SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Alan Tus, Alen Rakipović

BioMe - alat za statističku analizu biološki važnih metala

Zagreb, svibanj 2012.

Ovaj rad izrađen je na Sveučilištu u Zagrebu, Fakultetu Elektrotehnike i Računarstva, na Zavodu za elektroničke sustave i obradbu informacija pod vodstvom doc. dr. sc. Mile Šikića i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2011./2012.

SADRŽAJ

1. Uvod

Ovaj rad opisuje sustav za statičku analizu podataka o 3D strukturama biomolekula - *BioMe*. Sustav je potpuno automatiziran i samostalno svakog tjedna osvježava popis struktura i pokreće obradu podataka. Korisnici mogu koristiti postojeće statističke analize ili sami preuzeti cijelu bazu podataka te nad njome obavljati statističke analize. Alat je dostupan na `http://metals.zesoi.fer.hr/metals/`.

1.1. Obrazloženje teme

Jedno od preliminiranih istraživanja o utjecaju metala na proteinske komplekse napravljeno je u radu [10]. S odmakom vremena se javila potreba za osvježavanjem dobivenih rezultata i uvođenjem novih analiza. Rezultat spomenute potrebe je ovaj rad.

Usprkos činjenici da su zapisi o 3D strukturama javno dostupni postoji vrlo malo istraživanja o vezanju metala i proteina. Još manje je poznat utjecaj metala na RNA, a također i na DNA. Obzirom na navedeno, željeli smo istraživačima pružiti alat za statističku analizu kojim će u budućnosti moći na brz i jednostavan način analizirati velike količine podataka koje nudi naša baza i dobiti trenutni uvid u aktualno stanje rezultata.

BioMe je web sučelje za statističku analizu podataka. Prije nego što smo krenuli u izradu smo detaljno istražili postojeća rješenja za analizu biomolekula ([7], [16], [14], [6], [9], [22]). Svi spomenuti radovi su također web sučelja koja nude neku vrstu analize, ali se to svodi 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 našeg alata je velik broj statistika koje ostali alati nemaju. *BioMe* je također 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.

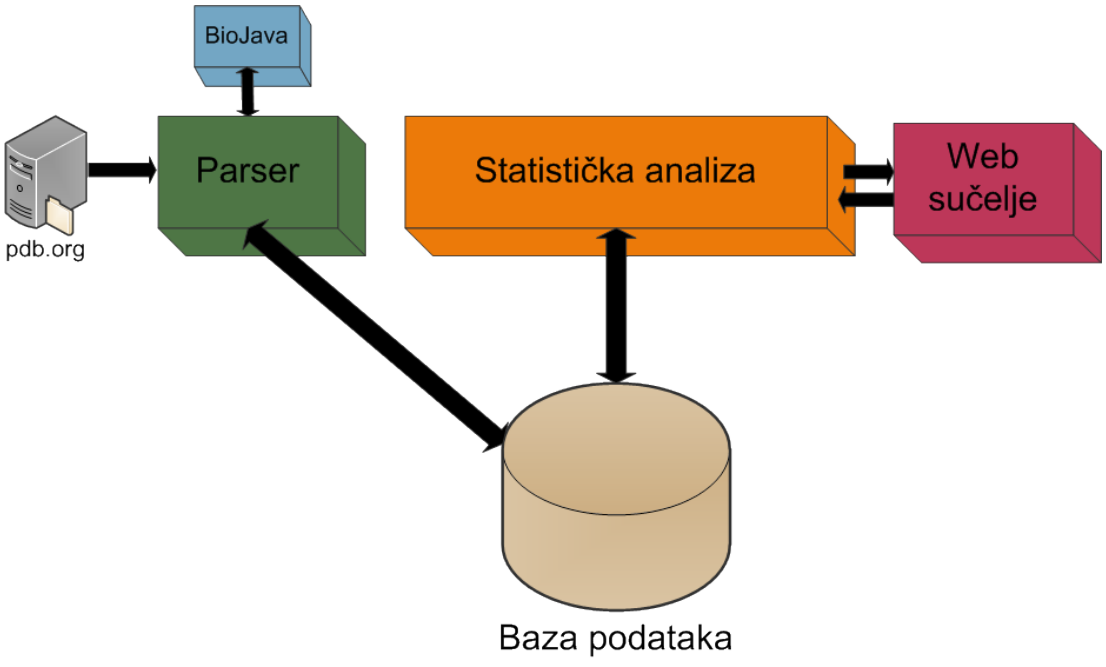
1.2. Podaci i obrada podataka

Podaci se prikupljaju iz središnje baze proteinskih struktura u *mmCIF* [17] formatu. To je 16 GB podataka koji sadrže sve dosada otkrivene i objavljene 3D strukture biomolekula. Ti podaci se potom učitavaju pomoću *BioJava* [15] alata u naš sustav koji filtrira strukture prema unaprijed zadanim kriterijima opisanim u poglavlju 3.4. Nad odabranim strukturama se vrši predobrada podataka. Ti podaci se spremaju u bazu podataka koja se koristi za statističku analizu i javno je dostupna za preuzimanje.

1.3. Statistička analiza

Web sučelje omogućuje korisniku da izabere metale, ligande, lance, razne pragove i metode obrade podataka na točno određenom skupu PDB zapisa. Statistička analiza započinje filtriranjem podataka iz baze na osnovu korisničkog upita. Nakon toga se računa sedam statistike za svaki odabrani metal, te jedna statistika za svaki odabrani ligand. Rezultati statistika se prikazuju u obliku liste stranica po odabranom metalu ili ligandu. Statistički podaci su ponuđeni u tabličnom obliku i popraćeni su grafovima.

Ligand, u ovom radu, definira sve molekule koje su povezane s metalima.



Slika 1.1: Dijagram sustava

Poglavlje 2 donosi pregled najvažnijih biomolekula te kratke opise istih. Zatim je u poglavlju 3 dan detaljan opis metoda prikupljanja podataka i početne obrade. U poglavlju 4 opisana je arhitektura sustava za statističku analizu te su objašnjeni načini računanja pojedinih statistika. Poglavlje 5 objašnjava kako se sustav automatski održava. Rezultati, te rasprava i zaključci su navedeni u daljnjim poglavljima.

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 ukupnost 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 [18].

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, znanstvenici se u eksperimentalnim metodama koriste modeliranje tijekom postupka dobivanja strukture.

2.2. Nukleinske kiseline

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

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.

Razvoj tehnologije je omogućio izolaciju, manipulaciju pa čak i sintezu nukleinskih kiselina korisnički definiranih sljedova i struktura što je dovelo do eksplozije generiranja podataka o ovim molekulama.

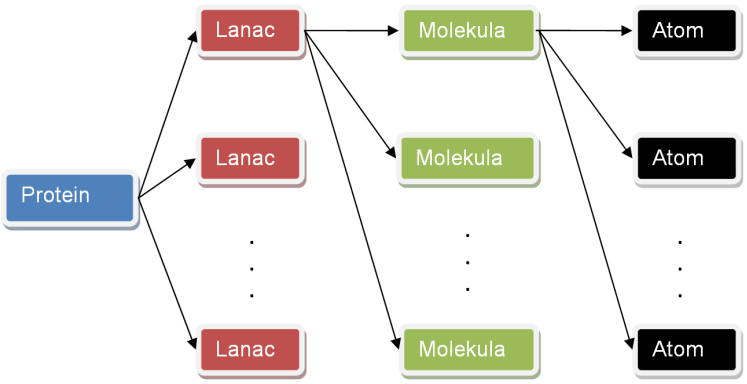
2.3. Uloga metala u biomolekulama

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 znali koliko su specifični za pojavu određenih metala i time možda stekli detaljniji uvid u funkciju proteina. Trećina ili četvrtina 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 DNA 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 podjedinica. 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. Shematski prikaz strukture biomolekule se može vidjeti na slici 2.1.



Slika 2.1: Struktura biomolekule

2.4.1. Vrste lanaca

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

- voda
- metal
- DNA lanac
- RNA 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 HOH.

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

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

Proteinski lanac 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, svrstavamo ga u **ostale** lance jer nam u ovoj inačici nije važan.

3. Prikupljane podataka

3.1. Ulazni podaci

Ulazni podaci se preuzimaju s web stranica baze proteisnkih struktura (engl. *Protein Data Bank*) [3] koje su dio Istraživačkog udruženja za strukturalnu Bioinformatiku (engl. *Research Collaboratory for Structural Bioinformatics (RCSB)*) [5]. Njihovi podaci sadrže sve dosada otkrivene i objavljene 3D strukture velikih bioloških molekula, uključujući proteine i nukleinske kiseline. Kako se svakog dana otkrivaju i objavljuju nove strukture, svakog tjedna se dodaju nove strukture u bazu podataka. Naš sustav prati ove izmjene i uvijek nudi najnovije podatke.

3.1.1. mmCIF

mmCIF (engl. *macromolecular Crystallographic Information File*) [17] je prilagodljiv i proširiv (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. Više o *mmCIF* strukturi podataka je moguće saznati na [2], a detalje o postupku kristalografije proteinskih struktura i načinima prikupljanja podataka na [8].

