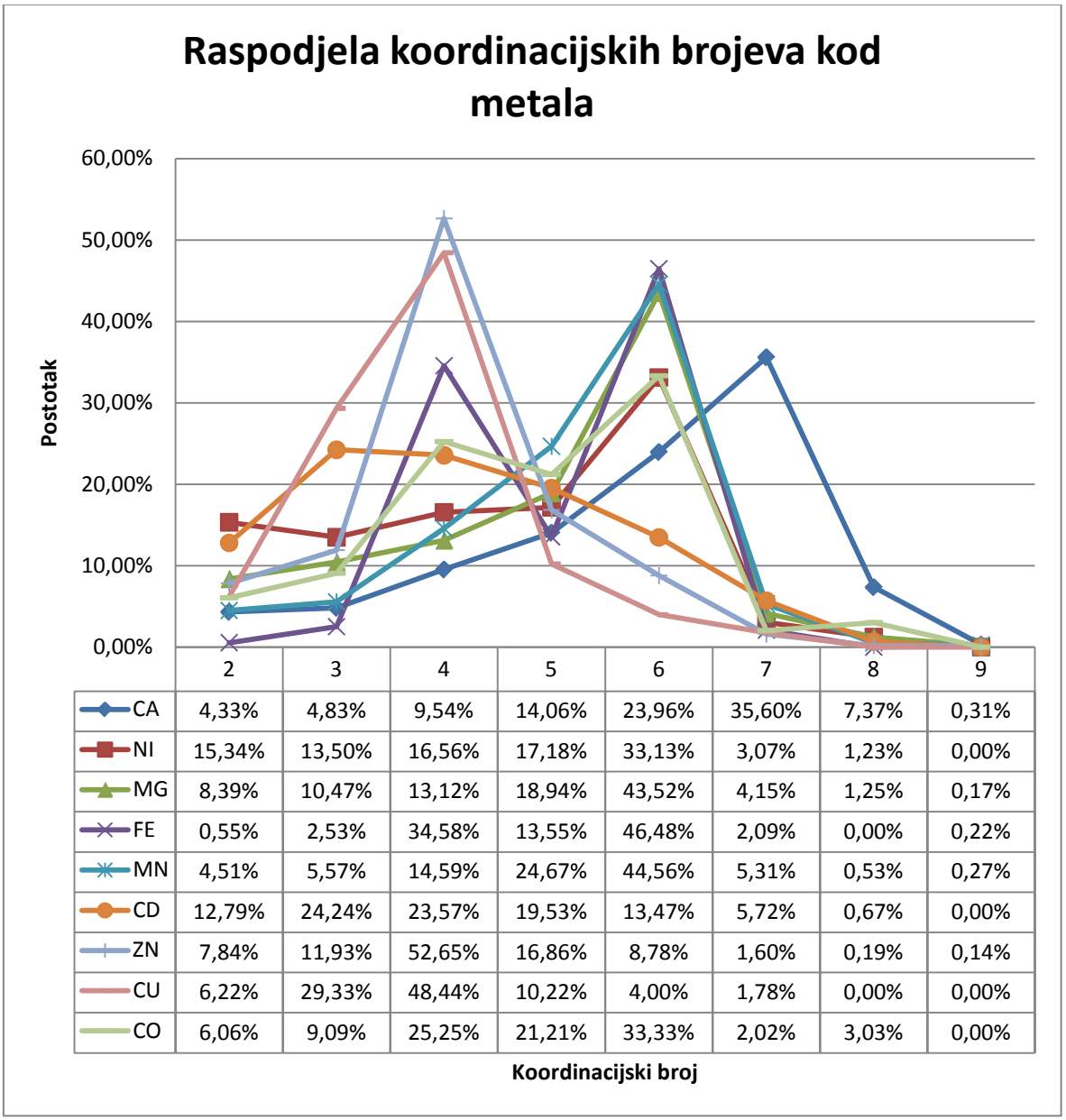
Grafikoni na slikama 5.3 i 5.4 predstavljaju gotovo iste rezultate uz vrlo male varijacije u udjelima pojedinih metala uz donore navedenih aminokiselina. Međutim na grafikonu prikazanom slikom 5.2 vidljive su drastične razlike također uzrokovane korištenjem različite metode opisane u prethodnom poglavlju. Za metale Cu, Fe i Ni se u odnosu na rezultate iz [5] pojavljuju ili nestaju neki zamjetni pikovi uz sve prikazane aminokiseline osim Gln.



