SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 372

BioMe - web sučelje baze biološki važnih metala

Alen Rakipović

Zagreb, lipanj 2012.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`.

Hvala mojoj obitelji i Sari na podršci. Veliko hvala Igoru Čanadiju na reviziji ovog teksta i korisnim savjetima.

SADRŽAJ

1. Uvod

Razvoj tehnologije je omogućio napredak u svim aspektima društva, a ponajviše se to odrazilo u prirodnim znanostima. Računalna moć i sve veća brzina računanja su udarili temelje novim znanostima i interdisciplinarnosti prirodnih znanosti i računarstva. Bioinformatika je jedan od predvodnika tog novog vala zbog mogućnosti ostvarivanja računalnih simulacija i analiza stvarnih pokusa i prirodnih pojava.

Razni se kemijski i biološki procesi mogu lakše opisati i analizirati ukoliko je poznato koji su proteini i metali uključeni u te procese. Identifikacija veze metala i proteina pomaže u određivanju uloge i značaja kako proteina, tako i pripadnog metala.

Ovaj rad pruža pregled statističkih podataka o povezanosti iona metala s amino i nukleinskim kiselinama, njihove distribucije po koordinacijskim brojevima te koordinacijske geometrije. Rad je dio većeg projekta (Tus i Šikić, 2012.) koji ima za cilj ostvariti jedinstvenu bazu podataka s informacijama o proteinskim lancima (Tus, 2010) te pružiti statističku obradu podataka i cjelovit prikaz navedenih statistika. Projekt je realiziran u suradnji sa dr. sc. Sanjom Tomić i B.Sc. Antonijom Tomić s Instituta Ruđer Bošković koje su s dr. sc. Mile Šikićem oblikovale funkcionalne zahtjeve sustava te pomogle u izgradnji kroz testiranje i ispravljanje netočnih rezultata statistika. Prvu verziju sustava izgradio je M.Sc. Goran Peretin (Peretin, 2010), a također je dao veliki doprinos svojim savjetima i testiranjima. Osim jezgre sustava koji čini baza podataka koja je polazište za statističku analizu, Alan Tus je izradio kostur nove verzije web sučelja. Na kraju, dr. sc. Mile Šikić je imao ključnu ulogu u vođenju tima i sudjelovao je u izgradnji sustava kroz sve etape sa svojim savjetima i idejnim rješenjima. Hvala svima!

U prvom poglavlju su opisane biomolekule kojima se bavimo u statističkoj analizi kao što su proteini i nukleinske kiseline. Također, ukratko je objašnjena uloga metala u biomolekulama te sama struktura biomolekula. Mnogo je sličnih pokušaja ostvarivanja web servisa za pregled biomolekula te je njihov pregled i usporedba s BioMe sustavom dana u drugom poglavlju. Sljedeće poglavlje donosi iscrpan pregled arhitekture statističkog analizatora, pripadnog web sučelja te opise statistika koje se računaju.

Nakon toga dani su najzanimljivi rezultati obrade podataka. Na kraju dolazi zaključak i popis korištene literature.

2. Biomolekule

2.1. Proteini

Proteini imaju ključnu ulogu u gotovo svim biološkim procesima. Većini proteina je uloga im je definirana funkcijom koju određuje njihova prostorna struktura. Proteomika je znanost koja se bavi proučavanjem svojstava, interakcija i funkcija proteina; to je znanstvena disciplina čiji je cilj opisati cjelokupnost proteina koji čine organizme (proteome).

Proteini su složene organske strukture koje se sastoje od aminokiselina povezanih peptidnim vezama, čiji je slijed određen genima koji ih kodiraju. Linearan niz aminokiselina koje tvore protein uvija se u specifičnu trodimenzionalnu strukturu koja određuje njegovu funkciju (Janjić, 2010).

Istraživanje genoma urodilo je spoznajom velikog broja aminokiselinskih sljedova koje kodiraju geni, međutim funkcija, struktura i interakcije proteina pripadnih sljedova uglavnom su nepoznate. Iz tog razloga se ulažu veliki napori kako bi se strukture odredilo eksperimentalno ili računski. Primjerice, znanstvenici se u eksperimentalnim metodama koriste modeliranjem sljedova tijekom postupka dobivanja strukture.

2.2. Nukleinske kiseline

Nukleinske kiseline su najveće organske molekule, prijeko potrebne komponente svake žive stanice. Neke od uloga su im očuvanje i prijenos genetičke informacije, biosinteza proteina, razmjena tvari i energije.

Nukleinske kiseline su polinukleotidi, koje se sastoje od velikog broja mononukleotida. Svaki nukleotid sadrži po jednu dušičnu bazu, molekulu šećera pentoze te molekulu fosforne kiseline. Dva su osnovna tipa nukleinskih kiselina u živim bićima: dezoksiribonukleinske kiseline (DNK) i ribonukleinske kiseline (RNK). DNK se sastoje od dušičnih purinskih baza adenina (A) i gvanina (G), dušičnih pirimidinskih baza

citozina (C) i timina (T), šećera dezoksiriboze i fosforne kiseline. RNK su građene od istih blokova kao DNK, osim što je u njima pirimidinska baza timin zamijenjena uracilom (U), a umjesto šećera dezoksiriboze je riboza.

Danas je moguće izolirati, manipulirati pa čak i sintentizirati nukleinske kiseline korisnički definiranih sljedova i struktura što je dovelo do eksplozije generiranja podataka o ovim molekulama.

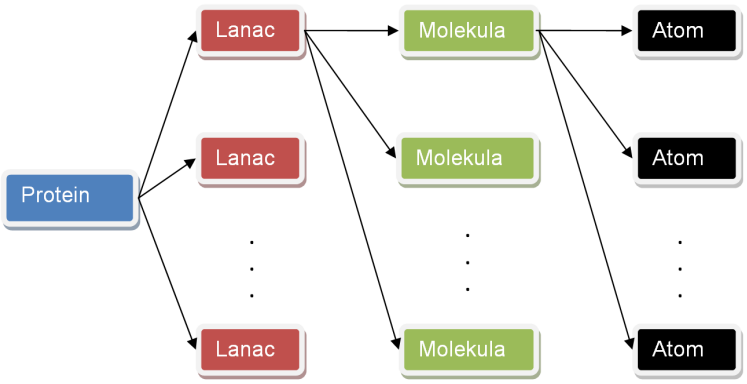
2.3. Uloga metala u biomolekulama

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 stekli detaljniji uvid u funkciju proteina te znali koliko su pojedini proteini specifični pri pojavi određenih metala. Trećina svih proteina treba metal kako bi mogli obavljati svoju funkciju tako da možemo susresti velik broj metala u proteinima.

Nukleinske kiseline se vežu s metalima na mnogo različitih načina kao što su kovalentna (koordinacija metala s DNK bazom, šećerom ili fosfatom) i nekovalentna vezanja (nekovalentno međusobno slaganje metalnih struktura) te vezanje s vodom. Vrlo bitna uloga metala u nukleinskim kiselinama je sposobnost djelovanja kao reducirajućeg agensa što pomaže u zaštiti organizma od štetnog djelovanja slobodnih radikala.

2.4. Struktura biomolekula

Većina biomolekularnih struktura u svojem sastavu sadrži lance različitih gradivnih podjedinica. Proteinski lanci su građeni od aminokiselina, no često u prirodi dolaze u obliku proteinskih kompleksa koji osim proteinskih lanaca mogu sadržavati RNK ili DNK lance, ligande, metale i molekule vode. Slika 2.1 prikazuje shematski prikaz strukture biomolekule.



Slika 2.1: Struktura biomolekule

2.4.1. Vrste lanaca

Lanci se sastoje od niza molekula i mogu biti različitih tipova. Razlikujemo 7 vrsta lanaca:

- voda
- metal
- DNK lanac
- RNK lanac
- proteinski lanac
- ostali

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 molekule vode (HOH).

Metal zapravo nije lanac, već samo jedan atom metala. Također ga je lako prepoznati jednostavnom usporedbom s popisom metala.

DNK i **RNK** predstavljaju nukleinske lance. Nukleinski lanci moraju zadovoljavati dva uvjeta: biti sastavljeni od najmanje 5 molekula i sadržavati samo i isključivo *DNK i RNK nukelotide* u sebi. Vrstu lanca određuje vrsta nukleotida od kojih se sastoji (DNK ili RNK).

Proteinski lanac je u sklopu ovog rada definiran kao takav ako se sastoji od najmanje 50 molekula i osnovnih 20 vrsta aminokiselina (ALA, ARG, ASN, ASP, CYS, GLU, GLN, GLY, HIS, ILE, LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR, VAL) mora činiti 95% tih molekula. Margina od 5% je dozvoljena jer se često pojavljuju neke neuobičajene aminokiseline (npr. MSE, ACE i slično).

Ukoliko lanac nije zadovoljio niti jedan od gore navedenih uvjeta, svrstan je u **ostale** lance jer nam u ovoj inačici nije važan.

3. Postojeća rješenja za analizu biomolekula

Prije izrade BioMe sustava istražena su postojeća rješenja za analizu biomolekula. U ovom poglavlju je dan kratak osvrt na pojedino rješenje te usporedba s BioMe sustavom.

U sklopu izrade web aplikacije MEPSUS (Hsin i Walkinshaw, 2008) provedeno je nekoliko istraživanja geometrije povezanosti metala u proteinima, ali na manjem skupu podataka i bez pokušaja odredeđivanja specifičnosti mjesta vezanja metala, kao što su distribucija atoma po koordinacijskim brojevima i kombinacija aminokiselina uključenih u koordinaciju. Aplikacija podržava skup od 10 metala te nudi mogućnost izrade statistika koje ukljucuju veze prema više molekula. Nedostaci su što ograničenje opsega pretraživanja na određeni protein ili lanac ne radi, vec se uvijek pretražuju svi te je aplikacija vrlo nestabilna u radu.