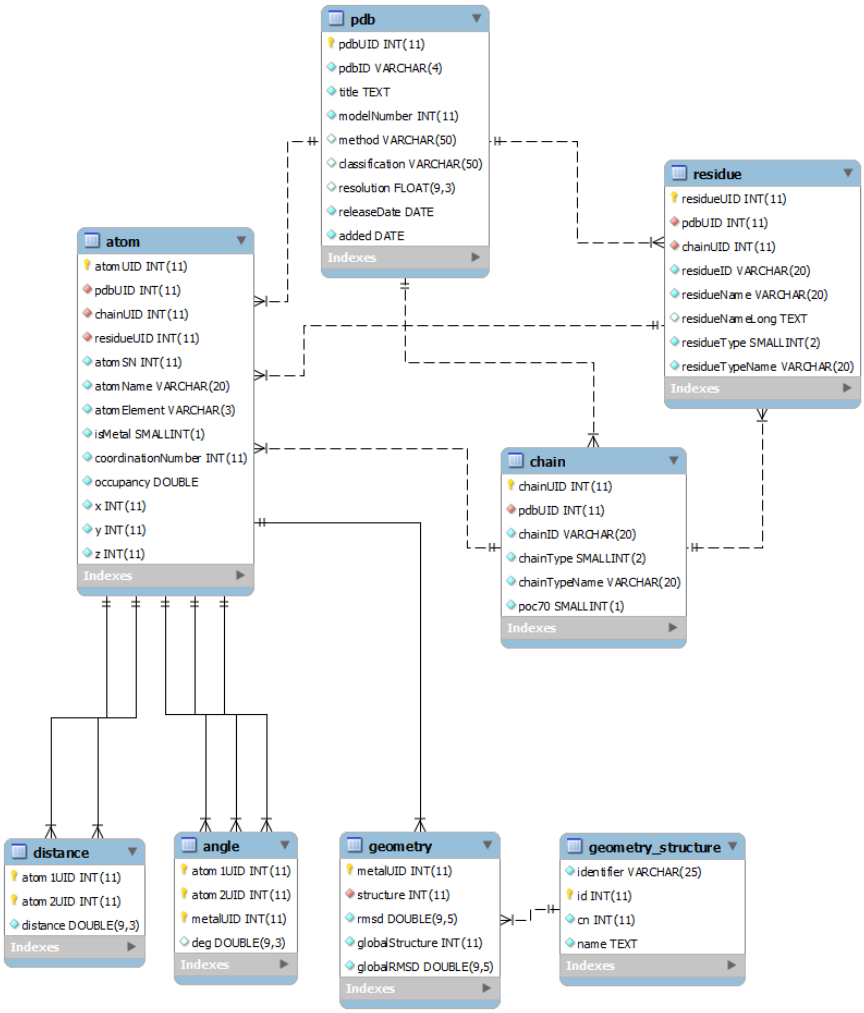
3.2. Baza podataka

Za implementaciju baze podataka korišten je *mySQL* [19] upravitelj podacima. *mySQL* je trenutno najpopularnija besplatna relacijska baza podataka.

Model baze podataka ne poštuje pravila o izradi relacijskih baza podataka u potpunosti. Relacije nisu normalizirane u treću normalnu formu. Trenutni oblik baze podataka sadrži redundantne podatke, ali to je napravljeno ciljano kako bi se smanjila

složenost upita. Radi se s velikim količinama podataka i upiti u bazu podataka bi bili vrlo složeni, a time i dugotrajni.

Baza podataka prati strukturu proteina. Vizualni prikaz odnosa relacija unutar baze podataka je vidljiv na slici 3.1.



Slika 3.1: Relacijski dijagram baze podataka

U bazi podataka se nalazi ukupno osam relacija. U nastavku je dan njihov kratki pregled i opis uloge.

pdb

Sadrži podatke o strukturama. Ti podaci uključuju jedinstvena oznaka strukture, naziv, metoda otkrivanja, klasifikaciju strukture, datum objave i slično. Na jedinstvenu oznaku strukture se vežu i stale relacije koje opisuju strukturu proteina (chain, residue i atom).

chain

Sadrži pojedinosti o svakom korištenom lancu svih unesenih struktura: jedinstvenu oznaku, oznaku strukture kojoj pripada, naziv i tip lanca (WATER, METAL, DNA, RNA, PROTEIN, OTHER). Ova relacija sadrži polje *poc70* o kojem će biti riječi u poglavlju 3.5 koje govori o *cluster 70* skupu. Na jedinstvenu oznaku lanca u relaciji chain se vežu relacije residue i atom.

Vrste lanaca i postupak njihova određivanja je opisan u poglavlju 2.4.1.

residue

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 (UNKNOWN, AMINO, DNA NUCLEOTIDE, RNA NUCLEOTIDE, WATER, METAL, OTHER). Na jedinstvenu oznaku lanca u relaciji residue se veže relacija atom.

atom

Sadrži podatke o svim korištenim atomima: jedinstvenu oznaku atoma, oznaku strukture, lanca i grupe atoma kojoj pripada, serijski broj atoma unutar strukture kojoj pripada, naziv, kemijski element, zastavicu koja označava radi li se o metalu ili nemetalu, koordinacijski broj, vjerojatnost da se atom stvarno nalazi na tom mjestu i koordinate izražene u angstromima (Å). Na jedinstvenu oznaku atoma unutar relacije atom se vežu relacije distance i angle.

Koordinacijski broj nam govori koliko se iona suprotnog naboja nalazi oko središnjeg iona.

angle

Sadrži podatke o svim mogućim kombinacijama kutova donora pojedinog metala: jedinstvene oznake dvaju atoma koji čine krakove kuta i vrh kuta te vrijednost kuta izraženu u radijanima.

distance

Sadrži podatke o udaljenosti svih donora metala koji su udaljeni manje od 3Åod metala. Udaljenosti su zapisane uz jedinstvene oznake dvaju atoma između kojih je izmjerena udaljenost. Osim udaljenosti donora, mogu se naći i udaljenosti između dvaju metala koji su udaljeni manje od 7Å.

Donori i odabrane vrijednosti za udaljenosti su objašnjene u poglavlju 3.4.2.

geometry

Sadrži sve geometrijske strukture otkrivene pomoću kutova koji su zapisani u relaciji *angle*. Relacija sadrži zapis o metalu oko kojeg je formirana geometrijska struktura, korijenu srednje kvadratne pogreške (engl. *root mean square error*) i vrstu strukture. Imena struktura i pripadajući koordinacijski broj su zapisani u relaciji **geometry structure**.

Metode određivanja geometrijskih struktura i njihov popis se nalaze u poglavlju 3.4.4.

3.3. Implementacija parsera

Parser je implemetiran u *Java* tehnologiji. Optimiziran je kako bi iskoristio višeprocesorku okolinu, stoga koristi više paralelnih dretvi pri izvršavanju. Dijagram razreda moguće je vidjeti na slici 3.2. gdje je prikazan dijagram razreda.

Izvorni kod je organiziran u tri osnovna paketa:

hr.fer.zesoi.biometals.model

Sadrži razrede MyStructure, MyChain, MyGroup i MyAtom koje odgovaraju modelu baze podataka. Ovaj skup razreda nije potrebno detaljnije opisivati jer je model baze podataka već detaljno objašnjen.

hr.fer.zesoi.biometals.parser

Sadrži razrede koje obavljaju zadatak parsiranja. Način parsiranja je detaljno obrazložen u poglavlju 3.3.3.

hr.fer.zesoi.biometals.services

Sadrži razrede koje obavljaju obradu podataka za vrijeme parsiranja i njihov detaljan pregled je dan u poglavlju 3.4.

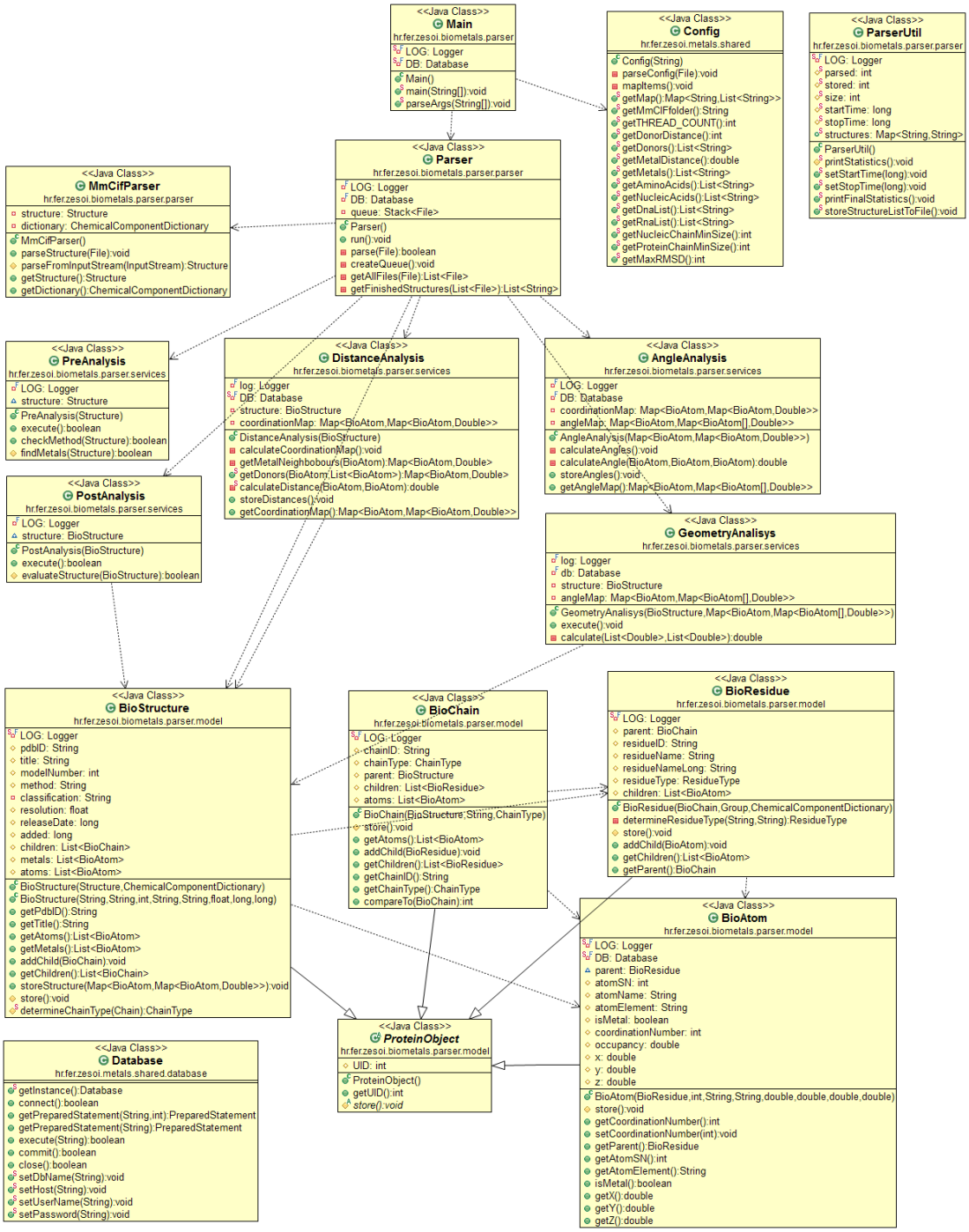
Od ostalih razreda koje se koriste u parseru, a nisu obuhvaćeni gornjom podjelom, možemo spomenuti slijedeće:

Main razred samo inicijalizira cijeli sustav, stvara zadani broj dretvi i pokreće obradu podataka. Po završetku izvršavanja gasi sustav.

Database razred obavlja sve poslove vezane uz bazu podataka: uspostavljanje veze, izvršavanje upita, potvrđivanje transakcija i zatvaranje veze.

