

SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 78

**BAZA ZASTUPLJENOSTI METALA U  
PROTEINIMA**

Goran Peretin

Zagreb, lipanj 2010.

*Zahvaljujem Mili Šikiću, Fakultet elektrotehnike i računarstva, za stručno savjetovanje i potporu prilikom izrade ovog rada.*

*Zahvaljujem svojoj obitelji na neizmjernoj podršci.*

## Sadržaj

1. Uvod .....	1
2. Pregled područja.....	2
3. Podaci.....	4
3.1. Struktura podataka .....	4
3.2. Baza podataka .....	5
3.3. Cluster 70 skup .....	6
3.4. Statistike.....	7
4. Implementacija .....	12
4.1. Arhitektura aplikacije .....	12
4.2. Važniji razredi poslužiteljske strane aplikacije.....	14
4.3. Korisničko sučelje.....	17
4.4. Važniji razredi korisničkog sučelja.....	19
4.5. Korišteni alati.....	20
5. Rezultati.....	22
6. Zaključak.....	26
7. Literatura.....	27

## 1. Uvod

Velik broj metala uključen je u razne kemijske procese kao i biološke aktivnosti proteina. Identificiranje veza između tih metala i proteina omogućuje točnije određivanje značaja proteina, ali i uloge samog metala. Povećanjem računalne moći otvaraju se nove mogućnosti analize podataka o metalima te njihovim funkcijama unutar proteina.

Cilj ovog rada je omogućiti pregled statističkih podataka o prisutnosti pojedinih metala u proteinima, kao i učestalosti pojavljivanja više metala u nekom proteinu. Potrebno je ostvariti analizu podataka o prisutnosti metala u proteinima, njihovom koordinacijskom broju i aminokiselinskim ostacima te ligandima na koje se vežu.

U drugom poglavlju rada dan je pregled prijašnjih istraživanja i radova u području analize povezanosti metala i proteina. Sljedeće poglavlje donosi opis strukture podataka, baze podataka te skupa *Cluster 70*. Također su opisane sve statistike koje se izračunavaju na definiranom skupu proteina ili proteinskih lanaca. Četvrto poglavlje sadrži objašnjenje arhitekture aplikacije, poslužiteljske strane i korisničkog sučelja, a dan je i kratki opis alata korištenih za implementaciju. Zadnji dio rada prikazuje rezultate izračuna svim statistika.

## 2. Pregled područja

Metali su vrlo važni za biološke aktivnosti proteina i njihova zamjena ili uklanjanje iz proteina je vrlo često praćena gubitkom ili smanjenjem biološke aktivnosti proteina. Također, metali su vrlo često uključeni u važne kemijske procese u prirodi, npr. magnezij u klorofilu je vrlo važan za fotosintezu, dok bakar i željezo imaju važnu ulogu u proteinima koji prenose kisik.

Poznavanje vrste i količine aminokiselina koje koordiniraju metal je vrlo važno za utvrđivanje njihove povezanosti s ulogom proteina. Svi ovi podaci su vrlo značajni za otkrivanje načina na koji protein odabire određeni metal, te ulogu tog metala u proteinu.

Ovaj rad, proširivanjem skupa metala i statistika koje se računaju, predstavlja nastavak istraživanja veza metala u proteinima i njihovih karakteristika koje su proveli Dokmanić i ostali [1].

Marjorie M. Harding je provela nekoliko istraživanja geometrije povezanosti metala u proteinima, ali na dosta manjem skupu podataka i bez pokušaja da odredi specifičnosti mjesta vezanja metala, kao što su distribucija atoma po koordinacijskim brojevima i kombinacija aminokiselina uključenih u koordinaciju. Njena istraživanja dostupna su na Internetu u obliku web aplikacije [2]. Podržan je skup od 10 metala te mogućnost izrade statistika koje uključuju veze prema više molekula. Nedostaci su što ograničenje opsega pretraživanja na određeni protein ili lanac ne radi, već se uvijek pretražuju svi. Također, aplikacija je vrlo nestabilna u radu.

MIPS (eng. *Metal Interactions in Protein Structures*) je još jedan sličan projekt [3]. Aplikacija se redovito održava te omogućuje detaljno specificiranje vrsta interakcija koje nas zanimaju. Nažalost, ne daje nikakve statistike povezanosti metala, već samo ispisuje PDB kodove proteina u kojima se nalaze metali koji zadovoljavaju zadane kriterije.

Rezultati prikazani ovdje vrlo su važni za identificiranje veza između metala i ostalih molekula u proteinima jer poznavanje geometrije i karakteristika tih veza u

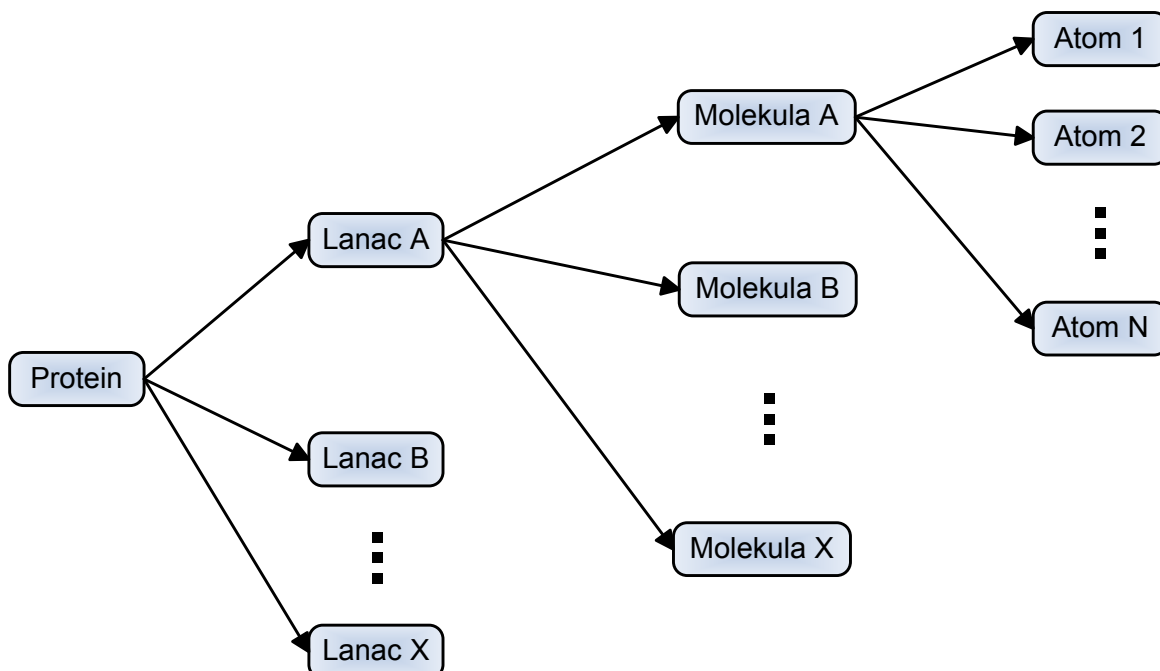
proteinima koji imaju poznatu ulogu pomaže da se točnije odrede biološke aktivnosti proteina s nepoznatom ulogom te izradi dizajn proteina koji će imati posebni afinitet prema određenim metalima.

Proteini se od nedavno razmatraju kao alat za uklanjanje teških metala iz okoliša kako bi se smanjilo zagađivanje. Stoga je vrlo bitno poznavati okolinu metala jer to može pomoći u odabiru proteina koji će biti učinkovitiji u uklanjanju određenih metala iz okoliša od onih već dostupnih.

## 3. Podaci

### 3.1. Struktura podataka

PDB (eng. *Protein Data Bank*) je baza proteinskih 3D struktura koja sadrži podatke o proteinima. Proteini se u aplikaciji identificiraju preko svog PDB koda. To je niz od četiri znaka pod kojim je struktura proteina zapisana u PDB datoteci na PDB portalu [4]. Jedan protein može imati više lanaca koji su unutar proteina identificirani jednim znakom. Lanac može biti proteinski, voda, ligand ili nukleinska kiselina. U ovom radu pretpostavljeno je da je lanac proteinski ako se sastoji od barem 50 aminokiselina. Ukoliko se sastoji samo od baza, tada se radi o nukleinskoj kiselini. Ukoliko lanac nije ni protein ni nukleinska kiselina, onda je ligand, odnosno voda ako se sastoji samo od molekula HOH. Svaki lanac može imati više molekula, a proteinski lanac može se sastojati samo od aminokiselina od kojih je barem 90% iz skupa standardnih 20 aminokiselina. Molekula se sastoji od više atoma koje dijelimo na metale i nemetale. Opisana struktura prikazana je na slici 1.



