Matematički fakultet
Univerzitet u Beogradu

Deskriptivni modeli istraživanja podataka – Primena u bioinformatici

Master rad

Mentor:
Prof. dr Gordana Pavlović-Lažetić

Kandidat:
Predrag Stefanović

Beograd, 2012.

SADRŽAJ

1 UVOD

Prema klasičnoj strukturno-funkcijskoj paradigmi dobro definisana prostorna struktura je preduslov za funkcionisanje proteina. Ovom uverenju u prilog ide više od 50 000 proteinskih struktura pohranjenih u proteinskoj bazi podataka (PDB). Posmatranjem ovih struktura na atomskom nivou dolazimo do elementarnih saznanja kako određeni proteini funkcionišu. Ipak, niz istraživanja u poslednjih desetak godina ukazuje da paradigma ne može da se primeni na sve proteine. Rastući skup eksperimentalno sakupljenih podataka pokazuje da postoji veliki broj funkcionalnih proteina koji ne zauzimaju jedinstvenu, uravnoteženu tercijalnu strukturu, već strukturu u kojoj se pozicije atoma i sam polipeptidni lanac vremenom menjaju. Za proteine sa takvom strukturom kažemo da su **suštinski nestruktuirani** ili **neuređeni**. U ljudskom proteomu ima oko 12% potpuno neuređenih proteina i oko 50% proteina sa barem jednim dugim (> 30AA) neuređenim regionom. Najveća javno dostupna baza neuređenih proteina je DisProt, koja sadrži više od 1300 neuređenih regiona (decembar 2011).

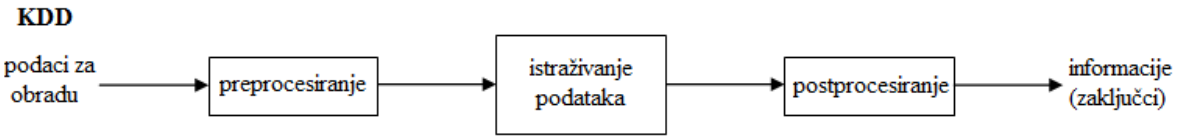
Uočeno je da regulatorni proteini, zatim oni koji su umešani u procese signalizacije i molekulskog prepoznavanja, češće sadrže neuređene regione. Funkcionisanje nekih neuređenih proteina, koji obavljaju upravo ove biološke procese, povezuje se sa teškim bolestima poput raka i neurodegenerativnim bolestima. Preduslov za konstruisanje izglednih terapija i lekova je proširivanje klasične strukturno-funkcijske paradigme, odnosno razumevanje kako neuređeni proteini funkcionišu i šta uzrokuje pojavu neuređenosti.

U ovom radu se upotrebom tehnike istraživanja podataka – analize pridruživanja, traže pravilnosti, potencijalne uzročno-posledične veze između fizičkohemijskih i biohemijskih svojstava aminokiselina i tendencije da se aminokiseline češće ili ređe pojavljuju u neuređenim regionima nego u ostatku, dobro struktuiranom delu proteina. Rezultati bi trebalo da doprinesu boljem razumevanju sastava neuređenih regiona.

Prvo poglavlje ovog teksta posvećeno je osnovama tehnika koje su korišćene u istraživanju i uvođenju najrelevantnijih bioloških pojmova – DNK, geni, proteini, proces sinteze proteina i struktura proteina. Drugo poglavlje opisuje računske modele i prezentuje prva merenja nad podacima iz *DisProt* i *AAindex* baza. Neophodne transformacije podataka, pretraga pravila pridruživanja i rezultati su predstavljeni u trećem.

1.1 Otkrivanje znanja i bioinformatika

Otkrivanje znanja iz podataka (eng. KDD – *Knowledge Discovery from Data*) je interdisciplinarna oblast čiji fokus čine metode za ekstrakciju netrivijalnih i potencijalno upotrebljivih informacija iz nekog skupa podataka. Ova oblast je nastala kao indirektna posledica tehnološkog razvoja u sferi prikupljanja podataka i njihovog sistematskog čuvanja. Naime, u medicini, meteorologiji ili nekoj drugoj nauci, često se podaci veoma brzo akumuliraju tako da nakon relativno kratkog perioda za njihovo skladištenje je potrebno nekoliko terabajta ili možda čak petabajta. Obrada tako velikih baza podataka predstavlja izazov i to neprikladan za uobičajene (tradicionalne) metode. Iz tih razloga nastaje KDD.