Config razred učitava konfiguracijsku datoteku opisanu u poglavlju 3.3.1 i u sebi čuva globalne varijable potrebne za vrijeme obrade podataka.

ParserUtil razred u sebi sadrži pomoćne funkcije za praćenje rada parsera. Pamti vrijeme izvršavanja, broj obrađenih i odbačenih struktura, te funkcije za ispis spomenutih podataka.



Slika 3.2: Class dijagram za parser

3.3.1. Konfiguracijska datoteka

Pri pokretanju parsera se definira putanja do konfiguracijske datoteke. Ona omogućava brzu i jednostavnu izmjenu programskih parametara o kojima će biti riječi u slijedećim poglavljima. U suprotnom bi ti parametri morali biti upisani u samom kodu sustava i samim time pri svakoj promjeni zahtijevali izmjenu koda i ponovno kompajliranje.

Sadržaj konfiguracijske definira slijedeće stvari:

- pristupne podatke za bazu podataka (ime baze, adresa, korisničko ime i lozinku)
- naziv direktorija u kojem se nalaze dohvaćene strukture u *mmCIF* obliku
- broj dretvi za pokretanje
- najveću udaljenost donora od metala u angstromima (Å) i popis donora
- najveću udaljenost metala od metala u angstromima (Å) i popis metala
- popis aminokiselina i nukleinskih kiselina
- najmanju duljinu proteinskog lanca i najmanju duljinu nukleinskog lanca
- najveći dozvoljeni korijen srednje kvadratne pogreške (engl. *root mean square error*) pri potrazi za geometrijskim strukturama

```
#BioMetalsParser config file

db_name = metali_dev
db_host = jdbc:mysql://localhost:3306/
db_user = metali
db_pass = lozinka

mmcif_folder = mmCIF

thread_count = 8

donor_distance = 3
donors = O, N, CL, S, F

metal_distance = 7
metals = FE, NI, MN, CA, CU, NA, MG, K, CO, ZN, CD, V,
        MO, W, PB, HG, AG, AU, SR, CS, PT, BA, TL, SM,
        RB, YB, LI, KR, LU, CR, OS, GD, TB, LA HO, GA,
        CE, RU, RE, PR, IR, EU, AL, TE, SB, PD

amino_acids = ALA, ARG, ASN, ASP, CYS, GLU, GLN, GLY,
        HIS, ILE, LEU, LYS, MET, PHE, PRO, SER, THR,
        TRP, TYR, VAL
nucleic_acids = A, T, G, C, U, DU, DC, DT, DG, DA

proteinChainMinSize = 49
nucleicChainMinSize = 5

maxRMSD = 50
```

Slika 3.3: Primjer konfiguracijske datoteke

3.3.2. BioJava

U implementaciji je korišten paket *BioJava* [15]. To je projekt otvorenog koda koji predstavlja razvojni okvir za obradu bioloških podataka u Javi. Pruža mogućnost učitavanje često korištenih oblika zapisa bioloških podataka (između ostalog i *mmCIF* datoteka), analitičke i statističke alate, dopušta manipuliranje biološkim sekvencama i 3D strukturama, te još mnoštvo funkcija. Cilj projekta jest omogućiti brz razvoj aplikacija iz područja bioinformatike. *BioJava* je dio Open Bioinformatics Foundation [20] udruženja, neprofitne organizacije sačinjene od volontera usredotočenih na razvoj programske podrške u bioinformatici. Iz iste organizacije su potekli i drugi slični projekti kao što su BioPython, BioPerl, BioRuby, BioPerl i drugi [20].

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 nedavno izašao iz razvoja u produkciju. U planu je ažuriranje na ovu inačicu.

Za potrebe ovog rada su nedostajale neke metode i članske varijable unutar struktura te parser nije vraćao potrebne podatke o strukturama. Kako je ovo projekt otvorenog koda preuzeli smo izvorni kod i prilagodili ga našim potrebama.

3.3.3. Višedretvenost

Parser razred inicijalizira popis zadataka i implementira metodu *run()* koja predstavlja posao jedne dretve. Metodu *run()* čini niz naredbi koji je vidljiv u pseudokodu 1.

```
begin
  dohvatiZadatak;
  BioJava.ucitajStrukturu;
  if (!Predanaliza.provjeriStrukturu())
    stop;
  end
  stvoriBioMeStrukturu;
  if (!Postanaliza.provjeriStrukturu())
    stop;
  end
  analizirajUdaljenosti;
  analizirajKutove;
  analizirajGeometriju;
  spremiPromjene;
end
```

Pseudokod 1: Pseudokod metode Parser.run()

Pseudokod 1 će u idućem poglavlju (3.4) biti detaljnije opisan.

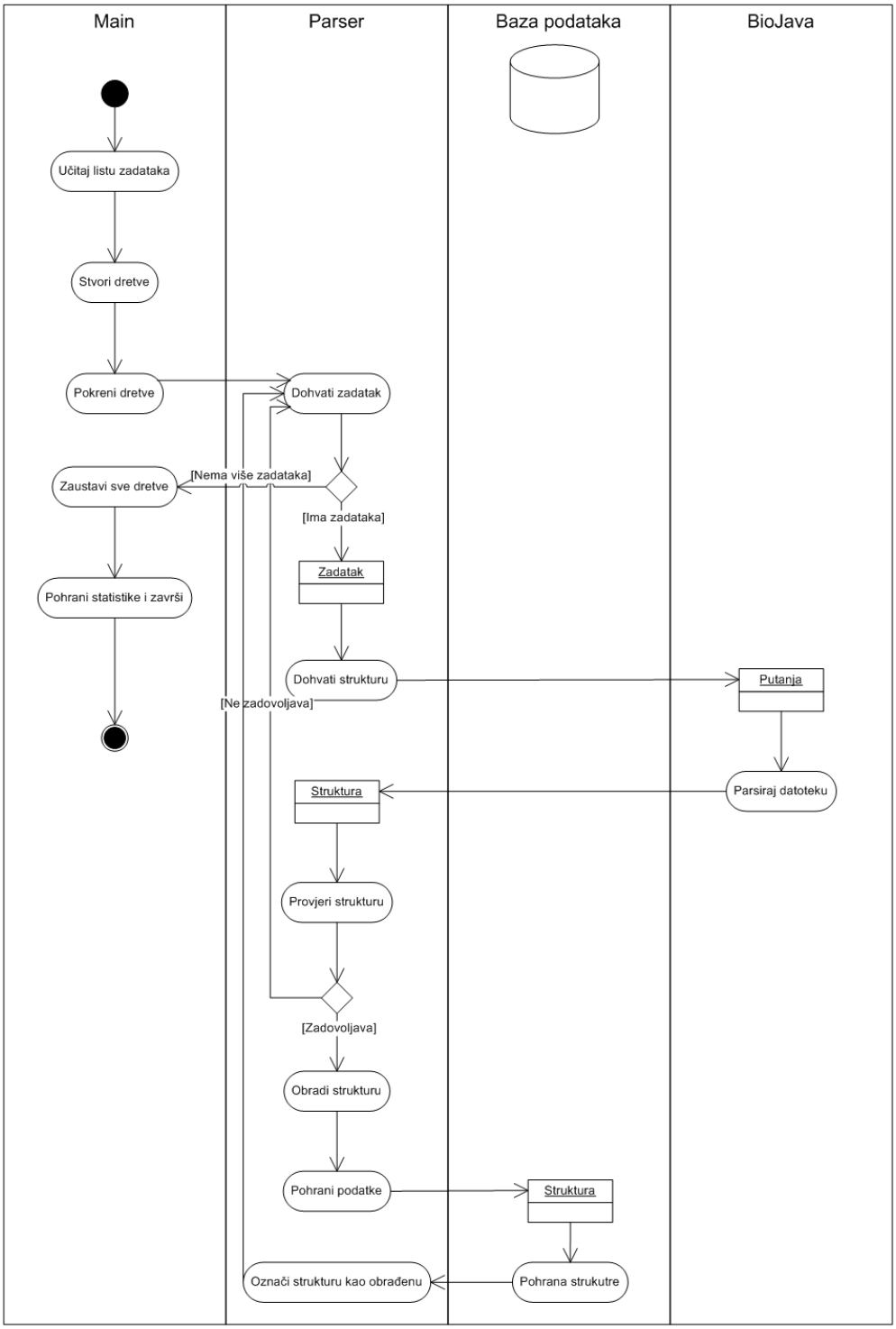
U svrhu ubrzanja procesa obrade podataka koristi se višedretvenost. Na taj način smo paralelizirali procese obrade ulaznih podataka, maksimalno iskoristili procesorske resurse te vrijeme koje smo inače izgubili čekajući da se struktura učita iz datoteke i da se podaci upišu u bazu podataka sada koristimo 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 te se može pokrenut i zaustaviti. Kada se stvaraju dretve pozivom *new Thread(Runnable class)*, dovoljno je predati referencu na razred *Parser* koji implementira *Runnable* sučelje. Pokretanjem dretve, pokrenuti će se i obrada podataka, tj. metoda *run()*.

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 pri inicijalizaciji programa stvara listu zadataka u obliku popisa putanja do datoteka sa strukturama koje je potrebno obraditi. Nakon toga se stvara unaprijed definirani broj dretvi. Eksperimentalno smo utvrdili da je broj dretvi za optimalan rad 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 iz liste zadataka preuzima po jedan zadatak, obrađuje ga i potom preuzima novi zadatak. Pristup listi zadataka je sinkroniziran tako da dvije dretve ne mogu istovremeno preuzeti isti zadatak.



Slika 3.4: Dijagram aktivnosti za parser

3.4. Metode obrada podataka

Svaka dretva zasebno vrši obradu podataka. Kao što je vidljivo iz pseudokoda 1, prvi korak je dohvaćanje novog zadatka iz liste zadataka. Zadatak se sastoji od putanje do datoteke sa strukturom koju se učitava pomoću *BioJava* paketa.

3.4.1. Analiza strukture

Svaka struktura prolazi predanalizu sadržanu u *PreAnalysis* razredu kojom se jednostavno i brzo eliminira velik broj struktura: **Struktura mora biti otkrivena korištenjem jedne od prihvatljivih metoda** Pri otkrivanju novih struktura, znanstvenici se ponekad u eksperimentalne svrhe služe metodama čiji rezultati zahtijevaju daljnju provjeru. Za potrebe ovog rada, prihvatljive metode su sve one koje uključuju korištenje rendgenskih zraka (engl. *X-ray*) i nuklearne magnetske rezonancije (*NMR*-a). **Struktura mora sadržavati barem jedan metal** Iteracijom po svim atomima strukture traži se barem jedan metal iz skupa metala koji su učitani u konfiguracijskoj datoteci. Popis metala se čuva u *Config* razredu.