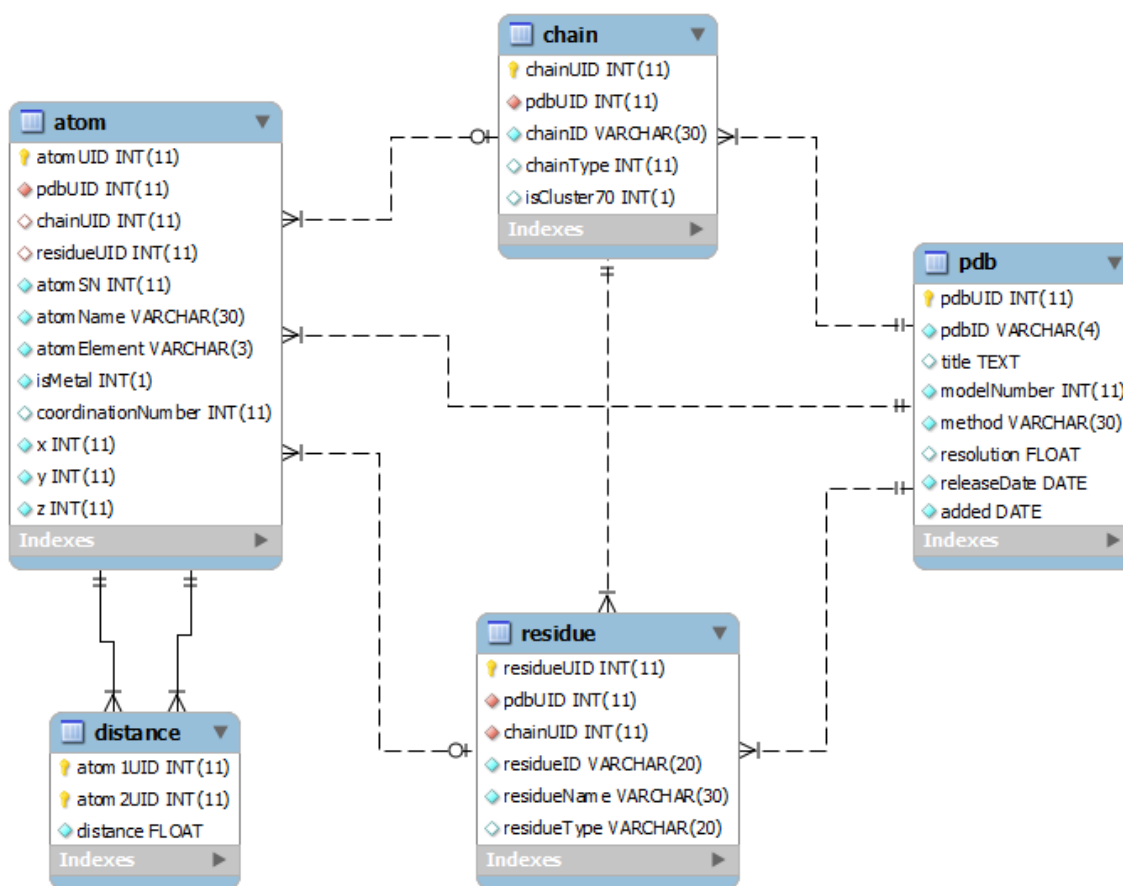
Slika 1. Struktura podataka

Tema ovog rada i pripadajuće aplikacije je odnos atoma metala prema drugim atomima i molekulama. PDB i pripadajući lanci su vrlo bitni jer se pomoću njih

definira opseg pretraživanja, tj. skup molekula i atoma koji se razmatraju prilikom izrade statistika.

### 3.2. Baza podataka

Kao izvor podataka koriste se podaci pohranjeni u relacijskoj bazi podataka [5]. Ta baza sadrži zapise o proteinima, dobivene parsiranjem PDB datoteka. Posebno je strukturirana za brzo pretraživanje podataka uvođenjem redundancije koja smanjuje potrebu za spajanjem tablica te samim tim omogućuje brži dohvat podataka. Tablice baze podataka s pripadajućim vezama prikazane su na slici 2.



Slika 2. Dijagram baze podataka

Tablica *atom* sadrži zapise o atomima, a bitni podaci su *atomName*, *atomElement*, *isMetal* te *coordinationNumber*. Polje *isMetal* označava da li se radi o atomu metala ili ne. Radi optimizacije, u ovu tablicu su ubačeni strani ključevi ostalih tablica čime je smanjena potreba za spajanjem tablica. Tablica *chain* sadrži zapise o proteinskim lancima, polje *chainType* označava da li je lanac protein, voda,



ligand ili nukleinska kiselina, *isCluster70* označava hoće li se lanac uzimati u obzir kod računanja statistika za *cluster 70* skup, objašnjen u poglavlju 3.3. Tablica *pdb* sadrži zapise o proteinima, a aplikacija za rad koristi polja *method* i *resolution*. *Method* označava metodu kojom je dobivena 3D struktura proteina. Postoje dvije metode određivanja strukture proteina, kristalografija koja koristi X-zrake (eng. X-RAY) i metoda koja koristi magnetsku rezonancu (eng. *Nuclear Magnetic Resonance*). Kada se koristi metoda kristalografije, rezultati su dobiveni s određenom rezolucijom koja označava preciznost, manja rezolucija znači veću preciznost. Rezolucija se nalazi u tablici *pdb*, polje *resolution*. Tablica *residue* sadrži podatke o molekulama. Iz te tablice koristi se podatak *residueName* što je troslovna oznaka za aminokiselinu. Tablica *distance* sadrži udaljenosti između svih atoma metala i atoma elektron donora u bazi i ključna je za određivanje da li neki metal interagira s drugim metalom ili aminokiselinom. Elektron donori su atomi elemenata (O, N, Cl, S) i bitni su zato što se metali vežu s njima te preuzimaju elektrone.

### 3.3. Cluster 70 skup

*Cluster 70* je skup proteinskih lanaca grupiranih u grozdove (eng. *cluster*) unutar kojih je sličnost sekvenci minimalno 70%. Datoteka koja sadrži taj skup se generira tjedno te se može preuzeti s RCSB (eng. *Research Collaboratory for Structural Bioinformatics*) poslužitelja [6].

*Cluster 70* datoteka je organizirana kao tablica gdje jedan redak označava jedan lanac. Stupci su broj grozda, rang i identifikator lanca. Manji rang označava bolji lanac, u smislu da oni lanci s najmanjim rangom najbolje predstavljaju grozd u kojem se nalaze. Identifikator lanca je oznaka oblika *pdb\_kod:lanac* te se pomoću nje identificira lanac i protein na koji se odnosi. Budući da su lanci unutar grozda vrlo slični, za potrebe ovog rada iz svakog grozda se uzima samo prvi lanac, tj. onaj s najmanjim rangom.

Budući da se popis *Cluster 70* osvježava tjedno, potrebno je osigurati da se prilikom računanja statistika koriste najnoviji podaci iz popisa. Pristupanje popisu na RCSB poslužitelju prilikom svakog izračunavanja bi bilo vrlo sporo i neefikasno pa je u tu svrhu razvijen program koji automatski ažurira lokalnu bazu podataka

jednom tjedno. Program se spaja na poslužitelj, preuzima popis te odgovarajućim lancima u bazi podataka postavlja ili briše zastavicu koja označava da lanac pripada skupu *Cluster 70*. Ažuriranje se obavlja redom po lancima u *Cluster 70* datoteci, od onog s najmanjim rangom do onog s najvećim. Prvi lanac na koji se naiđe, a da postoji u bazi podataka u tablici *chains*, označava se kao *cluster 70* lanac te obrada nastavlja s novim grozdom.