MIPS (Hemavathi i Sekar, 2009)(engl. *Metal Interactions in Protein Structures*) je još jedan slican projekt. Aplikacija se redovito održava te omogucuje detaljno specificiranje vrsta interakcija koje nas zanimaju. Nažalost, ne daje nikakve statistike povezanosti metala, vec samo ispisuje PDB kodove proteina u kojima se nalaze metali koji zadovoljavaju zadane kriterije.

Pokušaj klasifikacije metaloproteina i ostalih složenih proteina koristeći bioinorganske motive je istraživanje (Degtyarenko i Contrino, 2004) koje su proveli Degtyarenko i Contrino. Bioinorganski motivi su definirani kao zajednička strukturalna svojstva povezanih proteina koji uključuju metalne ione i koordinacije prve ljuske. Pripadna web aplikacija podržava pretraživanje malog broja struktura te nije dugo osvježavana.

MeRNA (Stefan i Holbrook, 2006) baza podataka je baza koja sadrži podatke o vezama metalnih iona s RNK te pruža informacije koje omogućuju klasifikaciju i pretraživanje određenih mjesta spajanja metala. Pruža informacije o dvadeset i tri različita iona metala. Pripadna web aplikacija je stabilna u radu, međutim informacije koje

pruža su ograničene na vrlo mali skup metalnih iona.

Svi spomenuti radovi, te ostali (Choi i Park, 2011), (Castagnetto i Pique, 2002), su također web sučelja koja nude neku vrstu analize, međutim analiza je svedena na unos paramatera i ispis popisa struktura koje zadovoljavaju unesene parametre. Ne postoji prava statistička analiza i podaci nisu osvježeni već dulje vrijeme.

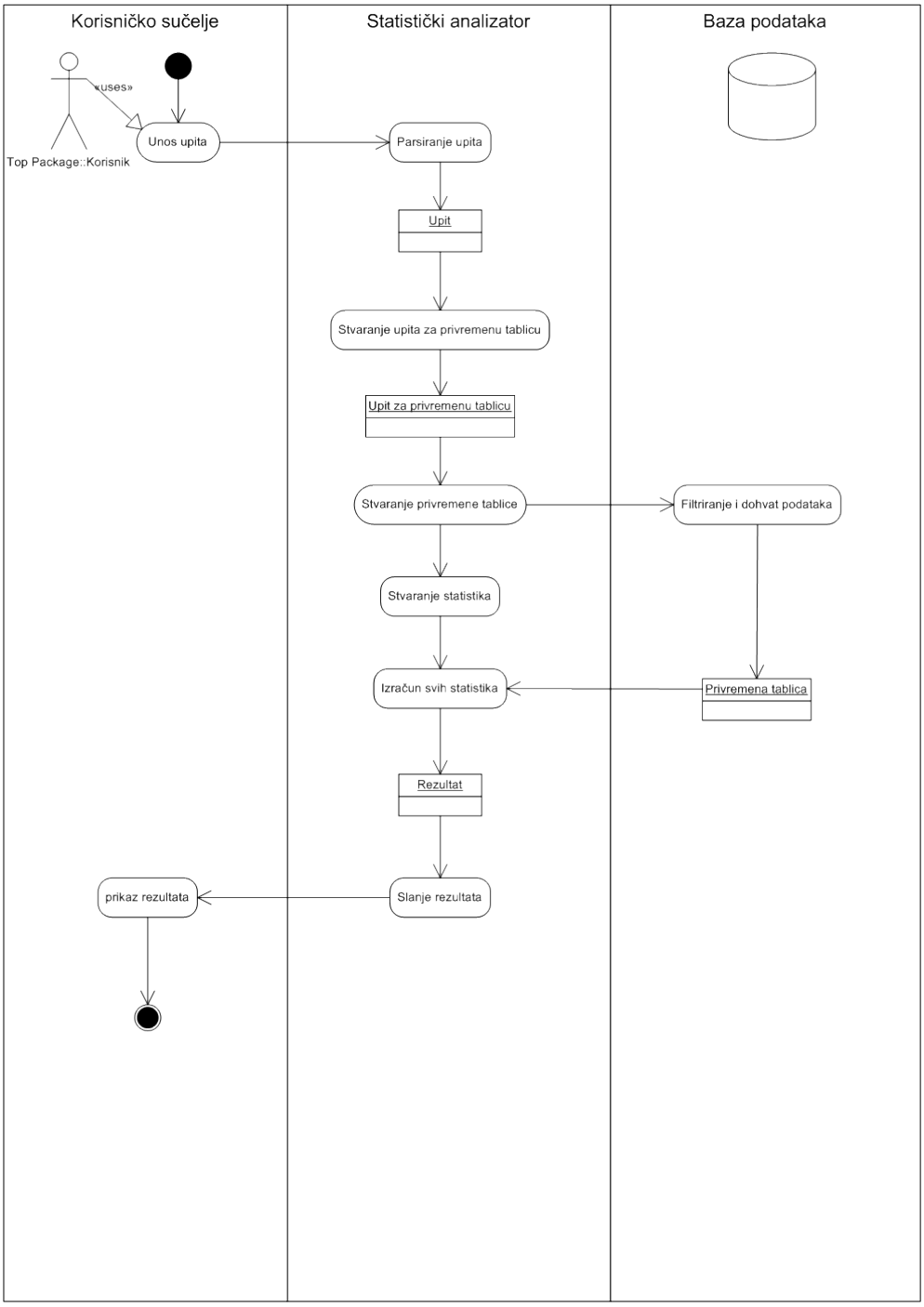
Osnovna prednost BioMe alata je velik broj statistika koje ostali alati nemaju. BioMe je jedini alat koji javno nudi cjelokupnu bazu podataka za preuzimanje. Time korisnici mogu samostalno provoditi vlastite statističke analize. Još jedna prednost jest da su ponuđene informacije uvijek najnovije. Pokušali smo stvoriti jedinstveni alat koji će popraviti sve uočene nedostatke, biti jednostavno proširiv i potpuno automatiziran.

4. Statistička analiza podataka

4.1. Arhitektura sustava za statističku analizu

Sustav se bazira na klijent-poslužitelj arhitekturi. Klijent je korisnik, odnosno njegov Internet preglednik, koji šalje zahtjev poslužitelju, BioMe sustavu, koji izračunava tražene podatke i šalje ih natrag.

Na slici 4.1 je prikazan dijagram aktivnosti statističke analize. Aktivnost započinje korisnik označavanjem željenih parametara na web sučelju. Pritiskom na gumb **submit** aktivira se statistički analizator koji prvo parsira označene podatke te potom generira upit koji se šalje bazi podataka. Iz baze podataka se filtriraju zapisi kako bi zadovoljili sva navedena ograničenja zadana na web sučelju. Navedeni zapisi se pohranjuju u privremenu tablicu u radnoj memoriji te služe za daljnje računanje statistika. Statistički analizator nastavlja s obradom rezultata te se oni šalju korisniku.



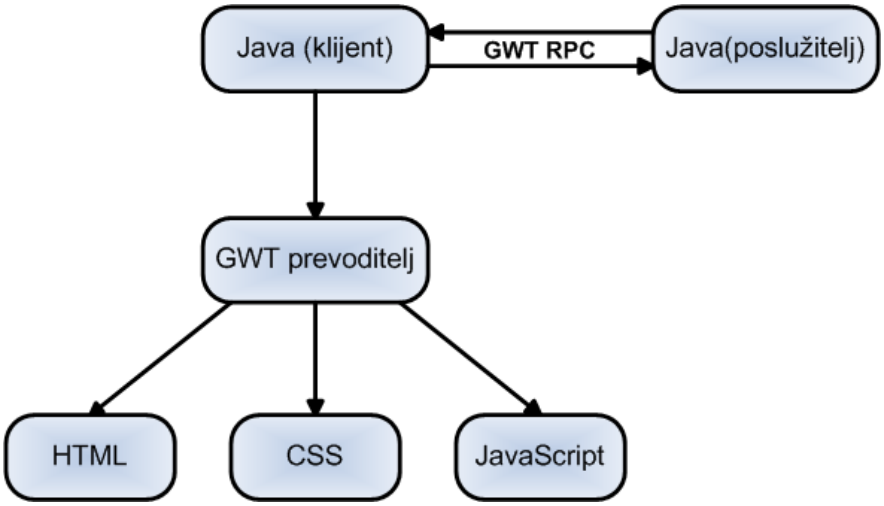
Slika 4.1: Dijagram aktivnosti za statističku analizu

4.1.1. Korištene tehnologije

Pozadinski sloj aplikacije (engl. *back-end*) je implementiran u *Javi*. Sučelje (engl. *front-end*) je implementirano pomoću GWT-a (engl. *Google Web Toolkit*) i *EXT-JS*.