Ukoliko struktura zadovolji ova dva uvjeta prelazi se na obradu podataka. Prvo se cijela struktura prevede iz *BioJava* strukture u našu implementaciju strukture. Pritom se zadržavaju samo korisni podaci. Ovo se radi kako bi se očuvala jednostavnost i brzina sustava. Za vrijeme pretakanja podataka se određuje vrsta svih lanca i izdvajaju se svi metali u posebnu listu. Vrste lanaca i uvjeti koje moraju zadovoljavati su objašnjeni u poglavlju 2.4.1.

Druga faza analize strukture se sastoji od slijedećih zahtjeva:

Struktura mora sadržavati barem jedan proteinski lanac odnosno DNA ili RNA lanac

Proteinski lanac mora imati barem dva donora, odnosno nukleinski barem jedan donor

Ukoliko proteinska struktura zadovoljava ova dva daljnja uvjeta prelazi se na daljnju obradu podataka i pohranu.

3.4.2. Analiza udaljenosti

Ova analiza prima cijelu strukturu i vraća popis metala i njihovih donora u obliku mape čiji su ključevi metali, a vrijednosti, liste donora. Rezultat se naziva *koordinacijskom mapom*.

Za svaki metal se određuje donore i susjedne metale. Nemetal može biti donor ako je na udaljenost manjoj od 3Åi ne mora biti na istom lancu.*Donori* mogu biti svi nemetali, ali u ovom radu se uzimaju u obzir samo slijedeći: O, N, CL, F i S. Dva metala su susjedi ukoliko se nalaze na udaljenosti manjoj od 7Åi ako su smješteni na istom lancu. Tako izračunate vrijednosti se vraćaju u obliku prethodno spomenute *koordinacijske mape*.

Odabrana je ista vrijednost za udaljenost metala i donora kao i u [10]. Vrijednost od 7Åza udaljenost metala od metala je određena eksperimentalnom metodom. Pokazalo se da je većina metala smještena unutar 7Åjedan od drugoga.

3.4.3. 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.