### 3.4. Statistike

Sve statistike računaju se na skupu proteina i/ili proteinskih lanaca. Taj skup određuje korisnik, a mogući skupovi su:

1. **Jedan protein** – kod izračunavanja statistika gledaju se samo atomi koji su unutar lanaca koji pripadaju ovom proteinu. Protein se zadaje unosom njegovog PDB koda.
2. **Više proteina ili lanaca** – ovdje je moguće zadati više proteina ili proteinskih lanaca koji će sudjelovati u računanju statistika. Svi atomi koji pripadaju zadanim lancima ili lancima koji pripadaju zadanim proteinima sudjeluju u računanju statistika. Za zadavanje proteina koristi se njihov PDB kod, dok se za zadavanje lanaca koristi oznaka *pdb\_kod:lanac*, npr: 2J7A:C.
3. **Cluster 70** – statistike se računaju samo za lance koji pripadaju skupu *Cluster 70*.
4. **Svi proteini** – statistike će se izračunavati na skupu svih proteina koji se nalaze u bazi podataka.

Statistike se izračunavaju po metalima. Dakle, za svaki metal koji korisnik odabere u korisničkom sučelju aplikacije, rade se sve statistike. Korisnik odabire i molekule za koje se rade statistike, dakle samo atomi iz označenih molekula se razmatraju prilikom računanja statistike. Budući da se statistike računaju za svaki metal odvojeno, u nastavku će se razmatrati računanje samo za jedan metal.

## 1. Ukupna količina pojedinog metala koordinirana s bar 2 atoma iz istog lanca proteina

U ovom radu podrazumijeva se da količina metala označava broj atoma tog metala. Atom je koordiniran s drugim atomom ako je njihova udaljenost manja od 3Å, odnosno manja od udaljenosti koju korisnik definira preko korisničkog sučelja. Ova statistika ujedno i ograničava sve ostale zato što se ostale statistike rade na skupu atoma koji su koordinirani s bar 2 atoma iz istog lanca, dakle, koji su rezultat ove statistike. Npr., ukoliko želimo dobiti statistike za metal Fe, statistike će se računati samo nad onim atomima Fe koji zadovoljavaju gornji uvjet.

Primjer rezultata statistike:

Total atoms of this metal coordinated with at least 2 atoms from same protein chain: 84

## 2. Količina veze između pojedinih metala i odabranih aminokiselina

Ova statistika govori koliko je metal povezan s određenim aminokiselinama, koje se također zadaju preko korisničkog sučelja. Za svaki od atoma koji zadovoljavaju uvjet iz prve statistike, računa se da li je u vezi s aminokiselinom. Atom je u vezi s aminokiselinom ako je barem jedan atom iz te aminokiseline udaljen manje od 3Å od tog atoma. Na kraju se zbroje veze svih atoma nekog metala te se onda kaže da metal ima toliko veza s aminokiselinom.

Primjer rezultata statistike:

Total connection between metal and aminoacids:

Aminoacid	Total connections
<b>HIS</b>	2604
<b>CYS</b>	1686

## 3. Postotak veze pojedine aminokiseline na metale

Ova statistika je jedina koja se ne računa po metalima, nego po aminokiselinama. Statistika govori koliko svaki metal sudjeluje u vezama s tom aminokiselinom i radi

se samo za označene aminokiseline. Veza između aminokiseline i atoma definira se isto kao i veza između atoma i aminokiseline, opisana u poglavlju 2.

Primjer rezultata statistike:

Percentage of metal connections for aminoacid:

Metal	Percentage
<b>Fe</b>	76%
<b>Cu</b>	10%
<b>Mn</b>	14%

#### 4. Distribucija metala po koordinacijskom broju

Ova statistika nam govori kako su raspoređeni atomi nekog metala po koordinacijskim brojevima. Dakle, ispisuje se koliko metal ima atoma za svaki koordinacijski broj.

Primjer rezultata statistike:

Distribution by coordination number:

Coordination number	Number of atoms
<b>4</b>	512
<b>5</b>	1043
<b>6</b>	785

#### 5. Kombinacije aminokiselina po koordinacijskom broju

Ova statistika daje prikaz kombinacija aminokiselina s kojima interagira metal, po koordinacijskim brojevima metala. Kombinacije aminokiselina koje se traže zadaju se preko sučelja. Za svaki koordinacijski broj zadanog metala, gledaju se svi atomi s tim koordinacijskim brojem. Za svaki od tih atoma dohvaćaju se označene aminokiseline. Ukoliko atom interagira sa svim aminokiselinama, on ulazi u statistiku. Ukoliko atom interagira samo s nekim od označenih aminokiselina (ne svima) on ne ulazi u rezultat, dakle nužno je da interagira sa svim označenim

aminokiselinama. Zbroj aminokiselina u kombinaciji ne mora uvijek biti jedna koordinacijskom broju. Moguće je da atom interagira s drugim aminokiselinama (koje nisu označene i ne ulaze u statistiku) ili vodama.

Primjer rezultata statistike:

Amino acid combinations by coordination number:

Coordination number	Combination	Total of that combination
<b>4</b>	2 HIS; 2 CYS	12
	1 HIS; 1 CYS	6
	1 HIS; 2 CYS	7
<b>5</b>	2 HIS; 3 CYS	10
	2 HIS; 2 CYS	4
<b>6</b>	3 HIS; 3 CYS	7
	1 HIS; 3 CYS	9

## 6. Distribucija metala vezanih za isti lanac

Ova statistika govori kako je metal povezan s drugim metalima unutar svog lanca. Statistika se računa za svaki atom metala, a na kraju se vrijednosti zbroje. Atom je povezan s nekim metalom ako je udaljen manje od 7Å od bilo kojeg atoma tog metala i ako su povezani istim proteinskim lancem. Ta udaljenost se može konfigurirati kroz sučelje aplikacije.

Primjer rezultata statistike:

Distribution of metals in same chain:

Metal	Connections
<b>Fe</b>	312
<b>Zn</b>	23
<b>Cu</b>	6

## 7. Srednja udaljenost i standardna devijacija po tipovima i imenima atoma

Računaju se srednja udaljenost i standardna devijacija između atoma zadanog metala i atoma elemenata O, N, S, Cl. Rezultati se ispisuju tipu i imenu atoma, a za svaki od njih je navedena srednja udaljenost i standardna devijacija.

Primjer rezultata statistike:

Average distance and standard deviation:

Atom element	Atom name	Average distance	Standard deviation
Cl	Cl	2.4095454	0.0483009
N	N	2.2782335	0.4201212
N	N A	2.0630001	0.0120000
O	O	2.2623243	0.3068362
O	O1	2.1740085	0.2987845
S	S	2.4127222	0.2252334
S	S1	2.2647468	0.0876632

## 8. Postotak mono i bidentalno koordiniranih metala s ASP i GLU aminokiselinama

Monodentalno koordiniran metal je onaj koji je koordiniran samo s jednim atomom kisika iz karboksilne skupine, a bidentalno koordiniran je onaj koji je koordiniran s oba kisika. Ova statistika se računa samo ako je označena ASP ili GLU aminokiselina ili obje. Atomi koji se gledaju kod računanja ove statistike su OE1 i OE2 kod GLU aminokiselina i OD1 i OD2 kod ASP aminokiseline.