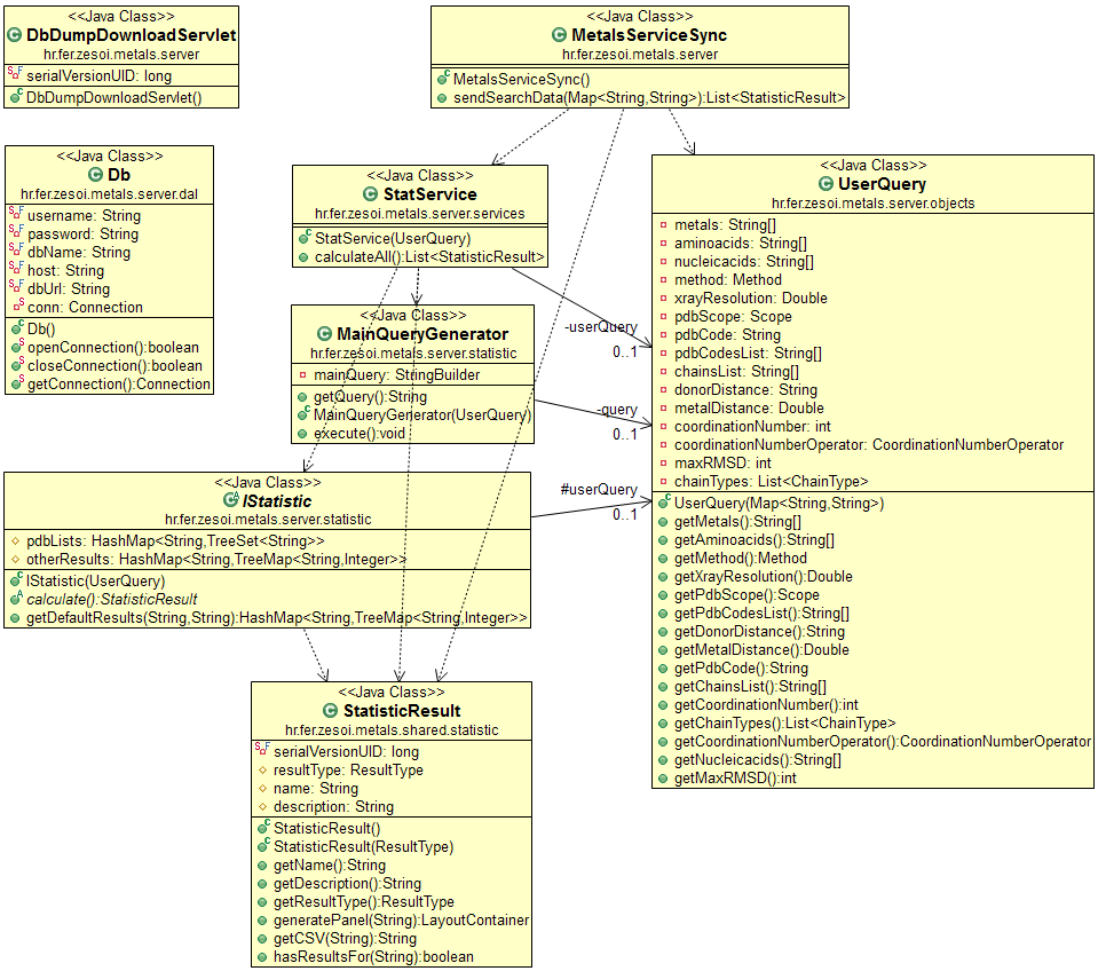
GWT alat za izradu web aplikacija temeljenih na programskom jeziku Java. Pred-

nost mu je prevođenje *Java* koda u *JavaScript* kod koji se izvršava u Internet pregledniku korisnika, točnije, HTML, CSS i JavaScript kod. Ovime je omogućen brz i jednostavan razvoj aplikacija u poznatom okruženju i poznatom programskom jeziku, bez potrebe za učenjem novog programskog jezika. GWT kompajler će generirati odgovarajući kod za prikaz (*HTML*, *CSS*) i funkcioniranje (*JavaScript*) web stranica kao što je prikazano na slici 4.2.



Slika 4.2: Prevođenje java koda pomoću *GWT*-a

EXT-JS je komercijalni dodatak za *GWT* koji je besplatan za korištenje ukoliko se radi o aplikaciji otvorenog koda. Pomoću ovog alata se dodatno pojednostavnjuje izrada aplikacija jer nudi niz gotovih komponenti koje je samo potrebno smjestiti na sučelje.



Slika 4.3: Dijagram razreda za serverski dio weba

Na slici 4.3 je prikazan dijagram razreda za serverski dio sustava za statističku analizu.

DbDumpDownloadServlet razred nasljeđuje razred *HttpServlet* te omogućava korisniku da preko Http request zahtjeva zatraži najnoviji dump BioMe baze podataka. Dump se vraća korisniku preko Http response zahtjeva. Navedena funkcionalnost je realizirana klikom na link u prozoru Database dumps.

Db razred je realiziran pomoću Singleton oblikovnog obrasca te sadržava metode za otvaranje i zatvaranje veze prema bazi podataka, te atribute za pristup.

MetalServiceSync razred koji centralno prima zahtjeve od klijenata i na njih odgovara.

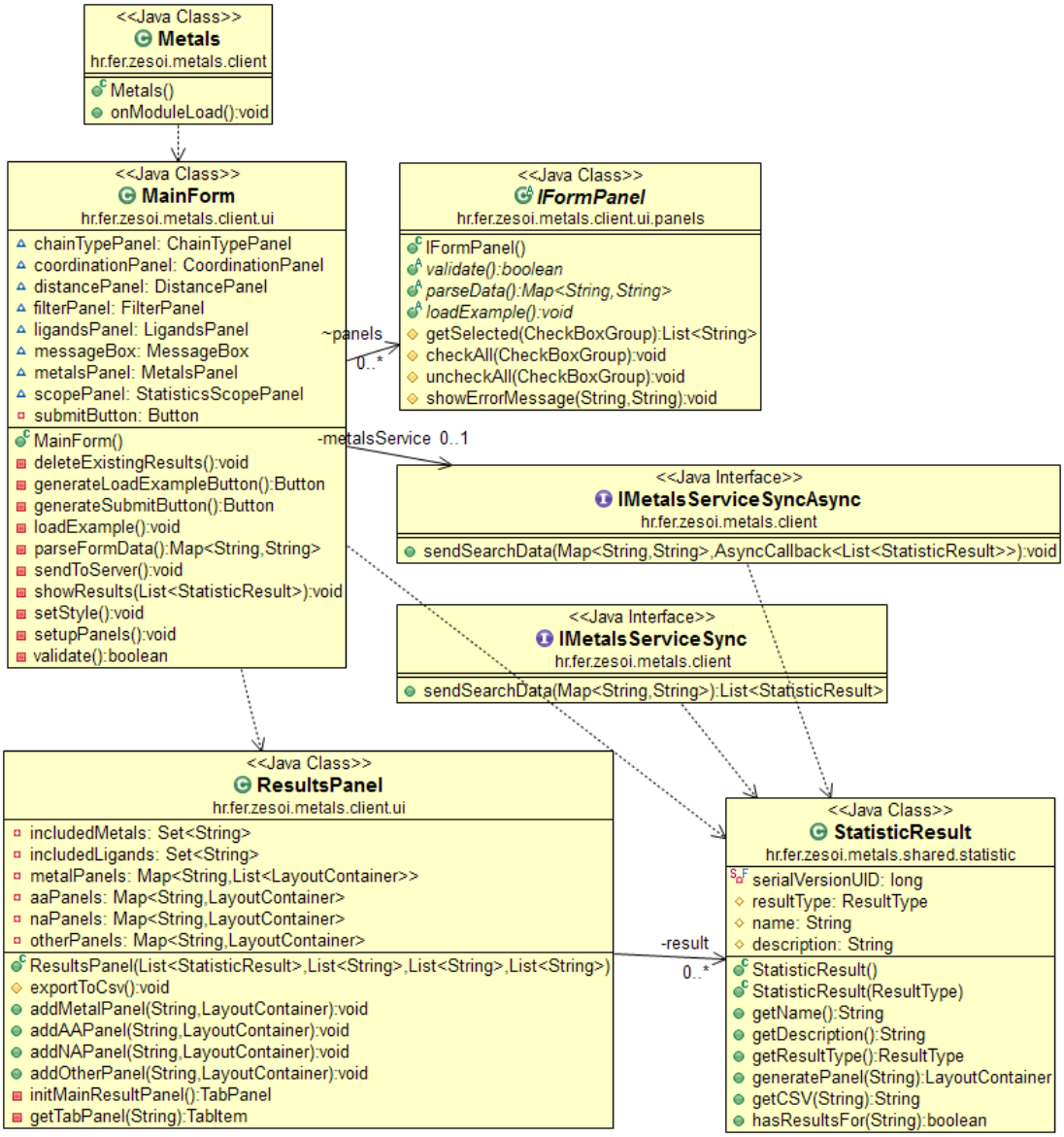
StatService razred predstavlja servis za stvaranje svih statistika te za svaku statistiku poziva metodu za izračunavanje rezultata.

MainQueryGenerator razred služi kako bi se generirao SQL upit koji će iz korisnikovog upita stvoriti privremenu tablicu koja se koristi u izračunavanju svih statistika. Razred sadrži metode za formatiranje upita po komponentama web sučelja.

UserQuery razred služi za pohranu svih podataka označenih na web sučelju.

IStatistic sučelje implementira svaki pojedini razred koji predstavlja statistiku. Svaki od ovih razreda ima svoj SQL upit za dohvat podataka potrebnih za izračun. Sučelje sadrži metode *calculate();* i *getDefaultResults();* koje služe za izračun rezultata.

StatisticResult razred je osnovni razred za pohranu rezultata. Svi razredi u kojima je pohranjen rezultat nasljeđuju ovaj razred. On sadržava metode za dohvat rezultata, generiranje strance za grafički prikaz rezultata, generiranje CSV prikaza rezultata te druge. Ovaj razred je zajednički serverskom i klijentskom dijelu sustava.