3.4.4. Geometrijska analiza

Geometrijska analiza na ulaz prima mapu kutova iz prethodne analize. Postupak određivanja geometrijske strukture je dan u pseudokodu 2.

```
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 2: Pseudokod metode Parser.run()

Za svaki metal i njegove kutove se izuzmu kutovi u posebnu strukturu gdje se poredaju po veličini od najvećeg do najmanjeg. Zatim se taj popis kutova uspoređuje sa svim strukturama koje imaju jednak broj kutova. Kutovi poznatih struktura su već poredani po veličini. Poredak po veličini nam je važan u slijedećem koraku kako bismo dobili što manji korijen srednje kvadratne pogreške (engl. *root mean square error, RMSD*). Pamtimo sve tako otkrivene geometrijske strukture i pogreške. Nakon završetka iteracija odaberemo strukturu s najmanjom pogreškom i zapišemo ju u bazu podataka.

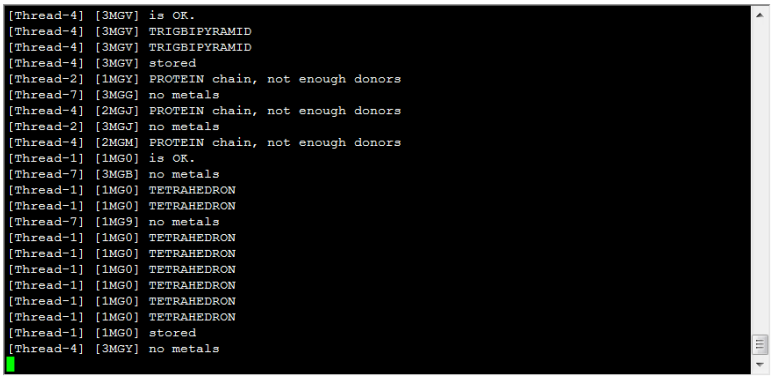
U tablici 3.1 su popisane sve geometrijske strukture koje su obuhvaćene ovim radom.

Ime strukture	Koordinacijski broj
tetrahedron	4
Squareplanar	4
Trigonalbipyramid	5
Squarepyramid(tetragonalbipyramid)	5
Octahedron	6
Trigonalprism	6
Octahedron,face-capped	7
Trigonalprism,squarefacemonocapped	7
Pentagonalbipyramid	7
Dodecahedron(bisdisphenoid)	8
Cube	8
Hexagonalbipyramid	8
Trigonalprism,squareface-bicapped	8
Squareantiprism	8
Trigonalprism,triangularface-bicapped	8
Squareantiprism,monocapped	9
Trigonalprism,squareface-tricapped	9
Squareantiprism,bicapped	10
Cuboctahedron	12
Anticuboctahedron	12
Icosahedron	12
Hexagonalantiprism,bicapped	14

Tablica 3.1: Popis svih geometrijskih struktura koje su obuhvaćene ovim radom

Nakon što su obavljene sve analize određuju se korisni atomi koji će biti pohranjeni u bazu podataka. Korisni atomi su svi oni koji se spominju u udaljenostima, kutovima i geometrijskim strukturama. Uz njih se pohranjuju i svi viši dijelovi strukture (molekule, lanci i struktura) koji ih sadrže. Izdvajaju se samo korisni atomi kako bi baze podataka bila što manja, a time upiti što brži. U bazu podataka se također pohranjuju sve izračunate udaljenosti, kutovi i geometrije. Ovime je završila obrada jedne strukture i dohvaća se novi zadatak.

Na slici 3.5 je vidljiv rad parsera.



Slika 3.5: Prikaz ispisa parsera za vrijeme rada

3.5. Cluster 70 skup

Cluster 70 predstavlja skup proteinskih lanaca grupiranih u grozdove (engl. *cluster*) unutar kojih je međusobna sličnost pojedinih lanaca 70% ili više. Datoteka sa opisom tog skupa proteinskih lanaca se generira jednom tjedno i može se preuzeti s PDB poslužitelja [4]. Budući da se popis *cluster 70* lanaca osvježava svakog tjedna potrebno je i ovu informaciju redovito osvježavati. Ovaj posao obavlja jednostavna *Python* skripta koja je u nastavku opisana.

Cluster 70 datoteka je organizirana kao tablica gdje jedan redak označava jedan lanac. Stupci su redom 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 pdbID: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.

Skripta se spaja na poslužitelj i sama preuzima najnoviju *cluster 70* datoteku. Potom odgovarajućim lancima u bazi podataka postavlja ili briše zastavicu koja označava da lanac pripada *cluster 70* skupu. Ažuriranje prati rang lanaca u *cluster 70* datoteci, od onog s najmanjim do onog s najvećim rangom. Prvi lanac na koji naiđe, a da postoji u bazi podataka, označava kao *cluster 70* lanac i prelazi na novi grozd.

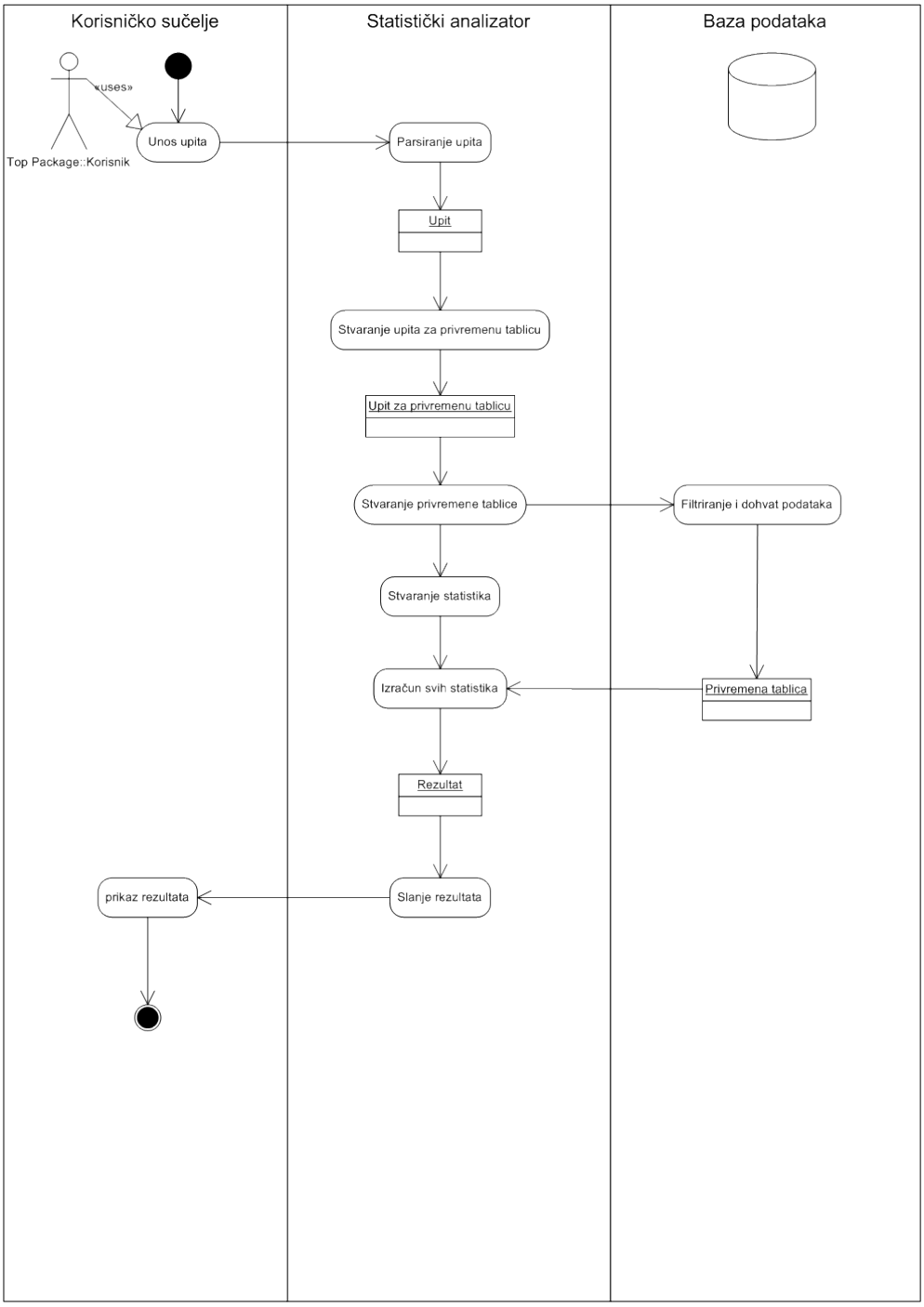
U budućnosti se planira funkcionalnost ove skripte uklopiti u rad parsera.

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, šalje zahtjev poslužitelju, našem 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.



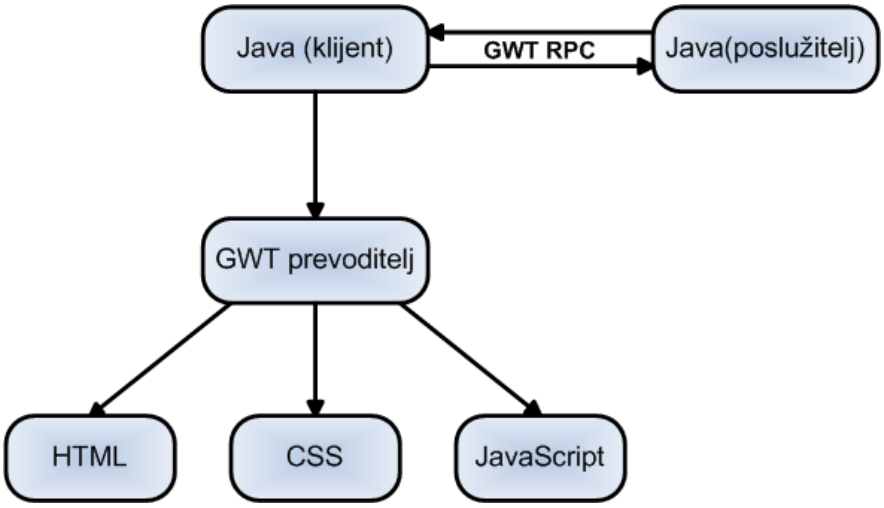
Slika 4.1: Dijagram aktivnosti za statističku analizu

4.1.1. Korištene tehnologije

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

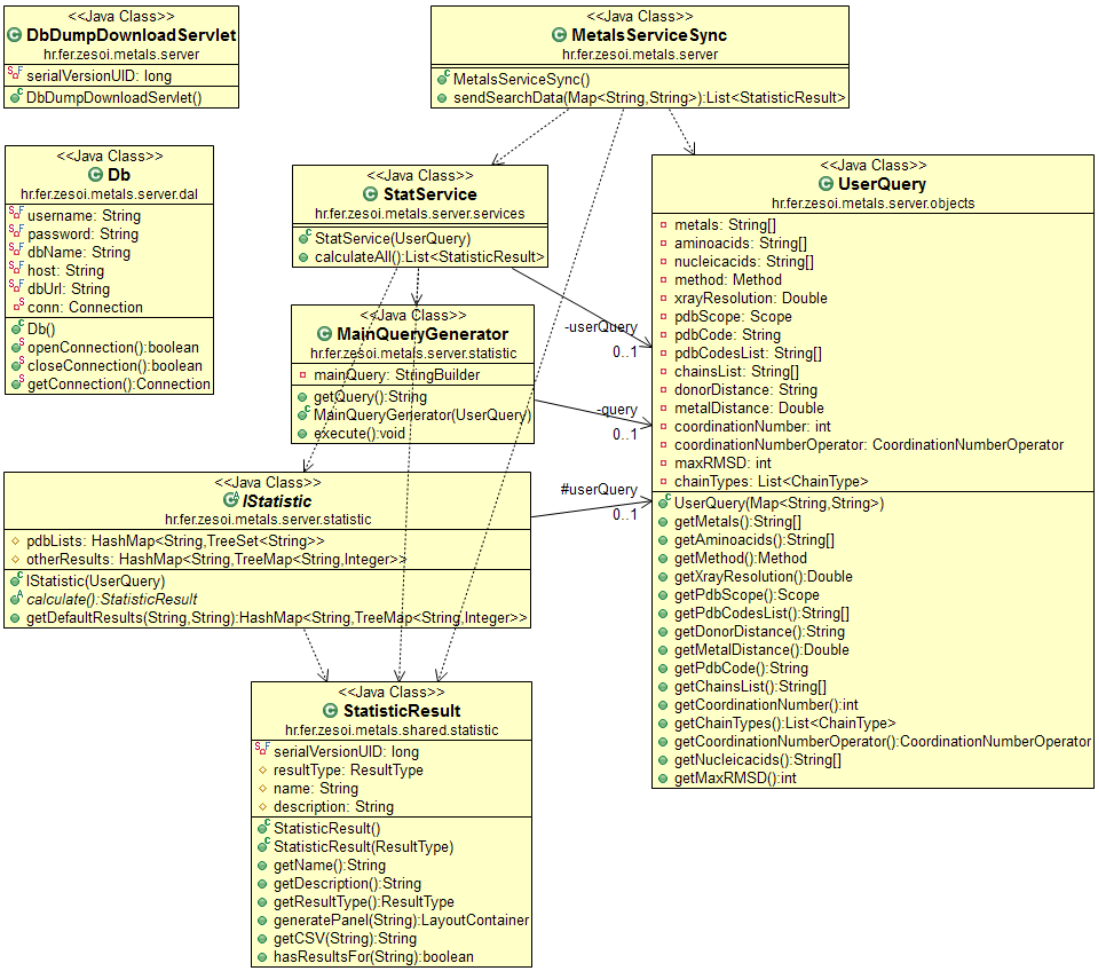
GWT alat za izradu web aplikacija temeljenih na programskom jeziku Java. Pred-

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.

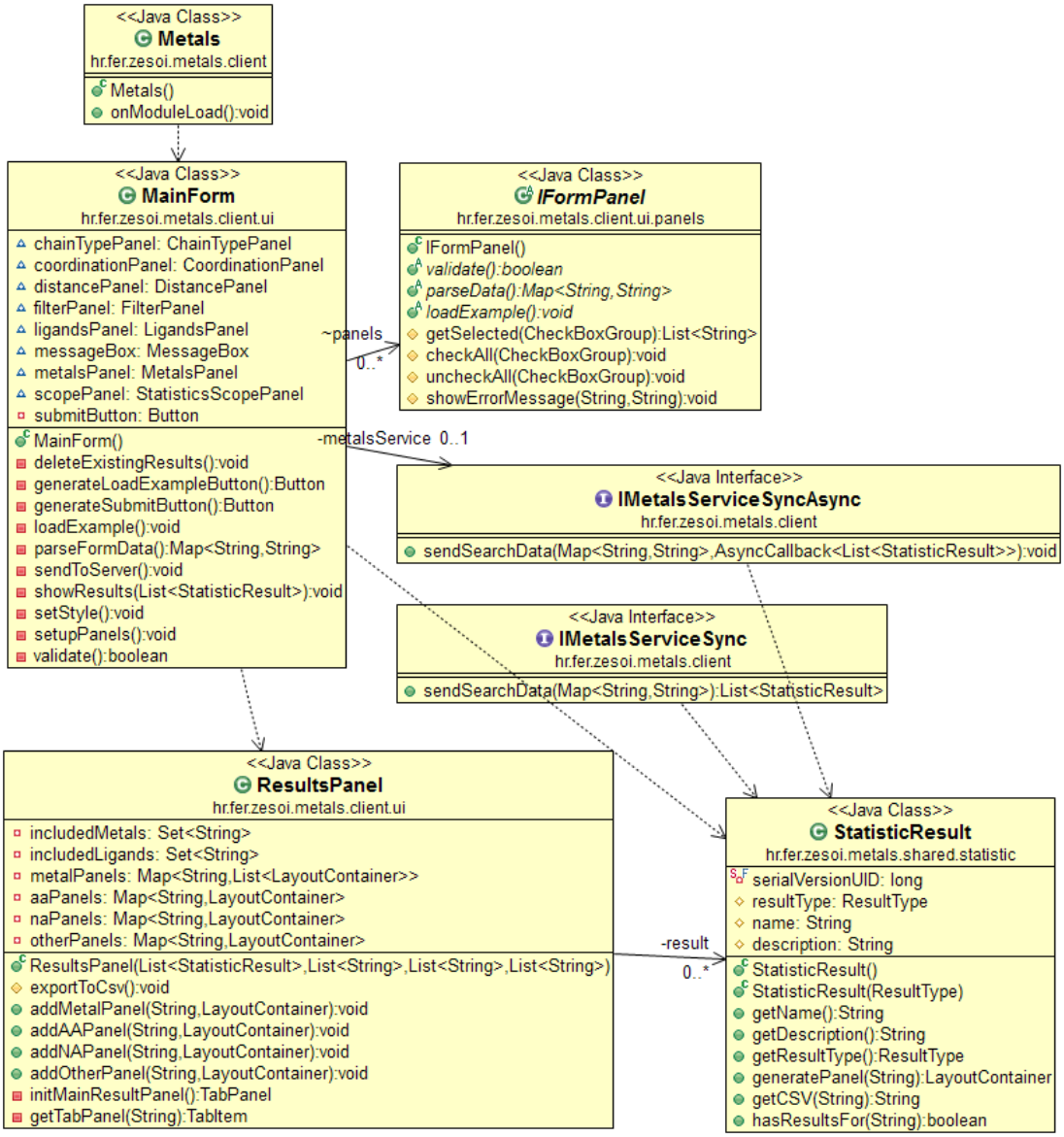
StatService razred predstavlja servis za stvaranje svih statistika te poziva za svaku statistiku poziva metodu za izračunavanje rezultata.

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

Na slici 4.4 je prikazan 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 [13] 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 vremenom 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 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 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 metala 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 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, korisnici je omogućen pristup dump-u podataka iz baze podataka gdje su zastupljeni svi metali iz baze proteina.

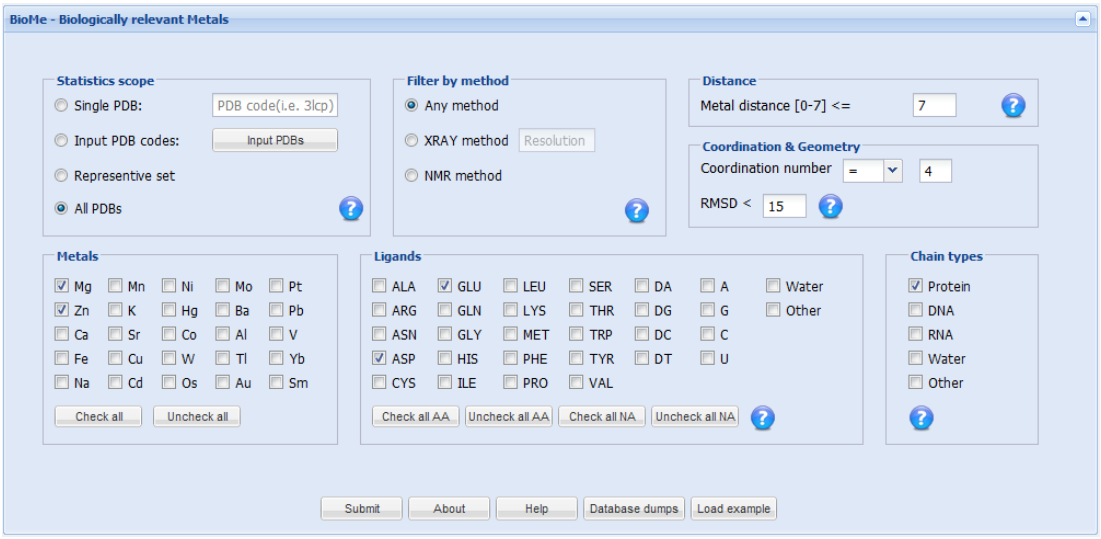