Primjer rezultata statistike:

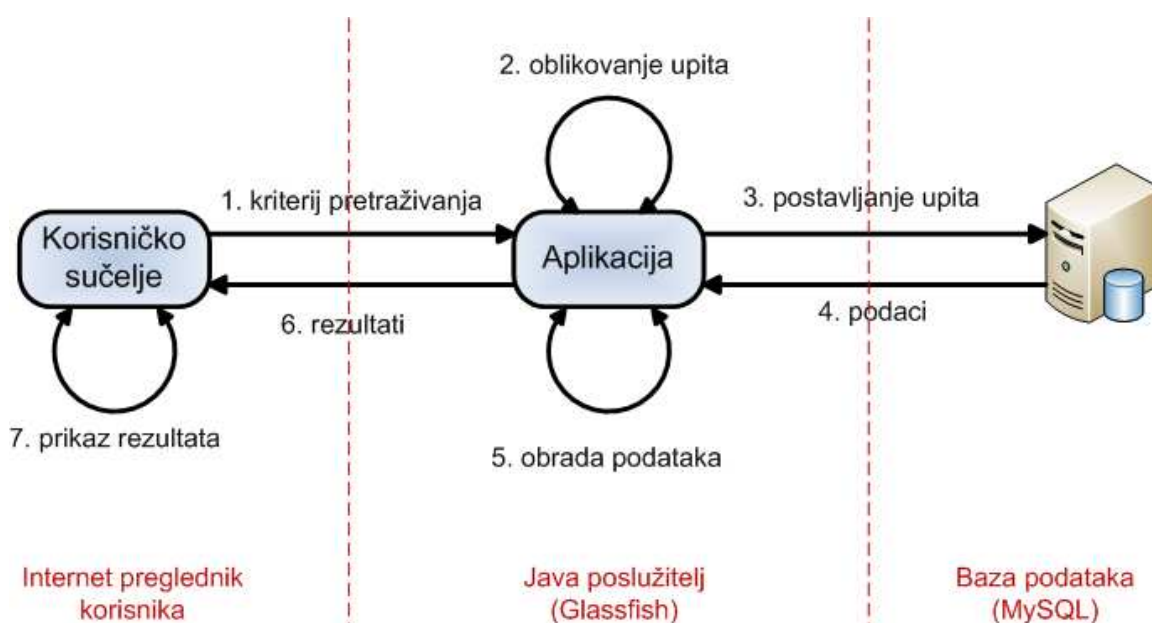
ASP	
Monodental	27
Bidental	4

GLU	
Monodental	13
Bidental	5

## 4. Implementacija

### 4.1. Arhitektura aplikacije

Aplikacija je implementirana kao web aplikacija te se bazira na klijent – poslužitelj arhitekturi. Klijent (korisnikov Internet preglednik) šalje kriterije za izračun statistika poslužitelju, koji dohvaća podatke iz baze, izračunava statistike te vraća rezultate pregledniku koji ih prikazuje korisniku. Proces obrade jednog zahtjeva prikazan je na slici 3.



Slika 3. Proces obrade jednog zahtjeva

Korisnik, koristeći internet preglednik, zadaje kriterije za izradu statistike te se oni šalju poslužitelju. Poslužitelj interpretira kriterije te sastavlja upite za bazu podataka. U ovom koraku se određuju proteini i lanci unutar kojih se pretražuje te metali i aminokiseline za koje se računaju statistike. Za svaki metal postavlja se 4 upita bazi podataka. Iako značenjem nisu povezani, upiti za više statistika često su spojeni u jedan upit radi smanjivanja opterećenja na bazu podataka te povećanja brzine rada. Nakon dobivanja rezultata upita, podaci se obrađuju te se u ovom koraku računaju svi podaci za sve statistike. Alternativa ovom pristupu je da se podaci dobiveni od baze podataka šalju direktno klijentskoj strani te se tamo računaju statistike. Prednost toga je što se od poslužitelja do klijenta prenosi manja količina podataka. Nedostatak je taj što je računanje statistika u

korisnikovom web pregledniku znatno sporije nego na poslužitelju, točnije, puno je efikasnije izračunati statistike na poslužitelju u Java programskom jeziku te prenositi veću količinu podataka, nego prenositi manju količinu podataka i računati u korisnikovom pregledniku koristeći JavaScript skriptni jezik. Nakon što se izračunate statistike pošalju klijentu, on ih samo prikazuje u pregledniku, bez ikakve obrade.

Jedan zahtjev za izračunom statistika računa se u jednoj dretvi. Većina današnjih aplikacijskih poslužitelja radi na način da za svaki HTTP zahtjev koji primi od klijenta stvara novu dretvu te aplikaciju koja ga obrađuje izvršava u njoj. Dakle, za svaki izračun statistika aplikacijski poslužitelj kreira novu dretvu.

Svi podaci koje korisnik unese validiraju se na klijentskoj strani aplikacije. Prednost ovog pristupa je što se podaci ne šalju na obradu na poslužitelj ukoliko nisu ispravni. Također, znatno se smanjuje vrijeme potrebno da se provjeri ispravnost podataka, jer nije potrebno iste slati na poslužitelj te čekati potvrdu.

Provjere koje se obavljaju su :

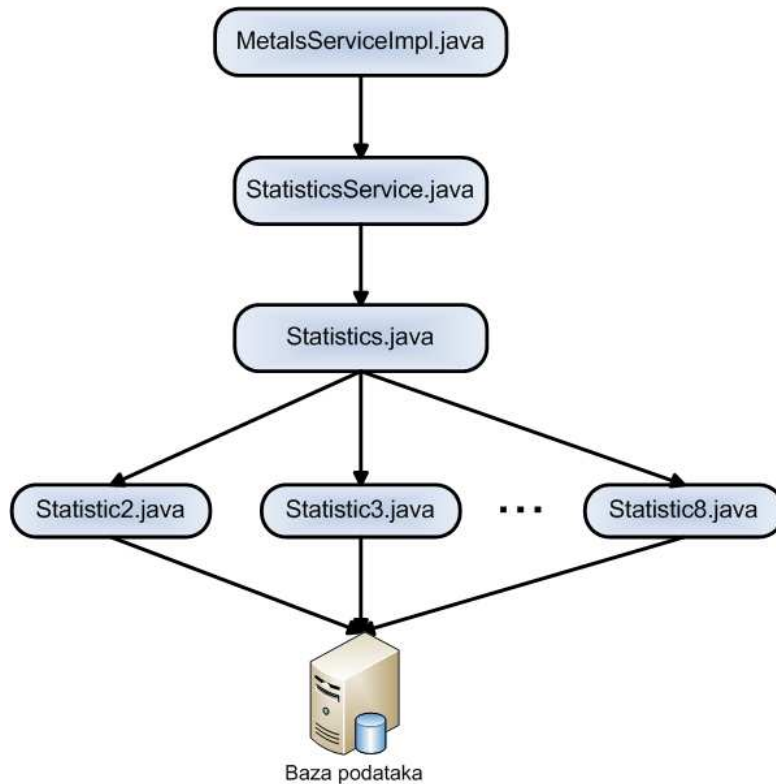
- Postojanje podatka – obavezni podaci moraju se unijeti
- Ispravni tip podatka (npr. nije moguće unijeti znakovni niz gdje se traži broj)
- Moguće vrijednosti podataka (npr. nije moguće unijeti negativnu udaljenost)

Unatoč provjeri korisničkih podataka, moguće su greške u radu aplikacije. Obrada grešaka obavlja se transparentno od korisnika. Zbog sigurnosnih razloga korisniku se ne prikazuje mjesto i razlog nastanka greške, već samo poruka da je došlo do greške u radu sustava. Svi podaci o greški zapisuju se u log datoteku koja je jedinstvena za cijelu aplikaciju. Lokacija log datoteke ovisi o aplikacijskom poslužitelju, a kod Glassfish poslužitelja nalazi se u direktoriju *instalacijski\_direktorij\_glassfisha/domains/ime\_domene/logs/server.log*. Podaci o svakoj grešci sadrže točno vrijeme greške te razred i metodu koja se izvršavala kada je došlo do prekida u radu aplikacije. Osim grešaka, u log datoteku se zapisuju i svi SQL upiti koji se izvršavaju nad bazom podataka, što vrlo često može dodatno olakšati traženje greške.



## 4.2. Važniji razredi poslužiteljske strane aplikacije

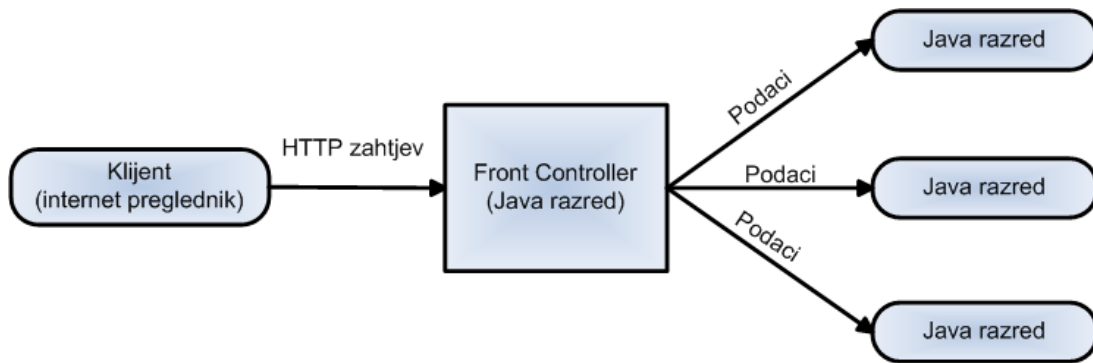
Poslužiteljska strana aplikacije sastoji se od nekoliko bitnih komponenata koje su prikazane na slici 4.