Slika 4.4: Dijagram razreda za klijentski dio weba

Slika 4.4 prikazuje dijagram razreda za klijentski dio sustava za statističku analizu.

Metals je ulazna točka klijentskog dijela sustava. Pri prvom pozivu sučelja se poziva ova klasa. Ona inicijalizira ostatak sučelja.

IFormPanel sučelje sadrži deklaracije metoda za parsiranje korisnikovog upita s web sučelja, validaciju unesenih podataka na web sučelje te metoda za ponašanje gumbi na panelu.

MainForm je razred koji implementira sučelje *IFormPanel*. Sadrži varijable koje predstavljaju panele od kojih je sastavljeno web sučelje.

IMetalsServiceSync sučelje za odašiljanje upita poslužitelju.

IMetalsServiceSyncAsync isto sučelje kao i prethodno, ali funkcionira asinkrono.

ResultPanel razred služi za prikaz rezultata. Sadrži metode za dodavanje liste stranica rezultata.

4.1.2. Performase

Sustav za statističku analizu je prošao kroz značajne promjene tokom svog razvoja. U prvotnoj verziji (Peretin, 2010) sustava korisnik je mogao odabrati manji broj metala te se tada računao i manji broj statistika. U ovoj nadogradnji implementirana je statistika koordinacijskih geometrija, a također su omogućen odabira nukleinskih kiselina. Uz to, vrijeme računanja starog sustava (pri najvećem opterećenju) je dosezalo nekoliko sati. Također, sustav je bio nestabilan te za pojedine upite sustav nije davao ispravne rezultate.

Način računanja rezultata je imao sljedeće korake:

- filtracija podataka iz baze s obzirom na korisnikov upit na web sučelju, rezultat ovog koraka je bila lista atoma koji zadovoljavaju upit (njih nekoliko tisuća pri najvećem opterećenju)
- za dio statistika, za svaki atom iz prvog koraka se izvršavao SQL upit u kojem se spajalo nekoliko tablica iz baze u svrhu dobivanja rezultata
- za ostale statistike se izvršavao po jedan SQL upit u bazu podataka

Ogroman broj SQL upita je postajao sve veći problem kako se BioMe baza nadopunjavala novim podacima, a praktična korist od aplikacije s dugačkim trajnjem postupka analize je nikakva.

Prva faza optimizacije je imala cilj smanjiti broj SQL upita na konačan broj, tj. u najboljem slučaju smanjiti broj SQL upita na broj statistika koje se računaju. Zadatak je uspješno izvršen, te se trenutno većina statistika računa s jednim SQL upitom, dok tri statistike imaju upita koliko ima označenih metala (maksimalno 25). Sveukupno vrijeme računanja se ovim postupkom smanjilo na 15-ak minuta.

Cilj druge faze optimizacije je bio smanjiti trajanje SQL upita dobivenih prvom fazom. Ovaj problem je uspješno riješen učitavanjem filtriranih podatka (onih koji zadovoljavaju korisnikov upit) u privremenu tablicu u radnoj memoriji računala. Pri najvećem opterećenju (označeni svi metali, ligandi i lanci, te područje pretrage postavljeno na sve PDB zapise) ova tablica zauzima malo više od 300 MB radne memorije.

Izračuni statistika se obavljaju čitanjem podataka iz privremene tablice te im je ovim korakom vrijeme računanja smanjeno na svega nekoliko sekundi. Sveukupno trajanje statističke analize trenutno je ispod jedne minute što je ogroman napredak u odnosu na prvotnu verziju.

4.2. Sučelje aplikacije

Korisničko sučelje aplikacije sastoji se od forme za odabir parametara koji služe da bi se odredilo kako se koja statistika računa te od forme za prikaz rezultata. Rezultati su prikazani u obliku liste stranica pri čemu svaka stranica ima rezultate za pojedini metal ili je na stranici prikaz izračunavanja statistike po lignadima. Rezultati su prikazani numerički i grafički pomoću tortnih te trakastih grafikona.

Forma za prikaza rezultata sastoji se od velikog broja opcija koje treba odabrati kako bi se oblikovao željeni upit. Ponuđene opcija su odabir željenog iona metala, način odabira strukture (kristalografija pomoću X-zraka, NMR ili obje) i rezolucije. Također, moguć je i odabir tipa lanca i kombinacije linganda kao i koordinacijski broj, maksimalna pogreška koordinantne geometrije te udaljenost između označenih iona metala. Forma sadržava i stranicu za pomoć gdje su objašnjeni svi navedeni parametri.

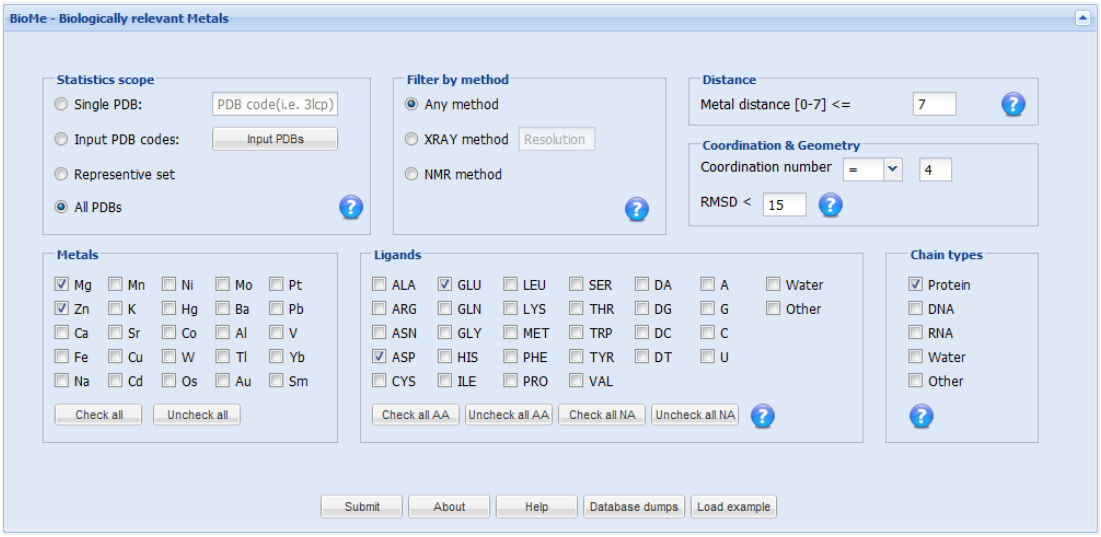
Pretraživanje je moguće izvršiti na četiri različita skupa PDB zapisa. Najjednostavnija pretraga podrazumijeva odabir samo jednog PDB zapisa, zatim se može odabrati skup od nekoliko PDB zapisa te cjelokupna lista PDB zapisa koji zadovoljavaju svojstvo da njihovi ioni metala imaju barem dva atoma donora iz lanca proteina ili jedan iz lanca nukleinske kiseline. Također, dostupno je i pretraživanje po reprezentativnom skupu naziva *cluster 70* skup.

Korisnik ima mogućnost izabrati do 25 najzastupljenijih metala (Mg, Zn, Ca, Fe, Na, Mn, K, Sr, Cu, Cd, Ni, Hg, Co, W, Os, Mo, Ba, Al, Tl, Au, Pt, Pb, V, Yb, Sm) za izračun statistika. Uz to, korisnicima je omogućen pristup dump-u podataka gdje su zastupljeni svi metali iz baze proteina.

Korisničko sučelje pruža odabir pet različitih vrsta liganda pa je moguće odabrati između aminokiselina, DNK i RNK nukleotida, vode te ostalih. Budući da je velik broj ostalih liganda, a njihova zastupljenost u lancima vrlo mala, oni su u prikazu statistika stavljeni u zajedničku skupinu. Međutim, kako u bazi podataka tako i u listi rezultata je omogućena opcija pregleda svakog pojedinog liganda.

4.3. Statistike

U ovom poglavlju bit će opisan način izračunavanja rezultata pojedinom statistikom te je dan primjer izračuna rezultata za svaku statistiku. Prikazani rezultati su dobiveni za odabrani metal Mg, te aminokiseline ASP i GLU za koordinacijski broj 4. Pretražuje se skup svih PDB-ova, a filtracija je postavljena na bilo koju metodu. Prag za grešku pri računanju koordinacijske geometrije je postavljen na 15. Na slici 4.5 je prikazana forma za unos navedenih paramatera s postavljenim vrijednostima kako je prethodno opisano.