	ASN	ASP	CYS	GLN	GLU	HIS	MET	SER	THR	TYR
CA	10,57%	49,67%	0,70%	2,63%	23,48%	1,67%	0,63%	4,30%	4,52%	1,83%
CO	0,64%	31,51%	7,07%	0,96%	20,26%	34,41%	0,96%	1,93%	0,96%	1,29%
MN	3,21%	41,05%	0,93%	0,59%	23,40%	28,46%	0,25%	0,93%	0,93%	0,25%
MG	7,14%	44,87%	0,37%	1,85%	23,17%	6,23%	0,49%	6,11%	8,12%	1,64%

Slika 5.4 Raspodjela aminokiselina po metalima za CA, CO, MN i MG

Na slici 5.5 je prikazana raspodjela koordinacijskih brojeva (2-9) za različite metale. U rezultatima iz ovog rada možemo primijetiti da se javlja pojačano grupiranje vrijednosti za koordinacijske brojeve 4 i 6 u području od 20% do 50%. U rezultatima prikazanim u [5] to nije slučaj i vrijednosti su razvučene do gotovo 70% za navedene koordinacijske brojeve.



	2	3	4	5	6	7	8	9
CA	4,33%	4,83%	9,54%	14,06%	23,96%	35,60%	7,37%	0,31%
NI	15,34%	13,50%	16,56%	17,18%	33,13%	3,07%	1,23%	0,00%
MG	8,39%	10,47%	13,12%	18,94%	43,52%	4,15%	1,25%	0,17%
FE	0,55%	2,53%	34,58%	13,55%	46,48%	2,09%	0,00%	0,22%
MN	4,51%	5,57%	14,59%	24,67%	44,56%	5,31%	0,53%	0,27%
CD	12,79%	24,24%	23,57%	19,53%	13,47%	5,72%	0,67%	0,00%
ZN	7,84%	11,93%	52,65%	16,86%	8,78%	1,60%	0,19%	0,14%
CU	6,22%	29,33%	48,44%	10,22%	4,00%	1,78%	0,00%	0,00%
CO	6,06%	9,09%	25,25%	21,21%	33,33%	2,02%	3,03%	0,00%

Slika 5.5 Raspodjela koordinacijskih brojeva po metalima

6 Zaključak

Prisutnost metala u proteinima možemo iskoristiti i primijeniti u različitim područjima poput čišćenja zagađenih područja od teških metala i drugih spojeva. Kako bismo to mogli potrebno je poznavati funkciju pojedinog proteina i ulogu metala u proteinu. Za poznavanje navedenoga su potrebna temeljita istraživanja i kvalitetan izvor podataka, a sistematizacija podataka uvelike doprinosi njihovoj čitljivosti i razumljivosti. Time se brže dolazi do željenih odgovora i povećava efikasnost. Upravo tome služi baza podataka dobivena kao rezultat ovog rada.

Sustav omogućuje brzu obradu velike količine podataka. Oni se pohranjuju u bazu podataka iz koje je moguće vršiti daljnju analizu ili prijenos podataka u druge sustave. Jednostavnom implementacijom omogućena je laka nadogradnja i brza prilagodba novim zahtjevima

7 Literatura

1. The Protein Data Bank (PDB). *An Information Portal to Biological Macromolecular Structures.* [Mrežno] lipanj 2010. http://www.rcsb.org/.