Slika 4. Arhitektura poslužiteljske strane aplikacije

Prilikom izrade aplikacije za postizanje bolje fleksibilnosti te omogućavanje proširivanja aplikacije novim zahtjevima, korišteni su oblikovni obrasci [7] (eng. *Design patterns*). Oblikovni obrazac je općenito rješenje nekog problema koji se često javlja u dizajnu programske podrške. To nije gotov dizajn razreda koji se može uklopiti u već postojeću aplikaciju, već obrazac po kojem se taj problem rješava na najbolji i najlakši način. Objektno orijentirani oblikovni obrasci su najčešće dijagrami razreda ili objekata koje programer implementira u svojoj aplikaciji kako bi riješio svoj problem.

`MetalsServiceImpl.java` je razred koji prima zahtjev od korisnikovog Internet preglednika. To je implementacija *front controller* oblikovnog obrasca prikazanog na slici 5.

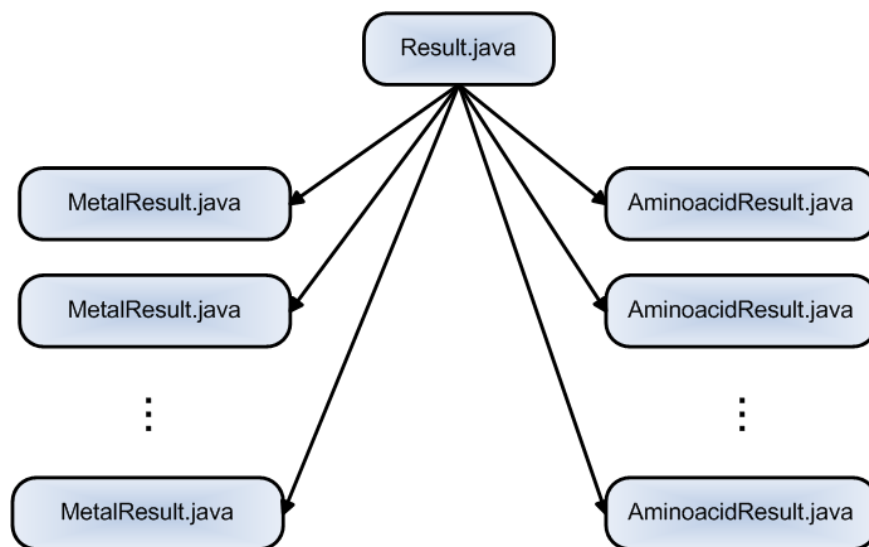


Slika 5. *Front controller* oblikovni obrazac

*Front controller* govori da svi zahtjevi od klijenta moraju stizati na jedno mjesto te se od tamo raspoređivati dalje u aplikaciju. Zahtjev se prosljeđuje razredu *StatisticsService.java* koji ga obrađuje i izvlači kriterije pretraživanja koje je korisnik zadao. Kriterije prosljeđuje razredu *Statistics.java* koji služi kao sučelje prema svakoj statistici posebno. Također, ovdje se skupljaju svi rezultati nakon što je obrada završena. Za svaku statistiku postoji poseban razred, osim za prvu koja se računa prilikom izračunavanja opsega pretraživanja budući da ona ograničava sve ostale. Ovakav način implementacije je odabran zato što statistike imaju različite ulazne podatke s kojima rade (stanje) te različite algoritme po kojima se izračunavaju (ponašanje) te bi bilo vrlo komplicirano održavati i nadograđivati aplikaciju da su statistike stavljene u isti razred. Nakon što su sve statistike izračunate, razred *Statistics.java* skuplja rezultate te se oni vraćaju istim putem kuda je zahtjev stigao.

Za održavanje veze s bazom podataka zadužen je razred *Db.java*. Ukoliko bi svaka statistika koristila zasebnu vezu na bazu podataka, bespotrebno bi se trošili resursi i opterećivao bi se sustav za upravljanje bazom podataka DBMS (eng. *Database Management System*). Stoga sve statistike koriste jedinstvenu vezu prema bazi podataka, a ta veza se održava u razredu *Db.java*. Budući da ne želimo da se u aplikaciji pojave dva takva razreda, on je realiziran kao *singleton* oblikovni obrazac, koji govori kako se u aplikaciji postiže da postoji samo jedan razred određenog tipa. To se ostvaruje tako da se u razred koji želimo da bude *singleton* stavlja privatni konstruktor koji onemogućuje instanciranje objekta tog razreda. Implementira se metoda *getInstance()* koja pri prvom pozivu interno kreira objekt te svakoj metodi koja ju pozove vraća taj lokalno kreirani objekt. Na taj način se u cijeloj aplikaciji koristi samo jedan objekt.

Za modeliranje rezultata statistika u sustavu razvijeni su posebni razredi prikazani na slici 6.



Slika 6. Struktura rezultata statistika

Razred *Result.java* je vršni rezultat, za svako računanje statistika radi se jedan ovakav razred. Taj razred sadrži dvije liste:

- Lista razreda *MetalResult.java* – za svaki metal za koji se radi statistika radi se jedan razred *MetalResult.java*. Taj razred sadrži sve statistike vezane za taj metal. Razreda ima onoliko koliko ima metala za koje se radila statistika. Statistike koje sadrži ovaj razred su sve osim statistike 3, koja se računa za aminokiseline.
- Lista razreda *AminoacidResult.java* – za svaku aminokiselinu koja sudjeluje u računanju statistika radi se jedan ovakav objekt koji sadrži rezultate za statistiku 3.

Statistika 1 je jednostavni *integer* tip podatka koji sadrži broj atoma koji zadovoljavaju prvu statistiku. Statistika 2 je realizirana kao *HashMap<String, Integer>* tip podatka. Ključ je znakovna vrijednost te označava aminokiselinu (npr. „HIS“), a vrijednost je broj koji označava broj veza između atoma za koji se radi statistika i te aminokiseline. Statistika 3 je realizirana koristeći istu strukturu podataka, ali ovdje ključ predstavlja metal (npr. „Fe“), a vrijednost je broj veza aminokiseline s tim metalom. Rezultati statistike 4 implementirani su koristeći *HashMap<String, Integer>* strukturu podataka, ključ je koordinacijski broj, a vrijednost broj atoma s tim koordinacijskim brojem. Statistika 5 je realizirana kao

*HashMap<Integer, HashMap<String, Integer>>* struktura podataka. Unutarnja *HashMap* struktura označava kombinaciju aminokiselina te broj pojavljivanja te kombinacije. Kako se te vrijednosti računaju za svaki koordinacijski broj, stavljene su u još jednu *HashMap* strukturu kojoj je ključ koordinacijski broj. Statistika 6 realizirana je kao *HashMap<String, Integer>* tip podatka, gdje ključ označava metal (npr. „Zn“), a vrijednost je broj veza s tim metalom. Rezultati statistike 7 realizirani su koristeći dvije strukture *ArrayList<String>*, jedna za rezultate grupirane po nazivu atoma te jedna za rezultate grupirane po elementu atoma. Iako bi se rezultati po elementima atoma mogli izračunati iz rezultata grupiranih po nazivu atoma te bi se time preko mreže slalo manje podataka, oni se grupiraju na poslužitelju jer se želi osloboditi klijenta od zahtjevnijih računskih operacija. Računanje na poslužitelju je znatno brže nego računanje na klijentskoj strani unutar korisnikovog Internet preglednika. Rezultati statistike 8 realizirani su kao dvije *HashMap<String, Integer>* strukture podataka, jedna za ASP aminokiselinu te jedna za GLU aminokiselinu. Ključ je oznaka monodentalno ili bidentalno, a vrijednost je broj atoma koji su koordinirani na taj način.