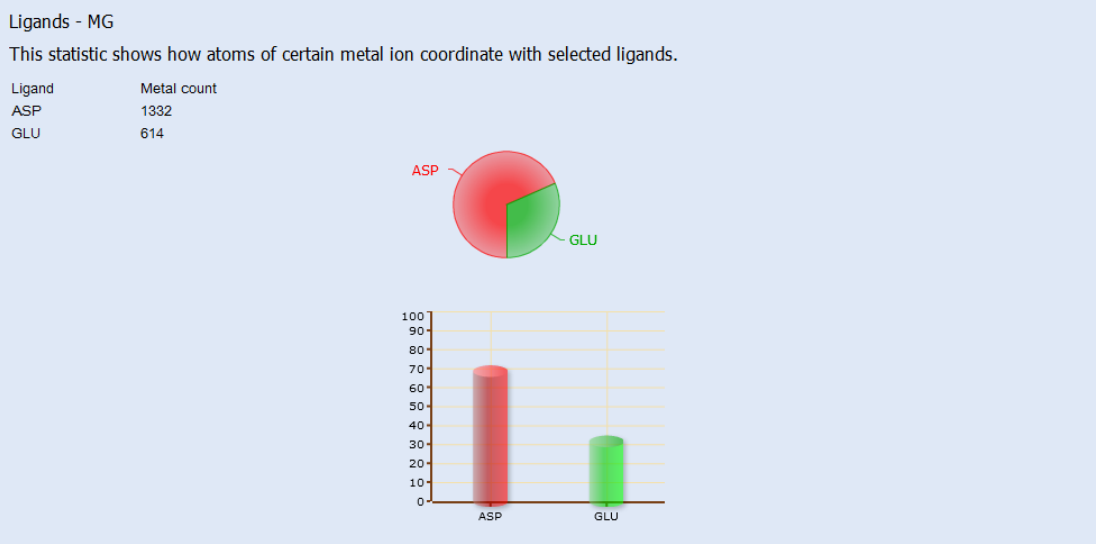


Slika 4.5: Prikaz forme za odabir parametara s označenim vrijednostima za pokazni primjer

Razlikujemo dvije vrste statistika: statistike koje se računaju za označeni metal (*M1 - M7*) te statistika koja se računa za označene ligande (*L1*). U nastavku su navedene i pobliže objašnjene sve statistike.

4.3.1. Distribucija veza metala s odabranim ligandima - M1

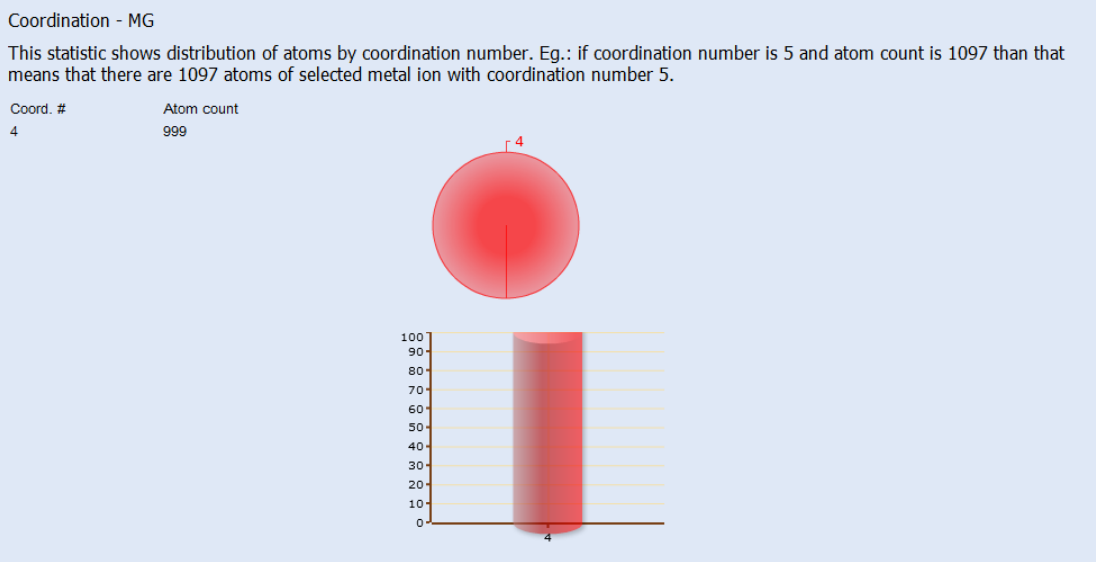
Ova statistika pokazuje kako se atomi pojedinog iona metala vežu na odabrane ligande. Metal je u vezi s ligandom ako je barem jedan atom iz tog liganda udaljen manje od 3Åod bilo kojeg atoma metala. Ukupni broj svih veza metala i odabranog liganda je zbroj veza svih atoma tog metala s ligandom.



Slika 4.6: Rezultat izračunavanja statistike M1

4.3.2. Distribucija broja atoma metala po koordinacijskom broju - M2

Ova statistika pokazuje kako su raspoređeni atomi metala po koordinacijskom broju, tj. računa se broj atoma metala za određeni koordinacijski broj.



Slika 4.7: Rezultat izračunavanja statistike M2

4.3.3. Kombinacije liganda po koordinacijskom broju - M3

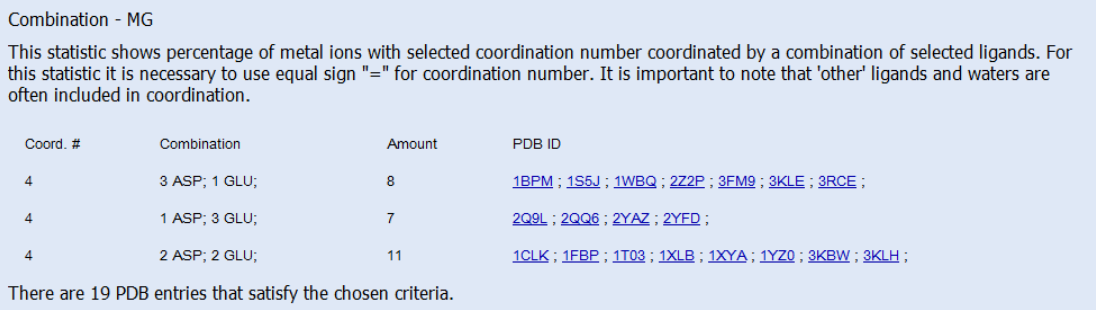
Ova statistika pokazuje kombinacije liganda s kojima metali međudjeluju. Za prikaz rezultata ove statistike potrebno je odabrati samo jedan koordinacijski broj (odabirom znaka jednakosti u korisničkom sučelju). Ukoliko postoji kombinacija svih označenih liganda s pojedinim atom označenog metala, izračunava se broj takvih kombinacija. Statistika također prikazuje i popis PDB-ova po nađenim kombinacijama. Budući da je ovo algoritamski najsloženija statistika, u nastavku je dan Pseudokod 1.

```
begin
  ListaAtomUIDs = dohvatiAtomUIDIzPrivremeneTablice();
  for (atomUID : AtomUIDs)
     MapaAtomUIDLigand = dohvatiSveLigande(atomUID);
  end
  AtomUIDLigand = filrirajPoKoordinacijskomBroju(AtomUIDLigand);
  MapaComboAtomUID = stvoriComboAtomUID(AtomUIDLigand);
  MapaRezultat = prebrojiComboAtomUID(ComboAtomUID);
end
```

Pseudokod 1: Pseudokod statistike M3

Izračun započinje dohvaćanjem svih atomUID-ova koji zadovoljavaju korisnikov upit. U složenu strukturu podataka (*HashMap<Integer, ArrayList<String>>*) se za svaki dohvaćeni atomUID sprema lista liganda kojoj taj atom pripada. U slučaju da je tip liganda "other" tada se u posebnu listu dohvaćaju i svi "other" ligandi. Metoda *filrirajPoKoordinacijskomBroju(atomUIDLigand);* prolazi kroz zapise u mapi *atomUIDLigand* te iz nje izbacuje one zapise kod kojih je broj liganda različit od koordinacijskog broja definiranog korisnikovim upitom (na pr. za koordinacijski broj 4 u mapi će ostati zapis {1234:{HIS, ASP, HIS, ASP}} gdje je 1234 atomUID, a {HIS, ASP, HIS, ASP} lista liganda). Isto tako, iz mape se izbacuju i oni zapisi u kojima nisu navedeni svi ligandi definirani korisničkim upitom. Metoda *stvoriComboAtomUID(atomUIDLigand);* vraća *HashMap<String, ArrayList<Integer>>* gdje ključ predstavlja vrijednost kombinacije svih označenih liganda (u gornjem primjeru bi to bilo: "2 HIS; 2 ASP"), a vrijednost je lista svih atomUID kod kojih je pronađena navedena kombinacija liganda. Ovaj postupak se ponavlja za svaki označeni metal. Rezultat je prikazan *HashMap<String, HashMap<Integer, HashMap<String, Integer>>>* strukturom podataka gdje je ključ vanjske mape metal za koji se računa statistika, a

vrijednost te mape je mapa u kojoj je ključ koordinacijski broj, a njezina vrijednost je mapa u kojoj je ključ combo zapis liganda, a vrijednost je broj takvih combo zapisa (shematski prikaz je: {metal:{coordination:{combo:cnt}}}).



Combination - MG

This statistic shows percentage of metal ions with selected coordination number coordinated by a combination of selected ligands. For this statistic it is necessary to use equal sign "=" for coordination number. It is important to note that 'other' ligands and waters are often included in coordination.