Korisničko sučelje pruža odabir pet različitih vrsta liganda tako je moguće odabrati između aminokiselina, DNA i RNA nukleotida, vode te ostalih. Budući da je velik broj ostalih liganda, a njihova zastupljenost u lancima je vrlo mala, oni su u prikazu statistika stavljeni u zajedničku skupinu. Međutim, 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.

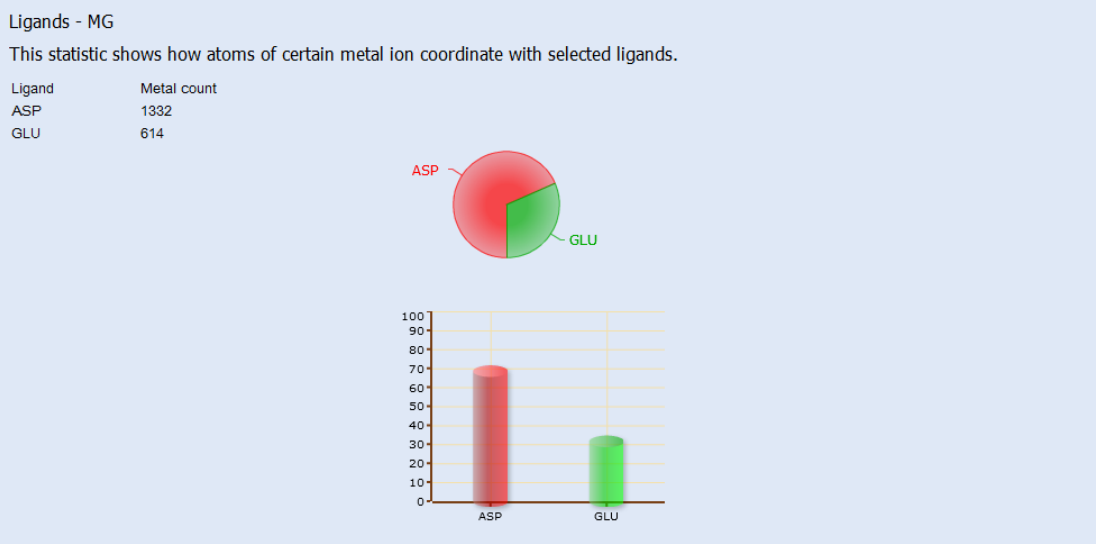
Razlikujemo dvije vrste statistika: statistike koje se računaju za označeni metal (*M1 - M7*) te statistika koja se računa za označene ligande (*L1*). U nastavku su navedene i pobliže objašnjene sve statistike.



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

4.3.1. Distribucija veza metala s odabranim ligandima - M1

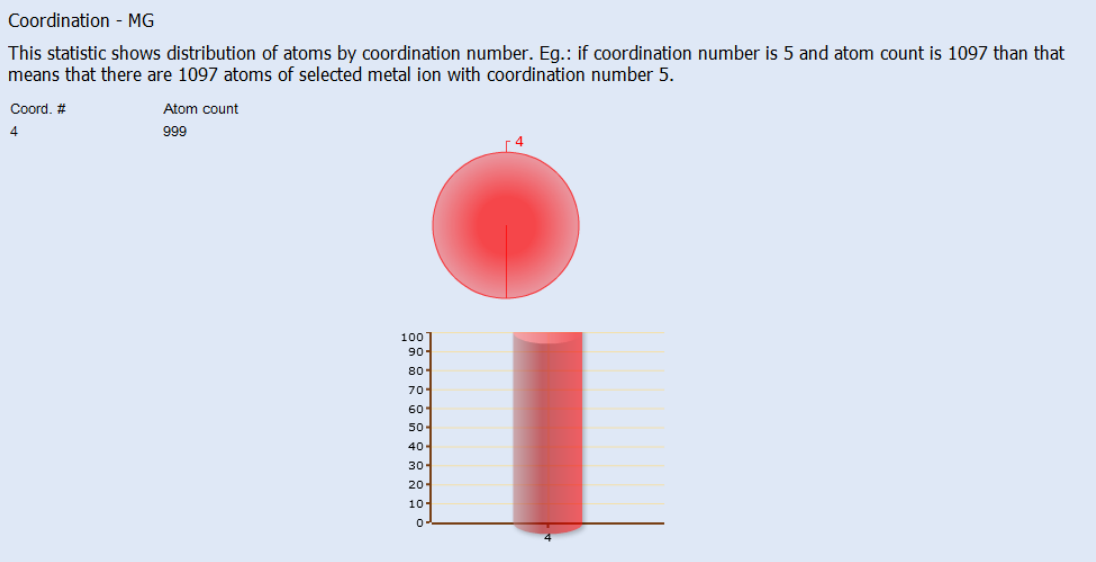
Ova statistika pokazuje kako se atomi pojedinog iona metala vežu na odabrane ligande. Metal je u vezi s ligandom ako je barem jedan atom iz tog liganda udaljen manje od 3Åod bilo kojeg atoma metala. Ukupni broj svih veza metala i odabranog liganda je zbroj veza svih atoma tog metala s ligandom.



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

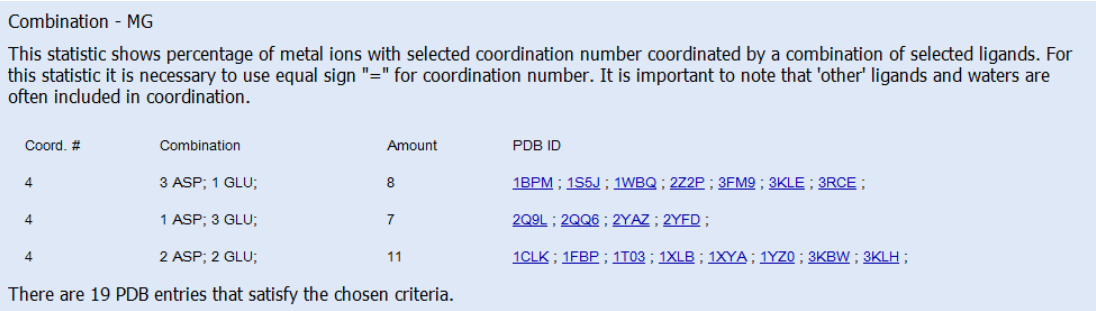
Ova statistika pokazuje kombinacije liganda s kojima metali međudjeluju. Za prikaz rezultata ove statistike potrebno je odabrati samo jedan koordinacijski broj (odabirom znaka jednakosti u korisničkom sučelju). Ukoliko postoji kombinacija svih označenih liganda s pojedinim atom označenog metala, izračunava 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 3.

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

Pseudokod 3: 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.

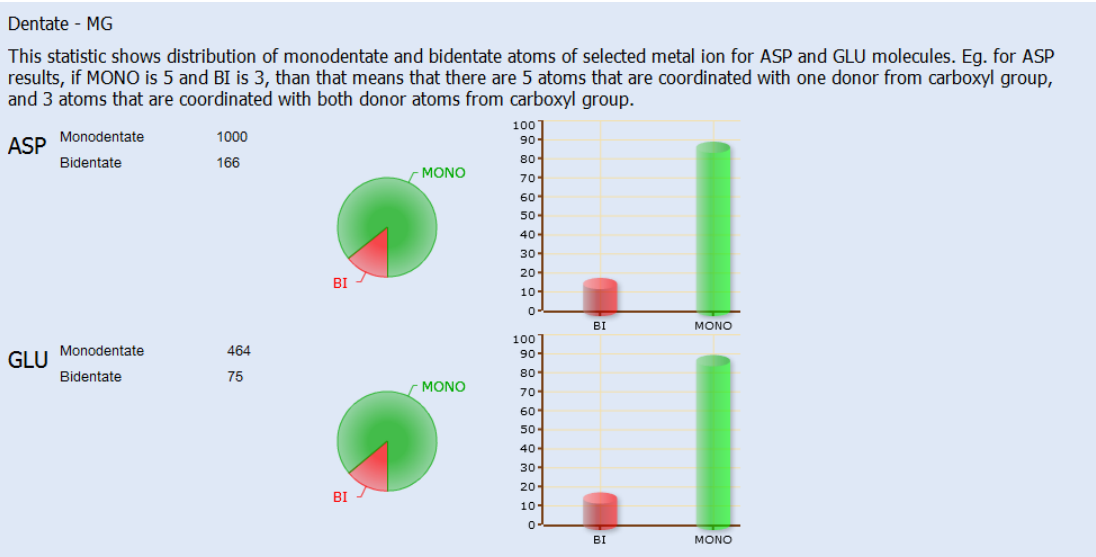
Coord. #	Combination	Amount	PDB ID
4	3 ASP; 1 GLU;	8	1BPM ; 1S5J ; 1WBQ ; 2Z2P ; 3FM9 ; 3KLE ; 3RCE ;
4	1 ASP; 3 GLU;	7	2Q9L ; 2QQ6 ; 2YAZ ; 2YFD ;
4	2 ASP; 2 GLU;	11	1CLK ; 1FBP ; 1T03 ; 1XLB ; 1XYA ; 1YZ0 ; 3KBW ; 3KLH ;

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 4 opisuje kako se izračunava ova statistika. Računanje započinje do-

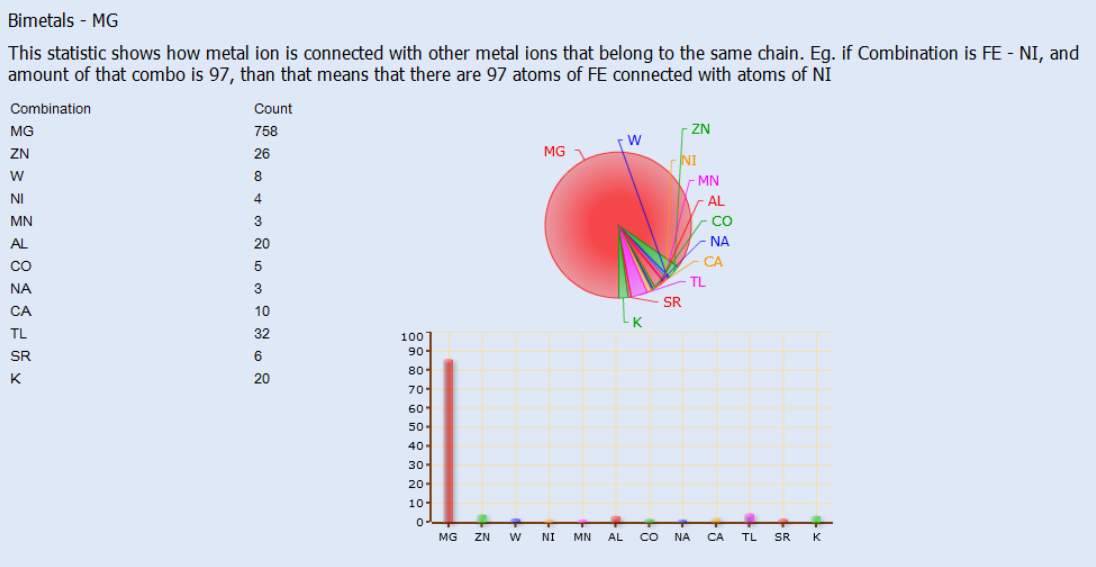
hvatom svih atomUID-ova iz privremene tablice koji zadovoljavaju korisnikov upit. Tada se za svaki atomUID prebrojavaju sve veze koje taj atom ostvaruje s aminokiselinom. Metoda *PrebrojiSveRazliciteVezeMetalAA();* jednim SQL upitom prebrojava sve različite veze metala i aminokiseline bez obzira na atom, 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 4: Pseudokod statistike M4

4.3.5. Distribucija atom metala vezanih za isti lanac - M5

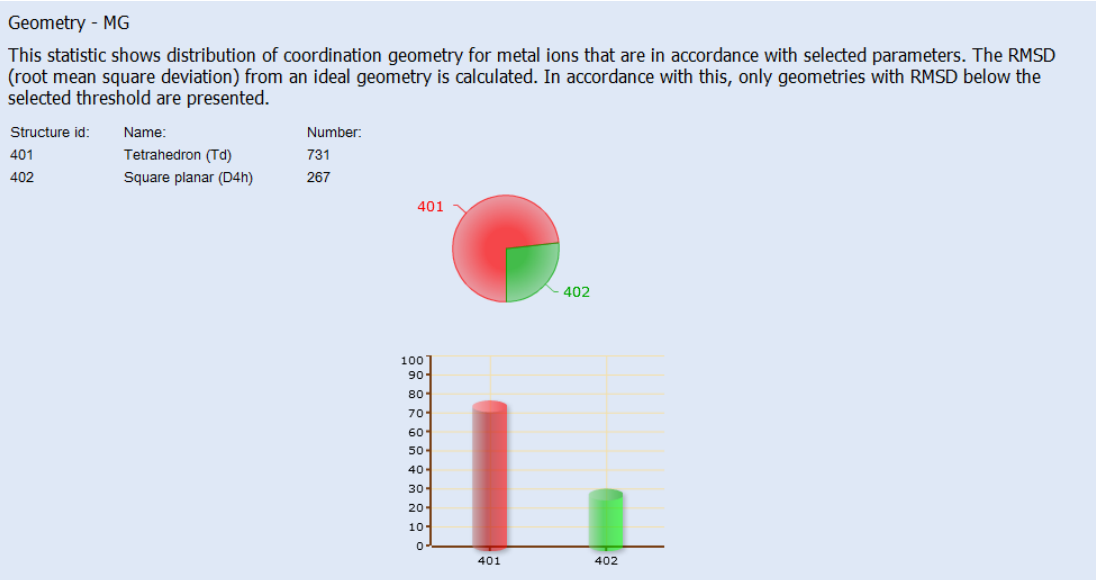
Ova statistika pokazuje broj atoma metala povezanih s odabranim metalom koji se nalaze na istom lancu. Atomi su povezani ako su na istom lancu te su udaljeni manje od željene udaljenosti u Åod bilo kojeg atoma tog metala. Željena udaljenost se može odabrati u korisničkom sučelju, a prepostavljena vrijednost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



Slika 4.10: Rezultat izračunavanja statistike M5

4.3.6. Distribucija koordinacijske geometrije po metalima - M6

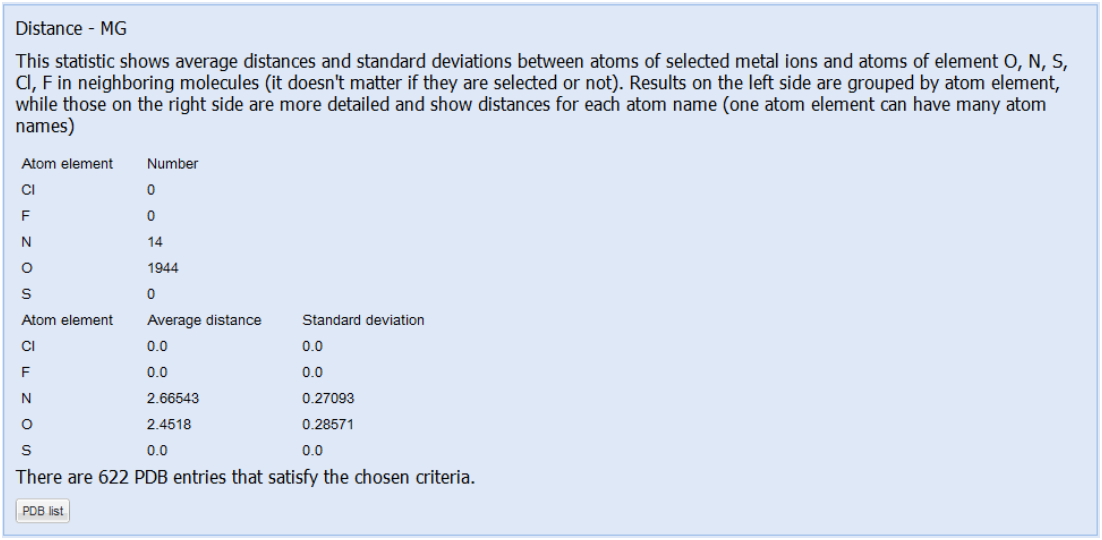
Ova statistika pokazuje broj atom metala u zadanoj geometrijskoj strukturi s obzirom na parametre zadane preko korisničkog sučelja. Vrijednost RMSD parametra predstavlja korijen srednje vrijednosti devijacije od idealne geometrijske strukture te se u rezultatu pribrajaju oni rezultati koji su ispod zadane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



Slika 4.11: Rezultat izračunavanja statistike M6

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.



Distance - MG

This statistic shows average distances and standard deviations between atoms of selected metal ions and atoms of element O, N, S, Cl, F in neighboring molecules (it doesn't matter if they are selected or not). Results on the left side are grouped by atom element, while those on the right side are more detailed and show distances for each atom name (one atom element can have many atom names)

Atom element	Number
Cl	0
F	0
N	14
O	1944
S	0

Atom element	Average distance	Standard deviation
Cl	0.0	0.0
F	0.0	0.0
N	2.66543	0.27093
O	2.4518	0.28571
S	0.0	0.0

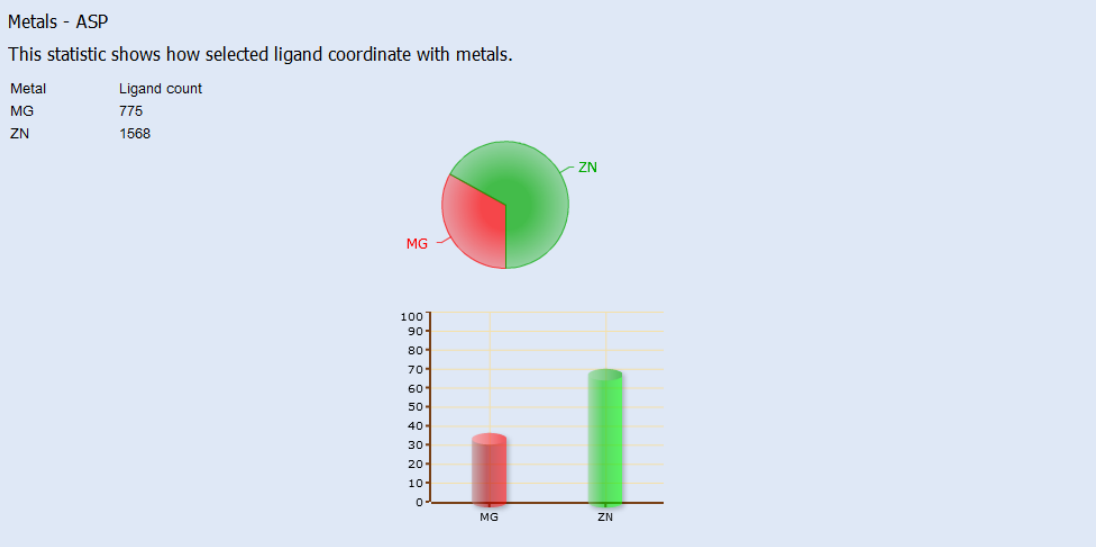
There are 622 PDB entries that satisfy the chosen criteria.

PDB list

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. Automatsko održavanje

Proteinska baza podataka [3] se ažurira svake srijede. Odmah nakon ovog ažuriranja se pokreće niže opisana skripta koja ažurira naš sustav. Za njeno pokretanje će se brinuti *cron*.

Cron je *GNU/Linux* alat koji omogućava precizno definiranje periodičnog pokretanje neke naredbe ili skupa naredbi.

Redoslijed koraka za ažuriranje sustava:

1. Preuzmi nove zapise o 3D strukturama s udaljenog poslužitelja.
2. Isprazni pomoćnu bazu podataka
3. Pokreni parsiranje novih podataka
4. Pokreni POC70 skriptu
5. Spremi pomoćnu bazu podataka u .sql file koji će biti dostupan za preuzimanje
6. Zaustavi web servisi
7. Isprazni glavnu bazu podataka i napuni ju novim podacima
8. Pokreni web servis

Gornja skripta nam omogućava da u roku tri do četiri sata od ažuriranja središnje baze podataka, imamo sustav s najnovijim podacima spreman za upotrebu.