### 4.3. Korisničko sučelje

Korisničko sučelje se sastoji od forme za unos kriterija po kojima će se raditi statistika te forme s rezultatima. Forma za unos kriterija prikazana je na slici 7, dok je forma s rezultatima objašnjena u poglavlju 5.

Slika 7. Forma za unos kriterija

Okvir *Statistics scope* služi za određivanje opsega pretraživanja, tj. koji će proteini i lanci sudjelovati u izradi statistike. Opcija *Single PDB* označava da će se statistika raditi samo nad jednim proteinom, njegov PDB kod se unosi u polje *PDB code*. Unosi se samo 1 PDB kod, velikim ili malim slovima, npr. 2J7A. *Input PDB codes* opcija omogućuje da se unesu proizvoljni PDB kodovi nad kojima će se obavljati statistika. Ovdje je moguće unijeti i pojedinačne lance koji će se pretraživati umjesto cijelog proteina. PDB kodovi se unose kao i kod *Single PDB* opcije, a lanci se unose u formatu *pdb\_kod:lanac*, npr. 2J7A:C. Opcija *Cluster 70* označava da će se pretraživati lanci iz *Cluster 70* skupa, a *All PDBs* pretražuje sve proteine.

*Filter by method* okvir omogućuje ograničavanje opsega pretraživanja po metodi kojom su dobiveni proteini. Moguće je pretraživati proteine dobivene *XRAY* metodom uz zadavanje rezolucije skeniranja, ili samo proteine dobivene *NMR* metodom. Kod pretraživanja *XRAY* metodom, ukoliko se ne unese rezolucija sustav će pretraživati sve proteine dobivene *XRAY* metodom, dok će kod unosa rezolucije sustav pretraživati samo one proteine koji imaju manju rezoluciju od unesene.

Okvir *Distances* služi za definiranje udaljenosti ispod kojih su 2 atoma u vezi. Ukoliko se ne unesu udaljenosti koriste se podrazumijevane vrijednosti, za nemetale 3Å, a za metale 7Å.

Okvir *Coord. # for aminoacid combinations* služi za ograničavanje statistike 5. Statistika 5 koja prikazuje kombinacije aminokiselina po koordinacijskom broju će se raditi samo za koordinacijski broj koji se ovdje unese. Ukoliko se unese 0, statistika se radi za sve koordinacijske brojeve.

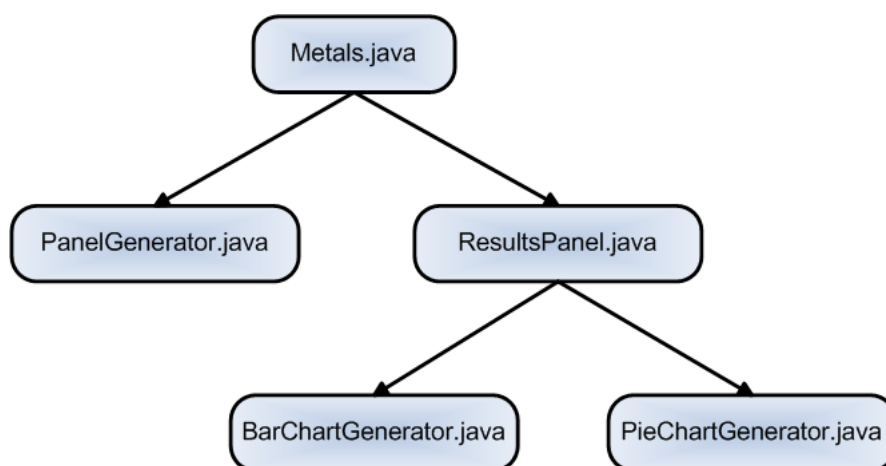
Okvir *Metals* služi za odabir metala za koje će se raditi statistika. Statistike se rade za svaki metal neovisno o drugima pa je ovdje moguće odabrati bilo koju kombinaciju. Statistika 3 koja za određenu aminokiselinu pokazuje kako je povezana s metalima će se računati samo za metale označene ovdje.

Okvir *Molecules* služi za odabir molekula koje će se uzimati u obzir kod računanja statistika. S lijeve strane ponuđeno je 20 standardnih aminokiselina, dok je s desno strane moguće odabrati pretraživanje voda i ostalih molekula. Ukoliko se ovdje označi voda, pri računanju statistika koristit će se HOH molekule, a ako se označe ostale molekule, koristit će se sve molekule koje nisu u skupu 20 standardnih aminokiselina i nisu vode (molekula HOH).

Okvir *chains* služi za ograničavanje opsega pretraživanja. U računanju statistika koristit će se samo atomi koji pripadaju označenim lancima.

#### 4.4. Važniji razredi korisničkog sučelja

Za prikaz i funkcioniranje korisničkog sučelja zaduženo je nekoliko glavnih razreda. Arhitektura je prikazana na slici 8.



Slika 8. Arhitektura korisničkog sučelja

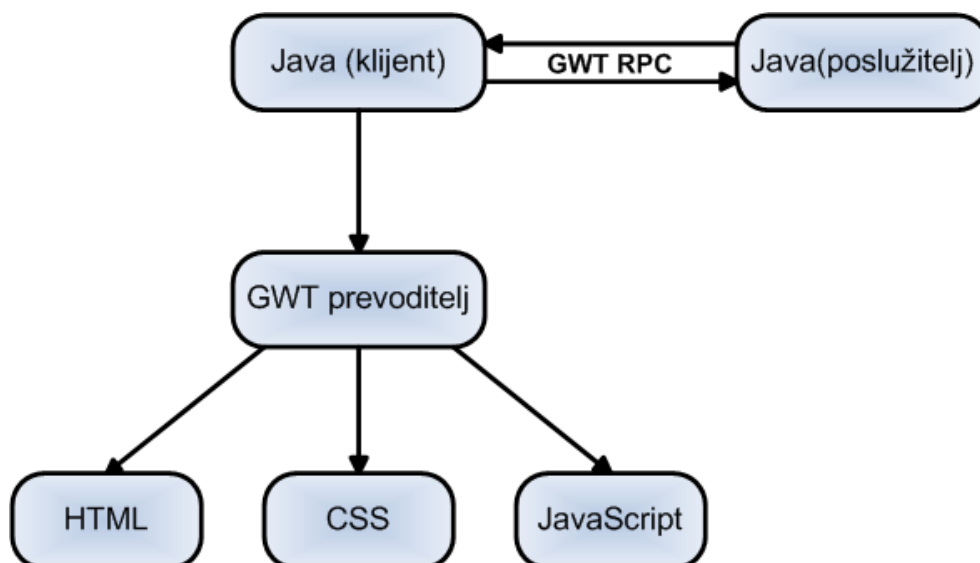
Glavni razred koji prikazuje i upravlja radom korisničkog sučelja je *Metals.java*. To je ujedno i ulazna točka programa, tj. metoda *onModuleLoad()* unutar tog razreda je prva metoda koje se pokreće kad korisnik posjeti web stranicu aplikacije. Za prikaz forme za unos kriterija ovaj razred poziva metode iz razreda *PanelGenerator.java*. Taj razred i pripadajuće metode su statičke, jer nas ne zanima stanje tog razreda, već ga koristimo kao servis. Dakle, treba nam samo da generira okvire i završi s radom.

Nakon što poslužitelj vrati izračunate statistike, razred *ResultsPanel.java* ih prikazuje korisniku. Budući da nekoliko statistika koristi grafički prikaz rezultata, napravljeni su generički razredi *BarChartGenerator.java* i *PieChartGenerator.java* koji kao primaju rezultat neke statistike te za njega naprave stupčasti odnosno tortni graf. Ovi razredi su statički iz istih razloga kao i *PanelGenerator.java*.

#### 4.5. Korišteni alati

Aplikacija je realizirana kao web aplikacija u programskom jeziku Java. Korištena je GWT (eng. *Google Web Toolkit*) biblioteka za izradu korisničkog sučelja, *OFCGWT* biblioteka za izradu grafova te *MySQL* Java biblioteka za spajanje na bazu podataka.