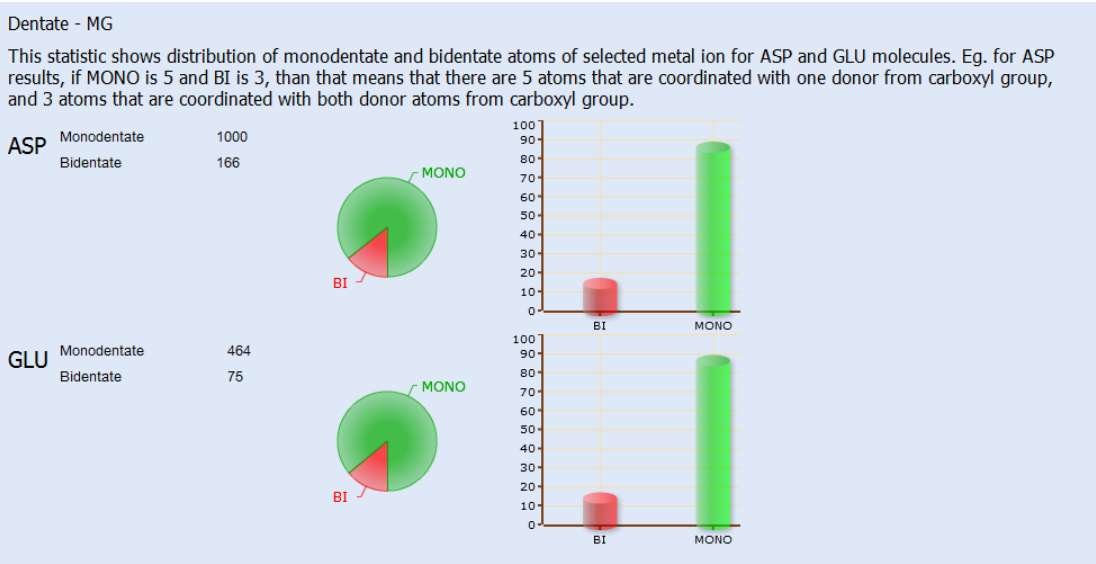
Coord. #	Combination	Amount	PDB ID
4	3 ASP; 1 GLU;	8	1BPM ; 1S5J ; 1WBQ ; 2Z2P ; 3FM9 ; 3KLE ; 3RCE ;
4	1 ASP; 3 GLU;	7	2Q9L ; 2QQ6 ; 2YAZ ; 2YFD ;
4	2 ASP; 2 GLU;	11	1CLK ; 1FBP ; 1T03 ; 1XLB ; 1XYA ; 1YZ0 ; 3KBW ; 3KLH ;

There are 19 PDB entries that satisfy the chosen criteria.

Slika 4.8: Rezultat izračunavanja statistike M3

4.3.4. Distribucija monodentatno i bidentatno koordiniranih metala s ASP i GLU aminokiselinama - M4

Ova statistika pokazuje broj monodentatno i bidentantno koordiniranih metala. Monodentatno koordiniran metal je onaj koji je koordiniran samo s jednim atomom kisika iz karboksilne skupine, dok je bidentatno koordiniran onaj koji je koordiniran s dva atoma kisika. Statistika se računa samo kada su označene ASP ili GLU ili obje aminokiseline. Atomi kisika koji se razmatraju su OE1, OE2 i O kod GLU aminokiseline, te OD1, OD2 i O kod ASP aminokiseline.



Slika 4.9: Rezultat izračunavanja statistike M4

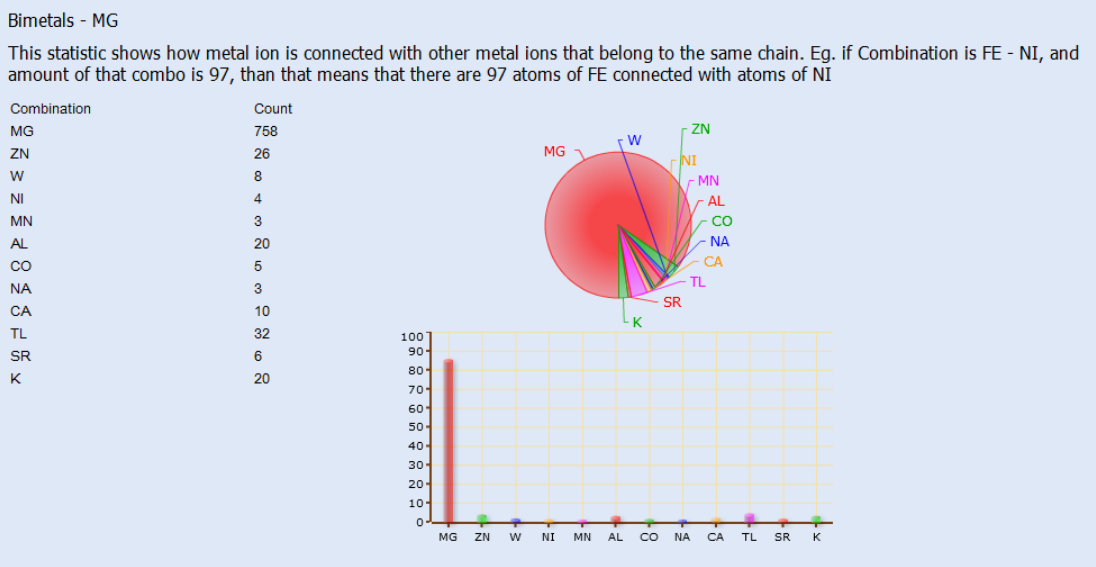
Pseudokod 2 opisuje kako se izračunava ova statistika. Računanje započinje dohvatom svih atomUID-ova iz privremene tablice koji zadovoljavaju korisnikov upit. Tada se za svaki atomUID prebrojavaju sve veze koje taj atom ostvaruje s aminokiselinom. Metoda *PrebrojiSveRazliciteVezeMetalAA();* jednim SQL upitom prebrojava sve različite veze metala i aminokiseline bez obzira na atom, tj. uključujući sve atome. Uz razumnu pretpostavku da nikada nema tri atoma pojedine aminokiseline koja se veže na metal tada je broj bidentatnih veza jednak razlici tih dviju vrijednosti, dok je broj monodentatnih veza jedan razlici ukupnog broja veza metala i aminokiseline i dvostrukog broja bidentatnih veza.

```
begin
  ListaAtomUIDs = dohvatiAtomUIDIzPrivremeneTablice();
  for (atomUID : AtomUIDs)
    ukupanBrojVezaMetalaIAA+ = PrebrojiSveVezeMetalAA(atomUID);
    ukupanBrojRazlicitihVezaMetalaIAA = PrebrojiSveRazliciteVezeMetalAA();
    brojBidentatnih = ukupanBrojVezaMetalaIAA − ukupanBrojRazlicitihVezaMetalaIA
    brojMonodentatnih = ukupanBrojVezaMetalaIAA − 2 ∗ brojBidentatnih;
  end
```

Pseudokod 2: Pseudokod statistike M4

4.3.5. Distribucija atom metala vezanih za isti lanac - M5

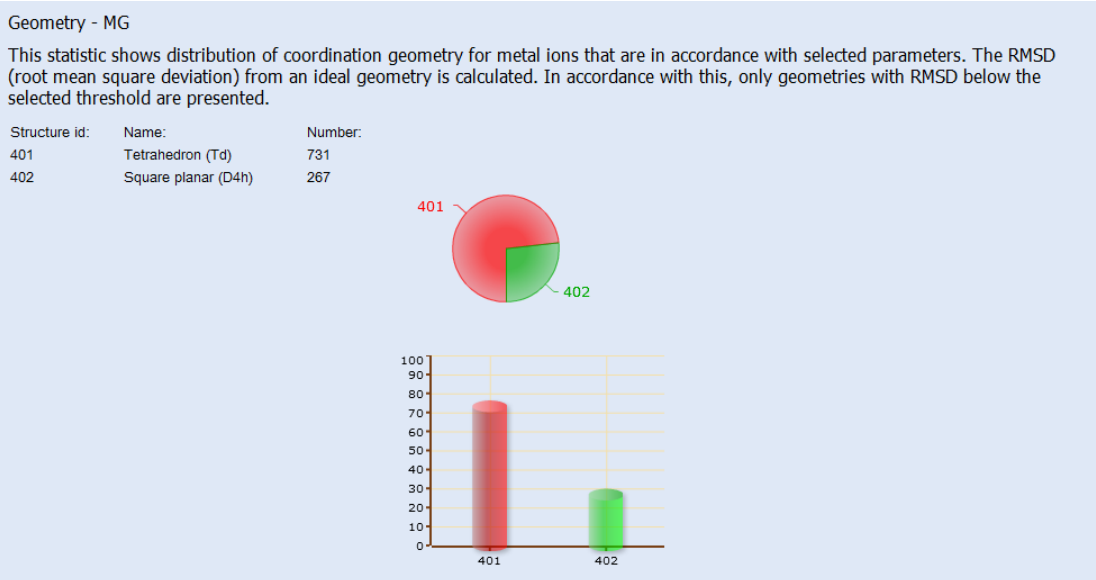
Ova statistika pokazuje broj atoma metala povezanih s odabranim metalom koji se nalaze na istom lancu. Atomi su povezani ako su na istom lancu te su udaljeni manje od željene udaljenosti u Åod bilo kojeg atoma tog metala. Željena udaljenost se može odabrati u korisničkom sučelju, a prepostavljena vrijednost je 7Å.



Slika 4.10: Rezultat izračunavanja statistike M5

4.3.6. Distribucija koordinacijske geometrije po metalima - M6

Ova statistika pokazuje broj atom metala u zadanoj geometrijskoj strukturi s obzirom na parametre zadane preko korisničkog sučelja. Vrijednost RMSD parametra predstavlja korijen srednje vrijednosti devijacije od idealne geometrijske strukture te se u rezultatu pribrajaju oni rezultati koji su ispod zadane granice.



Slika 4.11: Rezultat izračunavanja statistike M6

Analiza kutova

Kutovi su definirani tako da se metal nalazi u vrhu kuta, a donori čine krakove. Pri pokretanju analize, predaje se koordinacijska mapa, a rezultat je *mapa kutova*. Potrebno je izračunati sve vrijednosti kutova za sve permutacije parova njegovih donora bez ponavljanja.