Slika 1.1

Centralnu, analitičku ulogu u KDD ima **Istraživanje podataka** (eng. *Data Mining*), *Slika 1.1*, automatizovan proces za otkrivanje potencijalno zanimljivih informacija iz pripremljenog skupa podataka.[1] Tehnike istraživanja podataka dobijene su kombinovanjem metoda iz oblasti statistike, mašinskog učenja i sistema za upravljanje bazama podataka, pritom prevazilazeći osnovne izazove skalabilnost i otpornost na visoku dimenzionalnost podataka.

U poslednje dve decenije došlo je do gotovo eksplozivnog rasta biomedicinskih podataka, počev od onih prikupljenih tokom farmaceutskih studija i razvoja terapija protiv raka do onih dobijenih raznim istraživanjima u genomici i proteomici (ponavljajuće sekvence, genetske funkcije, interakcija proteina). Ovaj ubrzani rast jednako prati razvoj biotehnologije i metoda za analizu bioloških podataka što oblikuje naučnu granu koja obećava: *bioinformatiku*. **Bioinformatika** se neformalno može definisati kao primena informacionih tehnologija na obradu bioloških podataka, što podrazumeva smeštanje, ekstrakciju, organizaciju, analiziranje, tumačenje i korišćenje informacija iz bioloških sekvenci i molekula. Stalni cilj bioinformatike je da pomogne razumevanje bioloških procesa.

Istraživanje podataka i bioinformatika se idealno slažu, zapravo jedan od razloga razvoja KDD i tehnika istraživanja jeste upravo potencijalna primena u modernoj biologiji. Neke od primena istraživanja podataka u bioinformatici su pronalaženje i grupisanje gena, zaključivanje funkcija proteina, određivanje dijagnoza za neka oboljenja itd.

1.2 Istraživanje podataka

Kako istraživanje podataka – **IP** funkcioniše, odnosno kako se od velike skupine sirovih (neobrađivanih) podataka dolazi do željenih informacija i odgovora? Osnovna ideja jeste modeliranje. Za neko konkretno pitanje (npr. *Da li klijentu A treba odobriti kredit?*), formira se model – izdvajanjem reprezentativnih primera ili određivanjem relevantnih matematičkih veza nad podacima kod kojih je odgovor već poznat, nakon čega se taj model primenjuje na nove podatke i on na izlazu daje odgovor na postavljeno pitanje.

1.2.1 Prediktivni i deskriptivni modeli

Cilj IP-a je pronalaženje modela koji će najbolje opisati podatke sa kojima radi, a neke od tehnika koje se koriste u tu svrhu su *klasterovanje*, *klasifikacija, regresija*, *pretraga pravila pridruživanja* itd. Model po svojoj prirodi može biti **prediktivni** ili **deskriptivni**.

Prediktivni modeli predviđaju konkretnu vrednost jednog atributa na osnovu vrednosti ostalih atributa. U ovom kontekstu atribut čija se vrednost predviđa naziva se *zavisna promenljiva*, a atributi čije se vrednosti koriste za formiranje predviđanja nazivaju se *nezavisnim promenljivim*. Prediktivne modele možemo podeliti u dva tipa: **klasifikacija** (za predviđanje diskretnih zavisnih promeljivih) i **regresija** (za predviđanje kontinualnih zavisnih promenljivih). Cilj prediktivnog modeliranja je da se formira (nauči) model koji minimalizuje grešku između predviđene vrednosti i prave (poznato tačne) vrednosti zavisne promenljive.

Deskriptivni modeli otkrivaju šablone (eng. *patterns*) – korelicije, trendove, klastere itd, koji opisuju međusobne veze i zavisnosti u podacima. Za razliku od prethodnog, deskriptivni model služi za ispitivanje i opisivanje osobina posmatranih podataka i često dobijeni rezultati zahtevaju dodatno razmatranje i pojašnjavanje njihovog značaja.

1.2.2 Pravila pridruživanja

Analiza pridruživanja (eng. *Association analysis*) je metodologija koja se koristi za otkrivanje interesantnih i skrivenih veza između skupova stavki odnosno objekata. Ova metoda je poznata i pod imenom *analiza potrošačke korpe* (eng. *Market basket analysis*). Na primer, analiza pridruživanja omogućava da saznamo koju kombinaciju proizvoda ili usluge potrošači žele da kupe odnosno plate. Uočeni trendovi u kupovini potrošača dobijeni analizom pridruživanja mogu se koristiti za predviđanje ponašanja kupaca i u budućnosti.