GWT je alat za izradu web aplikacija temeljenih na programskom jeziku Java. Njegova glavna karakteristika je prevođenje Java koda u kod koji se izvršava u Internet pregledniku korisnika, točnije, HTML, CSS i Javascript kod. To omogućuje da se cijela aplikacija izradi u programskom jeziku Java, a GWT kompilator će generirati odgovarajući kod za prikaz (HTML, CSS) i ponašanje (Javascript) web stranica. Ovo omogućuje lakše održavanje i testiranje web aplikacija te oslobađa programera od optimizacije web stranica za pojedine preglednike. Opisana arhitektura prikazana je na slici 9.



Slika 9. Generiranje koda klijentske strane aplikacije

Druga važna karakteristika GWT alata je asinkronost. HTTP protokol je protokol koji radi na principu zahtjeva i odgovora – korisnikov Internet preglednik šalje zahtjev za određenom stranicom, a poslužitelj mu vraća kod te stranice. GWT omogućuje slanje zahtjeva za određenim podatkom, a ne cijelom stranicom te osvježavanje samo određenog dijela stranice. Budući da klijentska i poslužiteljska strana koriste Java programski jezik, ovo se u GWT-u manifestira kao jednostavan poziv Java funkcije, poznato i kao RPC (eng. *Remote Procedure Call*). GWT je otvorenog koda i besplatan za sve uporabe.

OFCGWT je besplatna biblioteka otvorenog koda koja služi za generiranje grafova. Podržano je desetak tipova grafova, a postoje i velike mogućnosti prilagodbe izgleda grafova. Biblioteka radi tako da se naprave podatkovne strukture koja ona prepoznaje te se ispune podacima koji se žele prikazati. Generirani dijagrami koriste Flash tehnologiju te omogućuju animiranje. Nedostatak OFCGWT biblioteke je nepostojanje dokumentacije.



## 5. Rezultati

U nastavku su prikazani rezultati izračuna statistika za sljedeće kriterije:

- Skup pretraživanja je skup svih proteina
- Nema ograničenja na metodu kojom je dobiven PDB
- Udaljenosti su standardne, 3Å i 7Å
- Statistike se rade za metal željezo (Fe)
- Statistike se rade za sve aminokiseline

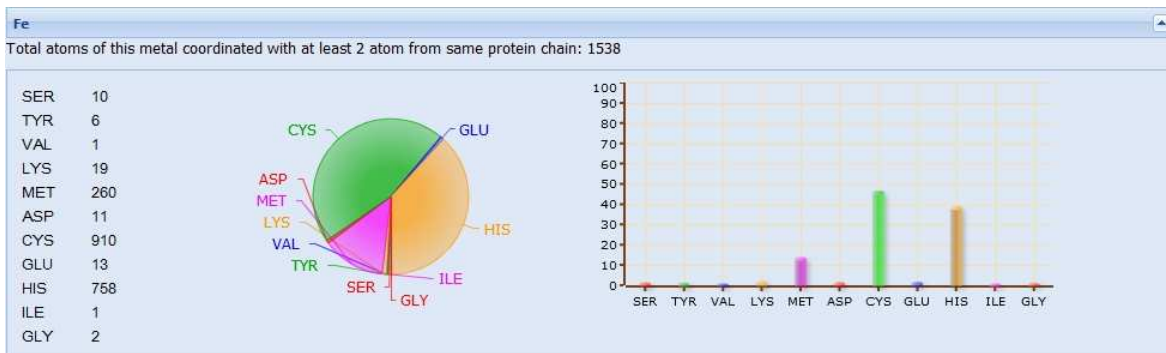
Slika 10 prikazuje izgled forme kojom se zadaje izračun statistika za koji će biti prikazani rezultati. Sve statistike u ovom poglavlju su izračunate prema toj formi, osim statistike 3 koja se računa za aminokiseline. Za tu statistiku su označeni metali Mg, Fe i Cu te HIS aminokiselina.

The image shows a web form for calculating protein statistics. It is divided into several sections:

- Statistics scope:** Includes radio buttons for 'Single PDB:', 'Input PDB codes:', 'Cluster 70', and 'All PDBs'. There are input fields for 'PDB code...' and 'Input PDBs'.
- Filter by method:** Includes radio buttons for 'Any method', 'XRAY method', and 'NMR method'. There is an input field for 'Resolution' next to the 'XRAY method' option.
- Distances:** Includes input fields for 'Donor distance is less than:' (value: 3) and 'Metal distance is less than:' (value: 7).
- Coord. # for aminoacid combinations:** Includes an input field for 'Coordination number:'.
- Metals:** A grid of checkboxes for various metals: Fe, Cu, Co, Mo, Ni, Na, Zn, W, Mn, Mg, Cd, Pb, Ca, K, V, Br. Below the grid are 'Check all' and 'Uncheck all' buttons.
- Molecules:** A grid of checkboxes for various amino acids: ALA, GLU, LEU, SER, Water, ARG, GLN, LYS, THR, Other, ASN, GLY, MET, TRP, ASP, HIS, PHE, TYR, CYS, ILE, PRO, VAL. Below the grid are 'Check all AA.' and 'Uncheck all AA.' buttons.
- Chains:** A grid of checkboxes for 'Protein', 'Ligand', 'Water', and 'Nucleic acid'.
- Submit:** A 'Submit' button at the bottom center.

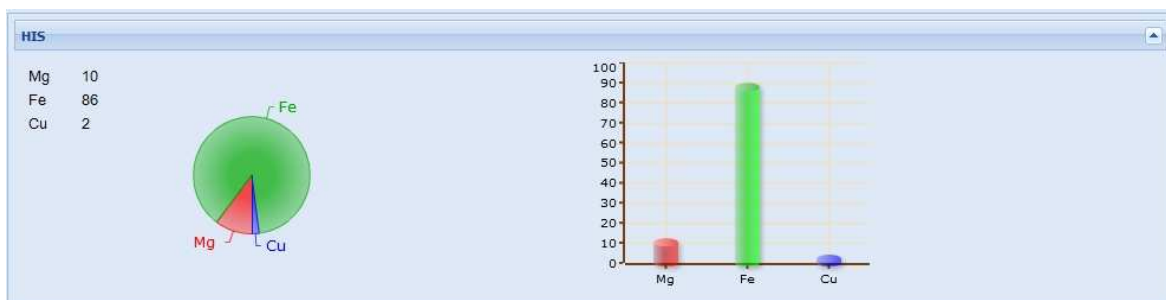
Slika 10. Izgled forme za primjer izračuna statistika

Statistike 1 i 2 prikazane su na slici 11. Statistika 1 govori da je 1538 atoma metala željeza koordinirano s barem 2 druga atoma iz istog lanca. Statistika 2 prikazana je u većem okviru. S lijeve strane je tablica koja govori koliko veza metal željeza ima s određenim molekulama. U sredini se nalazi tortni, a s desne strane stupčasti graf koji te podatke prikazuju grafički. Ovakav način prikaza gdje se s lijeve strane tablično prikažu podaci, a u sredini i desno grafički, slijede i ostale statistike.



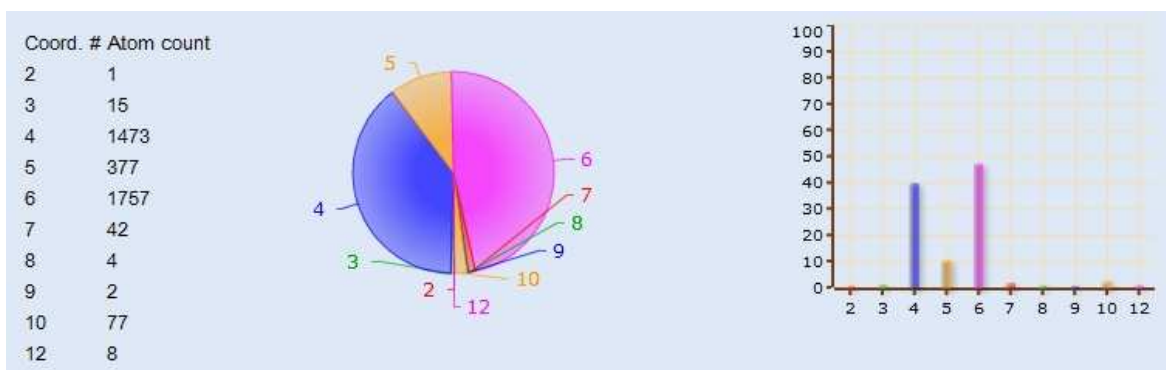
Slika 11. Statistike 1 i 2

Slika 12 prikazuje rezultate statistike 3. To je jedina statistika koja se ne računa po metalima nego po aminokiselinama. Iz ovog rezultata vidimo da HIS aminokiselina ima najviše veza sa željezom, a u manjem dijelu s magnezijem i bakrom.