Geometrijska analiza

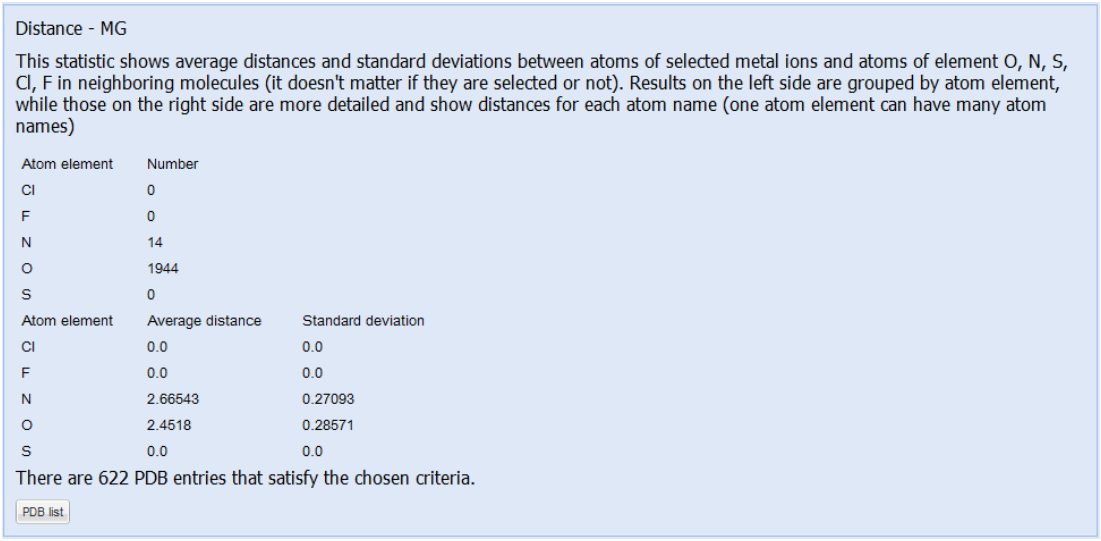
Geometrijska analiza na ulaz prima mapu kutova iz prethodne analize. Postupak određivanja geometrijske strukture je dan u Pseudokodu 3.

```
begin
  dohvatiMapuKutova();
  for (svaki_metal_i_njegove_kutove)
     poredajKutovePoVelicini();
     for (struktura : ListaStruktura)
        if (koordinacijsiBrojMetala == koordinacijskiBrojStrukture)
          poredajKutoveStrukturePoVelicini();
          izracunajRMSD();
          mapaKandidata = pohraniOdabranuGeometrijskuStrukturuIRMSD();
     else
          continue;
        end
        izaberiStrukturuSNajmanjimRMSD();
        spremiPromjene();
     end
  end
```

Pseudokod 3: Pseudokod računanja koordinacijskih geometrija

4.3.7. Srednja udaljenost i standardna devijacija po određenim elementima - M7

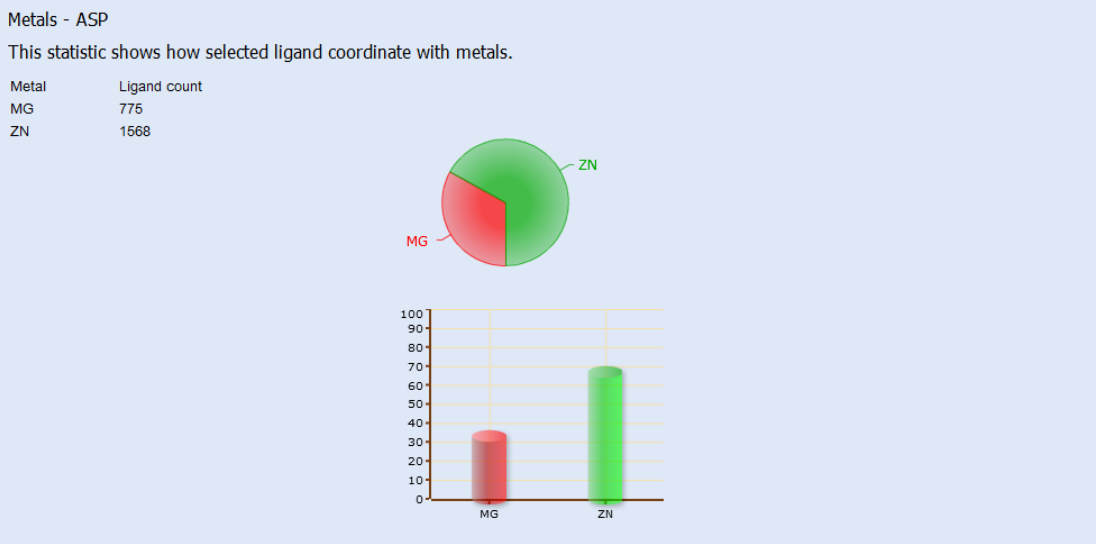
Ova statistika pokazuje broj atoma Cl, O, S, F i N povezanih s atomima odabranog metala. Uz to, izračunavaju se srednja udaljenost i standardna devijacija te udaljenosti od navedenih atoma do atoma odabranog metala.



Slika 4.12: Rezultat izračunavanja statistike M7

4.3.8. Distribucija atoma metala po ligandima - L1

Ova statistika pokazuje broj atoma odabranog metala koji su u međudjelovanju sa svakim pojedinim označenim ligandom. Metal je u vezi s ligandom ako je barem jedan atom iz tog liganda udaljen manje od 3Åod bilo kojeg atoma metala. Ukupni broj svih veza metala i odabranog liganda je zbroj veza svih atoma tog metala s ligandom.



Slika 4.13: Rezultat izračunavanja statistike L1

5. Podaci i rezultati

5.1. Podaci

U tablici 5.1 je dan broj gradivnih jedinica molekula. Od gradivnih struktura ističe se velik broj proteinskih, dok je nukleinskih struktra puno manje s obzirom da su nukleinske kiseline puno slabije istraženo područje.

Tablica 5.1: Struktura podataka u bazi

Ukupno struktura	21936
proteinskih struktura	21498
nukleinskih struktura	1137
proteinske i nukleinske	985
Lanaca	265563
Liganada	634460
Atoma	758595
metala	181006

BioMe baza podataka sadrži podatke o svim donorima koji su od metalnih iona udaljeni manje ili jednako 3Å. Tablica 5.2 sadrži broj takvih donora te isto tako i broj onih donora iz ulaznog skupa koje nisu zadovoljile navedeno ograničenje.

Tablica 5.2: Udaljenosti struktura od metalnih iona

Udaljenost	Broj udaljenosti
<= 3 A	558892
= 3	198
>3	1173491

Vrste zastupljenih lanaca dane su u tablici 5.3. Najviše je lanaca sastavljeno od metala, te je također velik broj proteinskih lanaca. Budući da je velik broj, a njihova

zastupljenost mala, uvedena je oznaka “other” koja označava lance koji ne pripadaju niti jednoj od ostalih navedenih kategorija.

Tablica 5.3: Vrste lanaca

Vrsta lanca	Broj lanaca
METAL	159837
PROTEIN	46326
OSTALI	30349
VODA	27133
DNK	1050
RNK	868

Budući da BioMe baza podataka sadrži više od dvije tisuće različitih liganada, u tablici 5.4 je dan pregled 15 najzastupljenijih. Esencijalna aminokiselina histidin koja je važna za mentalno i fizičko zdravlje čovjeka je prva na popisu.

Tablica 5.4: Top 15 liganada

Skraćeno	Broj	Puni naziv
HIS	51996	HISTIDIN
ASP	49531	ASPARTNA KISELINA
CYS	36259	CISTEIN
GLU	32642	GLUTAMINSKA KISELINA
G	19725	GUANOZIN-5’-MONOFOSFAT
A	13896	ADENOZIN-5’-MONOFOSFAT
ASN	11272	ASPARAGIN
U	9888	URIDIN-5’-MONOFOSFAT
C	8836	CITIDIN-5’-MONOFOSFAT
THR	6865	TREONIN
GLY	6840	GLICIN
SER	6309	SERIN
GLN	4668	GLUTAMIN
TYR	4119	TIROZIN
VAL	4035	VALIN

Najzastupljeniji metal je magnezij, a najzastupljeniji nemetal je kisik. S obzirom da je korisniku omogućen izbor između 25 metala za koje može dobiti po sedam statis-

tika, u tablici 5.5 je dana raspodjela po našem mišljenju najvažnijih metala i nemetala. Statistika M7 računa udaljenost i standardnu devijaciju svakog iona metala do ovdje navedenih nemetala.

Tablica 5.5: Razdioba metala i nemetala

Metal	Puni naziv	Broj atoma
MG	Magnezij	85714
ZN	Cink	19443
CA	Kalcij	18086
FE	Željezo	14882
NA	Natrij	12442
MN	Mangan	6015
K	Kalij	3769
CU	Bakar	3480
SR	Stroncij	3301
OS	Osmij	3118
CD	Kadmij	3014
NI	Nikal	1417
HG	Živa	1300
CO	Kobalt	1259
W	Volfram	1105

Nemetal	Puni naziv	Broj atoma
O	Kisik	421797
N	Dušik	108063
S	Sumpor	45965
F	Fluor	1073
CL	Klor	691

Koordinacijske geometrije su geometrijske tvorevine atoma oko središnjeg atoma. BioMe razlikuje 20 različitih geometrijskih struktura (tablica 5.6), od kojih su octahedron i tetrahedron najzastupljenije.