2. **Peretin, Goran.** Baza zastupljenosti metala u proteinima. *Diplomski rad.* Zagreb : Fakultet elektrotehnike i računarstva, 2010.

3. **Janjić, Saša.** Predviđanje mjesta sekundarne strukture proteina iz slijeda aminokiselinskih ostataka. *Diplomski rad.* Zagreb : Fakultet elektrotehnike i računarstva, 2010.

4. File:Myoglobin.png. *Wikipedia.* [Mrežno] lipanj 2010. http://en.wikipedia.org/wiki/File:Myoglobin.png.

5. *Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination.* **Dokmanic I., Šikić M., Tomić S.** Zagreb : Acta Crystallographica, srpanj 2007, Svez. Biological Crystallography. Acta Cryst. (2008). D64, 257–263.

6. Phytoremediation. *Wikipedia.* [Mrežno] lipanj 2010. http://en.wikipedia.org/wiki/Phytoremediation.

7. Metal Sites in Proteins - MESPEUS database. *University of Edinburgh.* [Mrežno] lipanj 2010. http://eduliss.bch.ed.ac.uk/MESPEUS/_1.jsp.

8. Metalmine. *NARA INSTITUTE of SCIENCE and TECHNOLOGY.* [Mrežno] lipanj 2010. http://metalmine.naist.jp/metalmine009/index.html.

9. MIPS. *Metal Interactions in Protein Structures.* [Mrežno] lipanj 2010. http://dicsoft2.physics.iisc.ernet.in/mips/.

10. Macromolecular Structure Database Group mmCIF Information. *European Bioinformatics Institute.* [Mrežno] lipanj 2010. http://www.ebi.ac.uk/msd/documentation/mmcif.html.

11. RCSB Protein Data bank. *The Macromolecular Crystallographic Information File (mmCIF).* [Mrežno] lipanj 2010. http://mmcif.rcsb.org/pubs/methenz.html.

12. Protein Crystallography. *Protein Crystallography.* [Mrežno] lipanj 2010. http://proteincrystallography.org/.

13. Sun Microsystems. *What is Java?* [Mrežno] lipanj 2010. http://www.java.com/.

14. The BioJava project. *BioJava.* [Mrežno] lipanj 2010. http://www.biojava.org/.

15. O|B|F. *Open Bioinformatics Foundation.* [Mrežno] lipanj 2010. http://open-bio.org/.

16. MySQL The world's most popular open source database. *MySQL.* [Mrežno] lipanj 2010. http://www.mysql.com/.

17. BioPython. *BioPython*. [Mrežno] lipanj 2010. http://biopython.org/

BAZA PODATAKA METALA U PROTEINIMA

Sažetak:

U ovom radu je pružen kratak uvod u ulogu metala u proteinima i nekim dosadašnjim upotrebama njihovih svojstava. Ostvarena je implementacija sustava i baze podataka koji iz centralne baze proteina preuzima sve strukture u mmCIF obliku, obrađuje ih te u bazu podataka pohranjuje strukturu proteina, udaljenosti između donora i metala te vrijednosti kutova između dva donora i metala.

Podatci se obrađuju tako da se odabiru oni proteini koji sadrže barem jedan metal i jedan proteinski lanac s dva donora. Ukoliko protein zadovoljava navedene uvjete izdvajaju se svi metali i donori, računaju se njihove međusobne udaljenosti i sve moguće vrijednosti kutova dvaju donora i metala tako da je metal u vrhu kuta i donori u krakovima.

Baza podataka dobivena ovim radom služi kao bogat izvor informacija o odnosima pojedinih metala s ostatkom proteinske strukture te je primjenjiva u biološkim i kemijskim istraživanjima.

Ključne riječi: metal, protein, lanac, koordinacijski broj, aminokiselina, nukleinska kiselina, ligand, donor, kut, udaljenost