6. Rezultati

6.1. Baza podataka i parsiranje

Ulazni skup koji čini 16 GB podataka je sveden na bazu podataka veličine 100-150 MB oblikovanu na način koji pogoduje računanju statistika. U nastavku je dan pregled filtriranih podataka. Važno je napomenuti da je ovo trenutno stanje baze koje se mijenja svaki tjedan (zbog promjene podataka iz izvornog skupa [3]).

U tablici 6.1 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 6.1: 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 3A. Tablica 6.2 sadrži broj takvih donora te isto tako i broj onih donora iz ulaznog skupa koje nisu zadovoljile navedeno ograničenje.

Tablica 6.2: Udaljenosti struktura od metalnih iona

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

Vrste zastupljenih lanaca dane su u tablici 6.3. 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 6.3: Vrste lanaca

Vrsta lanca	Broj lanaca
METAL	159837
PROTEIN	46326
OTHER	30349
WATER	27133
DNA	1050
RNA	868

Budući da BioMe baza podataka sadrži više od dvije tisuće različitih liganada, u tablici 6.4 je dan pregled 15 najzastupljenijih. Esencijalna aminokiselina histidin koja je važna za mentalno i fizičko zdravlje čovjeka je prva na popisu.

Tablica 6.4: Top 15 liganada

Skraćeno	Broj	Puni naziv
HIS	51996	HISTIDINE
ASP	49531	ASPARTIC ACID
CYS	36259	CYSTEINE
GLU	32642	GLUTAMIC ACID
G	19725	GUANOSINE-5'-MONOPHOSPHATE
A	13896	ADENOSINE-5'-MONOPHOSPHATE
ASN	11272	ASPARAGINE
U	9888	URIDINE-5'-MONOPHOSPHATE
C	8836	CYTIDINE-5'-MONOPHOSPHATE
THR	6865	THREONINE
GLY	6840	GLYCINE
SER	6309	SERINE
GLN	4668	GLUTAMINE
TYR	4119	TYROSINE
VAL	4035	VALINE

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 statistika, u tablici 6.5 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 6.5: 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 6.6), od kojih su octahedron i tetrahedron najzastupljenije.

Tablica 6.6: Koordinacijske geometrije

Ime strukture	Koordinacijski broj	Broj struktura
Octahedron (Oh)	6	30911
tetrahedron (Td)	4	24160
Trigonal bipyramid (D3h)	5	15775
Trigonal prism	6	7836
Square planar (D4h)	4	4893
Trigonal prism, square face monocapped (C2v)	7	4819
Square pyramid (tetragonal bipyramid, C4v)	5	3918
Octahedron, face-capped (C3v)	7	2165
Pentagonal bipyramid (D5h)	7	1258
Square antiprism (D4d)	8	921
Trigonal prism, square face-bicapped	8	574
Square antiprism, bicapped (D4d)	10	203
Dodecahedron (bisdisphenoid, D2d)	8	171
Square antiprism, monocapped	9	166
Hexagonal bipyramid (D6h)	8	165
Trigonal prism, triangular face-bicapped	8	34
Anticuboctahedron	12	26
Cube (Oh)	8	18
Hexagonal antiprism, bicapped	14	8
Trigonal prism, square face-tricapped	9	5

Tablica 6.7 prikazuje raspodjelu geometrijski struktura po koordinacijskim brojevima, tj. koliko atoma čini strukturu.

Tablica 6.7: 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

7. Rasprava i Zaključci

7.1. Problemi u razvoju

Krenuli smo s razvojem parsera. Prva inačica je bila napisana u skriptnom jeziku *Python*. Obrada podataka nije sadržavala sve današnje elemente i trajala je gotovo pet dana. Slijedeća inačica je bila u *Javi*. Zahvaljujući višedretvenosti i malo optimizacije korištenih metoda, vrijeme izvršavanja je smanjeno s početnih devet sati na samo tri sata. Ovo je detaljnije opisano u [1]. Daljnja optimizacija je zasigurno moguća, no dobitci su zanemarivi tako da trenutno nije prioritet.

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 [11] i prelaskom na *Tomcat* [12] web poslužitelj taj problem je riješen. Problem je bio uzrokovan ograničenjima u implementaciji *Glassfish* [21] 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. Do tada se nismo pridržavali ovog pravila i zato su se javljale iznimke. Oba problema su uspješno riješena.

S početnih 15 minuta koliko je trajao dohvat najkompliciranijeg rezultata na prvoj inačici statističkog analizatora, postigli smo da najkompliciraniji upit traje 60 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).

7.2. Daljnje nadogradnje

Slijedeći prioritet u razvoju statističkog analizatora jest uvođenje višedretvenosti pri računanju statistika. Ovime bi omogućili paralelno izračunavanje statistika i dodatno ubrzali prikaz rezultata. Trenutna arhitektura analizatora omogućava jednostavan prijelaz na višedretveni način rada.

SQL upite kojima se dohvaćaju rezultat iz baze podataka se u pojedinim statistikama može dodatno optimizirati, kao i samu bazu podataka.

Budući da je trenutno prvi korak pri statističkoj analizi učitavanje podatka u privremenu tablicu, kao logično rješenje se nameće baza podataka pohranjena u radnoj memoriji računala. Za ovu vrstu nadogradnje bi se prvo moralo pojednostavniti podatkovne strukture kojima se sustav koristi, što bi predstavljalo dodatnu uštedu radne memorije.

Sva navedena poboljšanja su planirana u budućnosti i predstavljaju slijedeći korak u izgradnji ovog sustava.