Slika 12. Statistika 3

Slika 13 prikazuje rezultate statistike 4. Rezultati pokazuju koliko atoma metala željeza ima određeni koordinacijski broj. Iz ovog primjera vidimo da je najviše atoma željeza s koordinacijskim brojem 6, njih 1757.



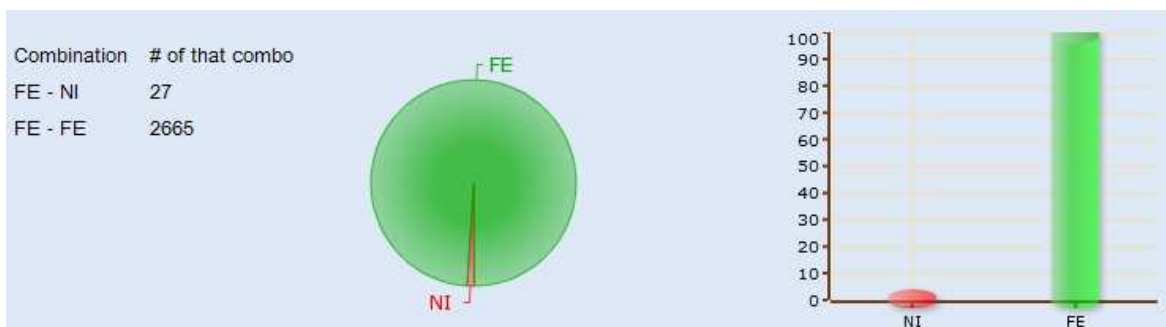
Slika 13. Statistika 4

Slika 14 prikazuje rezultate statistike 5. Ova statistika prikazuje kombinacije aminokiselina s kojima koordinira metal po koordinacijskim brojevima. Ovaj primjer pokazuje da postoji 2 atoma željeza koji imaju koordinacijski broj 6 i povezani su s jednom molekulom ASP i dvije HIS.

Coord. #	Combination	# of that combo
6	1ASP;2HIS;	2
7	1ASP;2HIS;	5

Slika 14. Statistika 5

Na Slika 15 su rezultati statistike 6 koji prikazuju veze između metala željeza i ostalih metala. Primjerice, vidimo da se atomi željeza najčešće vežu s drugim atomima željeza, a vrlo malo i s metalima nikla.



Slika 15. Statistika 6

Slika 16 prikazuje rezultate statistike 7. Prikazane su udaljenosti atoma željeza od atoma O, N, S, Cl. S lijeve strane su rezultati grupirani po elementu atoma, dok su s desne strane dani rezultati u cijelosti po imenima atoma. Ispis je skraćen radi preglednosti.

Atom element	Average distance	Standard deviation	Atom element	Atom name	Average distance	Standard deviation
N	2.1042990	0.1689749	N	N	2.5402983	0.4002938
O	2.2229762	0.2577827	N	N1	2.1670001	0.0982929
S	2.2762230	0.1063674	N	N10	2.1789999	0.0000000
			N	N2	2.9273334	0.0936815
			N	N3	2.0360000	0.0074833
			N	N8	2.1840000	0.0000000
			N	NA	2.0335441	0.0942732
			N	NB	2.0174494	0.1147882
			N	NC	2.0193689	0.1363559
			N	ND	2.0398673	0.1004568
			N	ND1	2.2349956	0.1518009
			N	ND2	1.9129999	0.1464172
			N	NE2	2.1665924	0.1551720
			N	NX	1.8940833	0.0185111

Slika 16. Statistika 7

Slika 17 prikazuje rezultate statistike 8. Vidimo da atomi željeza koordiniraju s jednim kisikom iz karboksilne skupine molekule GLU 13 puta, dok s oba ne koordiniraju. U slučaju ASP aminokiseline, postoji 6 monodentalnih atoma i 2 bidentalna.

Metal coordinated with GLU	
Monodental	13
Bidental	0
Metal coordinated with ASP	
Monodental	6
Bidental	2

Slika 17. Statistika 8

## 6. Zaključak

Prikladan i sustavan prikaz statističkih podataka o povezanosti metala i proteina omogućuje jednostavnije shvaćanje njihove uloge u biološkim aktivnostima. Pravilnom analizom podataka o metalima razvijene su statistike koje uključuju razne distribucije (npr. metala po koordinacijskim brojevima) i analize povezanosti (poput količine veze metala i odabrane aminokiseline). Također su implementirani izračuni srednje udaljenosti i standardne devijacije među atomima odabranog metala i nekoliko nemetala.

Omogućen je odabir opsega pretraživanja unosom jednog proteina ili zadavanjem proteinskog lanca, uz što je moguće definirati i pretraživanje po jednoj od metoda kojom proteini nastaju. Dohvat podataka može se ograničiti i zadavanjem udaljenosti veza dvaju atoma (odvojeno za nemetale i metale) te definiranjem nekog lanca. Statistički podaci generiraju se za proizvoljno odabran broj metala i molekula. Svi rezultati prilagođeni su za danju obradu i analizu kako bi poslužili za poboljšanje istraživanja proteina te poboljšali njihovu primjenu u nove svrhe.

## 7. Literatura

1. Dokmanić, I., Šikić, M., Tomić, S. Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. *Biological crystallography*. D64 (2008), str. 257.-263.
2. Marjorie M. Harding, Metal site sin proteins – MESPEUS database, 12.1.2008, <http://eduliss.bch.ed.ac.uk/MESPEUS/1.jsp>, 15.5.2010.
3. K.Hemavathi, M.Kalaivani, A.Udayakumar, G.Sowmiya, J.Jeyakanthan, K.Sekar – Metal Interactions in Protein Structures, <http://dicsoft2.physics.iisc.ernet.in/mips/>, 15.5.2010.
4. Protein Data Bank, *An Information Portal to Biological Macromolecular Structures*, <http://www.pdb.org/pdb/home/home.do>, 15.5.2010.
5. Tus, A., Baza podataka metala u proteinima, završni rad, Fakultet Elektrotehnike i Računarstva, 2010.
6. Clusters, <ftp://resources.rcsb.org/sequence/clusters/>, 11.6.2010.
7. Gamma, E., Helm, R., Johnson, R., Vlissides, J.M., *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley Professional, 1994.

# Baza zastupljenosti metala u proteinima

## Sažetak:

U ovom radu dan je uvod u značenje proteina i metala u biološkim procesima te pregled ostvarenih istraživanja na tom području. Implementirana je aplikacija koja analizira podatke o povezanosti metala i proteina te pruža statistički pregled dobivenih informacija.

Glavni zadatak je omogućiti pregled podataka o prisutnosti metala u proteinima, učestalosti njihova pojavljivanja te vezanju za aminokiseline. Takvi podaci mogu poslužiti u danjoj analizi i određivanju uloge proteina u okolišu.

Na kraju rada prezentirani su rezultati svih statističkih pretraga za jedan metal (željezo) i sve aminokiseline.

## Ključne riječi:

metal, aminokiselina, statistika, koordinacijski broj, lanac, protein, ligand

## Representation of metals in proteins

### Summary:

This paper outlines meaning of metals and proteins in biological processes. Also, previous achievements in this area of science are briefly introduced. Web application for analysis of connections between metals and proteins was implemented. Application presents user with statistical overview of given results.

Main task was to provide overview of statistics about metals in proteins, including how often and with which proteins these metals coordinate. These kinds of data can provide valuable insight in further analysis and determination of role that proteins have in environment.

Last chapter presents result statistics for sample Fe metal and standard 20 amino acids.

### Keywords:

metal, amino acid, statistics, coordination number, chain, protein, ligand