Tablica 5.6: Koordinacijske geometrije

Ime strukture	Koordinacijski broj	Broj struktura
Oktaedar (Oh)	6	30911
Tetraedar (Td)	4	24160
Trostrana bipiramida (D3h)	5	15775
Trostrana prizma	6	7836
Četverostrana planarna (D4h)	4	4893
Trostrana prizma (C2v)	7	4819
Četverostrana piramida (C4v)	5	3918
Oktaedar(C3v)	7	2165
Peterokutna bipiramida(D5h)	7	1258
Kvadratna antiprizma (D4d)	8	921
Trostrana prizma, bipoklopljena	8	574
Kvadratna antiprizma, bipoklopljena (D4d)	10	203
Dodekaedar (D2d)	8	171
Kvadratna antiprizma, bipoklopljena	9	166
Heksagonalna bipiramida (D6h)	8	165
Trostrana prizma, trokutasto poklopljena	8	34
Antikockahedron	12	26
Kocka (Oh)	8	18
Heksagonalna antiprisma	14	8
Trostrana prizma, tripoklopljena	9	5

Tablica 5.7 prikazuje raspodjelu geometrijski struktura po koordinacijskim brojevima, tj. koliko atoma čini strukturu.

Tablica 5.7: Geometrije po koordinacijskom broju

Koordinacijski broj	Broj struktura
6	38747
4	29053
5	19693
7	8242
8	1883
10	203
9	171
12	26
14	8

5.2. Tijek razvoja i rezultati rada

U poglavlju 4.1.2 opisana su dva dijela optimizacije statističke analize. U prvom dijelu problem je predstavljala količina memorije koju je statistički analizator zauzimao. Pri najvećem opterećenju je prelazila nekoliko gigabajta i uzrokovala rušenje sustava. Korištenjem programa za profiliranje (Ej-technologies, 2012) i prelaskom na *Tomcat* (Foundation, 2012) web poslužitelj taj problem je riješen. Problem je bio uzrokovan ograničenjima u implementaciji *Glassfish* (Oracle, 2012) poslužitelja spojen s činjenicom da *GWT* ne oslobađa zauzete resurse odmah po završetku izvođenja.

U drugom dijelu je korištenje radne memorije umanjeno tri puta, ali se pri najvećem opterećenju nisu prikazivali rezultati. Otkrivene su i uklonjene dvije greške u kodu analizatora. Prva pogreška je bila uzrokovana nepravilnom inicijalizacijom strukture podataka za rezultate jedne od statistika. Zbog toga se pojavljivala iznimka null pokazivača. Druga pogreška je bila u serijalizaciji objekata. Svakom objektu koji se može serijalizirati je dodijeljen jedinstveni broj koji se mora osvježiti pri svakoj promjeni podatkovne strukture tog objekta. Oba problema su uspješno riješena.

S početnih petnaest minuta koliko je trajao dohvat najkompliciranijeg rezultata na prvoj inačici statističkog analizatora, postignuto je da najkompliciraniji upit traje šezdeset sekundi. Valja napomenuti da su u međuvremenu dodane nove statistike.

Navedena vremena trajanja su izmjerena na računalu s Gentoo Linux operacijskim sustavom (Intel Core2 Quad CPU Q6600 at 2.40GHz, 4 GB).

6. Zaključak

Web sučelje opisano u ovom radu je dio BioMe sustava za statičku analizu podataka o biomolekulama. Proteini i nukleinske kiseline su biomolekule koje su ključne u gotovo svim biologijskim i kemijskim procesima. Interakcija i međusobni položaj biomolekula i metala određuju njihovu ulogu i značaj. Ovo područje je jedno od žarišta istraživanja zbog svoje važnosti za životne procese, kako čovjeka tako i ostalih živih bića.

Web sučelje statističkog analizatora omogućuje odabir između 5 vrsta lanaca, dvadeset i pet metala, dvadeset osnovih aminokiselina te nukleinskih kiselina te mnoge druge. Odabirom parametara započinje statistička analiza koja za rezultat ima 8 statistika kao što su statistički podaci o povezanosti iona metala s amino i nukleinskim kiselinama, njihove distribucije po koordinacijskim brojevima te koordinacijske geometrije.

Kroz svoj razvoj sustav je prošao kroz brojne promjene. U trenutnoj verziji sustava postignuta je velika optimizacija na području vremena izvođenja kao i na iskoristivosti radne memorije.

U usporedbi s ostalim dostupnim sličnim alatima BioMe je superioran po broju statistika, vremenu izvođenja, količini podataka koja se obrađuje te po automatiziranosti cjelokupnog procesa te automatskom osvježavanju podataka za statističku analizu.

LITERATURA

Hennessy S.W. Roberts V. a Getzoff E.D. Tainer J. a Castagnetto, J.M. i M.E. Pique. Mdb: the metalloprotein database and browser at the scripps research institute. *Nucleic acids research*, 2002.

Kang H. Choi, H. i H. Park. Metligdb: a web-based database for the identification of chemical groups to design metalloprotein inhibitors. *Journal of Applied Crystallography*, 2011.

K. Degtyarenko i S. Contrino. Come: the ontology of bioinorganic proteins. *BMC structural biology*, 2004.

Ej-technologies. Jprofiler, 2012. URL `http://www.ej-technologies.com/index.html`.

The Apache Software Foundation. Apache tomcat, 2012. URL `http://tomcat.apache.org/`.

Kalaivani M. Udayakumar a. Sowmiya G. Jeyakanthan J. Hemavathi, K. i K. Sekar. Mips: metal interactions in protein structures. *Journal of Applied Crystallography*, 2009.

Sheng Y. Harding M.M. Taylor P. Hsin, K. i M.D. Walkinshaw. Mespeus: a database of the geometry of metal sites in proteins. *Journal of Applied Crystallography*, 2008.

S. Janjić. Predviđanje mjesta sekundarne strukture proteina iz slijeda aminokiselinskih ostataka. Magistarski rad, FER Zagreb, 2010.

Oracle. Glassfish - open source application server, 2012. URL `http://tomcat.apache.org/`.

G. Peretin. Baza zastupljenosti metala u proteinima. Magistarski rad, FER Zagreb, 2010.

Zhang R. Levitan A.G. Hendrix D.K. Brenner S.E. Stefan, L.R. i S.R. Holbrook. Merna: a database of metal ion binding sites in rna structures. *Nucleic acids research*, 2006.

A. Tus. Baza podataka metala u proteinima, 2010.

Rakipović A. Peretin G. Tomić S. Tus, A. i M. Šikić. Biome: biologically relevant metals. *Nucleic Acids Research*, 2012.

BioMe - web sučelje baze biološki važnih metala

Sažetak

U ovom radu je opisano web sučelje za statističku analizu metala u biomolekulama. Web sučelje je dio sustava za statističku analizu izrađenog na Fakultetu elektrotehnike i računarstva na Sveučilištu u Zagrebu. Sustav se sastoji od tri dijela: baze podataka, parsera i web sučelja za statističku analizu.

MySQL baza podataka u konačnici sadrži podatke o svim strukturama, lancima, metalnim ligandima i atomima te odgovarajućim udaljenostima, kutovima i geometrijskim strukturama. Tako stvorena baza podataka, veličine oko 150 MB, služi za statističku analizu i javno je dostupna za preuzimanje. Prvenstveno su dostupni podaci za proteinske, RNK i DNK lance, ali i za sve ostale komponente biomolekula.

Statističku analizu moguće je obaviti kroz web sučelje implementirano pomoću GWT-a. Korisnik može saznati razne statističke podatke kao što su prisustvo odabranog liganda u koordinaciji s metalom, raspodjelu koordinacijski brojeva, postotak metalnih iona koordiniran s kombinacijama odabranih ligada, raspodjelu monodentatnih i bidentatnih metalnih karboksila, raspodjelu po koordinatnim geometrijama te druge. Rezultati su dostupni u brojevnom i grafičkom formatu.

Baza podataka i sučelje za statističku analizu služe kao bogat izvor informacija za istraživačku zajednicu. Uloga metala u proteinima je još uvijek slabo istraženo područje, stoga vjerujemo da će ovaj alat poslužiti znanstvenicima u budućim istraživanjima. U usporedbi s postojećim alatima, ovaj alat je daleko nadmoćniji zato što uvijek nudi najsvježiju informaciju, objedinjuje sve ponuđene podatke na jednom mjestu i nudi mnoštvo novih podataka. Osim navedenoga, sustav je brži i stabilniji. Dostupan je na `http://metals.zesoi.fer.hr`.

Ključne riječi: PDB, proteini, nukeinske kiseline, biometali, statistička analiza

BioMe - statistical analyzer for biologically relevant metals

Abstract

This paper offers an overview of the web interface of statistical analyzer for biometals. Web interface is a part of system for statistical analysis that was developed at the Faculty of Electrical Engineering and Computing at the University of Zagreb, Croatia. The system consists of three parts: a database, file parser and web interface for statistical analysis.

MySQL database (approx. 150 MB) contains information about all selected structures, chains, residues and corresponding distances, angles and geometric structures. This database is later used for statistical analysis and it is publicly available for download. Data about protein, DNA and RNA chains are mainly available as well as data about other components of the biomolecule.

Statistical analysis is carried out using a web based interface implemented using GWT. Users can obtain the following statistical properties: presence of selected ligands in a metal coordination sphere, distribution of coordination numbers, the percentage of metal ions coordinated by the combination of the selected ligands, the distribution of monodentate and bidentate metal-carboxyl, the distribution of coordination geometry and others.

The database and statistical analysis interface can be used as a rich source of information for the scientific community. The role of metals in proteins is still an area where very little research has been carried out, so we believe that this tool will be of use to researchers. In comparison with other available tools, this tool is better since it always offers up-to-date information, consolidates all the information in one place and offers a broad variety of new details. Additionally, the system is faster and more stable. It is available at `http://metals.zesoi.fer.hr`.

Keywords: PDB, proteins, nucleic acids, biometals, statistical analysis