Mnogi upiti se često ponavljaju i nema potrebe da se svaki put ponovno izračunavaju, ako se baza podataka nije osvježavala. Ovo se može postići uvođenjem *cacheiranja rezultata*, tj. privremene pohranu već izračunatih rezultata u bazu podataka po sistemu ključ-vrijednost.

U slučaju da *BioMe* postane servis koji će koristiti milijuni korisnika ili da *BioMe* baza podataka višestruko naraste ovo bi sigurno bilo nužno. Autori ovog sustava ovo trenutno ne vide kao potreban korak izgradnje sustava, ali su spremni na njega kada za to dođe vrijeme.

8. Zahvale

Ovim putem autori zahvaljuju Goranu Peretinu na izgradnji prve verzije statističkog analizatora, te pomoć u početnim mjesecima rada na sustavu.

Hvala Sanji i Antoniji Tomić za biološka i kemijska objašnjenja, definiranje zahtjeva, te pomoć u provjeri ispravnosti rada BioMe sustava.

Hvala Antoniji Burčul što nam je sa strpljenjem objašnjavala kompliciranе biološke termine i nekoliko puta lektorirala radove.

Hvala Marku Čupiću što nas je naučio Javu.

Također, želimo zahvaliti mentoru Mili Šikiću na beskonačno izgubljenih živaca, sati i svega ostalog u ove dvije-tri godine rada na sustavu.

LITERATURA

[1] Tus A. Baza podataka metala u proteinima. Magistarski rad, FER Zagreb, 2010.

[2] RCSB Protein Data bank. The macromolecular crystallographic information file (mmcif), 2012. URL http://mmcif.rcsb.org/pubs/methenz.html.

[3] The Protein Data Bank. The protein data bank, 2012. URL http://www.pdb.org.

[4] The Protein Data Bank. Clusters, 2012. URL ftp://resources.rcsb.org/sequence/clusters/.

[5] The Worldwide Protein Data Bank. The worldwide protein data bank (wwpdb), 2012. URL http://www.wwpdb.org/.

[6] 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.

[7] 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.

[8] Protein Crystallography. Protein crystallography, 2012. URL http://proteincrystallography.org/.

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

[10] Šikić M. Dokmanić I. i Tomić S. Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. *Biological Crystallography*, 2008.

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

[12] The Apache Software Foundation. Apache tomcat, 2012. URL http://tomcat.apache.org/.

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

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

[15] R. C. G. Holland, T. A. Down, M. Pocock, A. Prlić, D. Huen, K. James, S. Foisy, A. Dräger, A. Yates, M. Heuer, i M. J. Schreiber. Biojava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–2097, 2008. doi: 10.1093/bioinformatics/btn397. URL http://bioinformatics.oxfordjournals.org/content/24/18/2096.abstract.

[16] 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.

[17] European Bioinformatics Institute. Macromolecular structure database group mmcif information, 2012. URL http://www.ebi.ac.uk/msd/documentation/mmcif.html.

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

[19] MySQL. Mysql the world's most popular open source database, 2012. URL http://www.mysql.com/.

[20] OBF. Open bioinformatics foundation, 2012. URL http://open-bio.org/.

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

[22] 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.

BioMe - alat za statističku analizu biološki važnih metala

Sažetak

Autori: Alan Tus, Alen Rakipović

U ovom radu je opisan sustav za statističku analizu metala u biomolekulama izrađen na Fakultetu elektrotehnike i računarstva na Sveučilištu u Zagrebu. Sustav se sastoji od tri dijela: baze podataka, parsera i web sučelja za statističku analizu. Potpuno je automatiziran te se osvježava svakog tjedna nakon ažuriranja centralne proteinske baze podataka (http://www.pdb.org/) iz koje se preuzimaju podaci. Razvoj traje već treću godinu i u planu su još mnoge nadogradnje i proširenja.

Višedretveni parser mmCIF datoteka je implementiran u Java tehnologiji i optimiziran tako da iskoristi prednosti višeprocesorskog okruženja. Obavlja posao odabira, analize i spremanja odabranih struktura iz oko 16 GB ulaznih podataka. 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, RNA i DNA 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

Authors: Alan Tus, Alen Rakipović

This paper offers an overview of the statistical analyzer for biometals which 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. It is completely automated, so it does a weekly refresh after the update of the central Protein Data Bank (http://www.pdb.org/), which is also our main data source. The system has been in development for almost 3 years and many further upgrades and expansions are yet to come.

The multi-threaded mmCIF file parser is implemented in Java technology and optimized so that it utilizes the multiprocessor environment. It selects and analyses approx. 16 GB of structures and then stores the ones which fulfill the predetermined criteria into a database. The resulting 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 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