Same veze među skupovima stavki se prikazuju uz pomoć pravila koja nazivamo **pravila pridruživanja**, kao npr:

- 72% potrošača koji kupuju mleko takođe kupuju hleb i jaja. Ovo pravilo se može primeniti na 20% svih transakcija.
- U 80% slučajeva kada potrošač kupuje pivo, on kupuje i čips.

Formalno, neka je $I = \{i_1, i_2, \dots, i_m\}$ skup svih stavki. Neka je D skup svih transakcija, gde se svaka transakcija $T \in D$ sastoji od stavki iz skupa I, tako da uvek važi $T \subseteq I$. Svaka transakcija ima svoj jedinstveni identifikator koji nazivamo TID. Neka je A skup stavki. Kažemo da transakcija T sadrži A ako i samo ako $A \subseteq T$.

Pravilo pridruživanja je implikacija oblika $A \Rightarrow B$, gde je $A \subset I$, $B \subset I$ i $A \cap B = \emptyset$. Značaj pravila $A \Rightarrow B$ može se posmatrati kroz njegovu **podršku** (eng. *support*) i **pouzdanost** (eng. *confidence*):

$$\text{Podrška, } s(A \Rightarrow B) = P(A \cup B) = \frac{\sigma(A \cup B)}{N}$$

$$\text{Pouzdanost, } c(A \Rightarrow B) = P(B|A) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

gde je:

$$\sigma(X) = |\{T|\ X \subseteq T,\ \ T \in D\}|$$

$$N\ =\ |D|$$

Odavde vidimo da se pouzdanost pravila $A \Rightarrow B$ može izraziti preko podrške skupa A i skupa $A \cup B$, tj:

$$c(A \Rightarrow B) = \frac{s(A \cup B)}{s(A)}$$

Pritom, problem istraživanja (otkrivanja) pravila pridruživanja se može formalno definisati: Za dati skup transakcija D, pronaći pravila za koja važi *podrška* $\geq$ *minsup* i *pouzdanost* $\geq$ *minconf*, gde su *minsup* i *minconf* unapred zadati pragovi podrške i pouzdanosti.[1] Pravila koja ispunjavaju zadate pragove *minsup* i *minconf* nazivamo **jakim** ili **dobrim pravilima** (eng. *strong rules*), a skupove čija podrška prelazi prag *minsup* nazivamo **čestim skupovima** (eng. *frequent itemsets*).

Dakle, ukoliko znamo podršku skupova A, B i $A \cup B$ onda je trivijalno odrediti da li su pravila $A \Rightarrow B$ ili $B \Rightarrow A$ dobra pravila.

Obično se proces pretrage pravila pridruživanja izvršava u dva koraka:

1. Pronalaženje svih čestih skupova.
2. Generisanje dobrih pravila na osnovu pronađenih čestih skupova.

Kako je izvršavanje prvog koraka generalno skuplje od izvršavanja drugog, efikasnost pretraživanja pravila pridruživanja se svodi na efikasnost pronalaženja čestih skupova.

Značajan problem u pretrazi čestih skupova predstavlja potencijalno ogroman broj čestih skupova, posebno ako je prag *minsup* mali. Naime, za svaka dva skupa stavki X i Y važi – *princip antimonotonosti*:

$$X \subseteq Y \ \ \Rightarrow \ \ s(X) \geq s(Y)\ .$$

Npr. ako je skup dužine 100 čest, onda su i svi njegovi podskupovi česti, a on ima jako puno podskupova:

$$\binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{100} = 2^{100} - 1 .$$

Samo pamćenje informacija o podršci za sve ove skupove je jako skupo. Da bi se ovaj problem prevazišao uvode se koncepti *zatvorenog čestog skupa* i *maksimalno čestog skupa*.

Skup stavki X je **zatvoren** nad skupom transakcija D, ako ne postoji njegov pravi nadskup Y takav da ima istu podršku kao skup X. Skup stavki X je **maksimalno čest** skup nad skupom transakcija D, ako je X čest i ako ne postoji njegov pravi nadskup Y tako da je Y čest nad D.[2]

Neka je C skup svih zatvorenih i čestih skupova stavki, a M skup svih maksimalno čestih skupova stavki nad skupom podataka D. Pretpostavimo da znamo podršku svih elemenata iz skupova C i M. Primetimo da na osnovu informacija o podršci elemenata skupa C možemo tačno da odredimo podršku svih čestih skupova stavki nad skupom podataka D, dok na osnovu informacija o podršci elemenata skupa M možemo da odredimo koji su skupovi česti i da njihovu podršku ograničimo odozdo. U oba slučaja broj indeksiranih čestih skupova je znatno smanjen u odnosu na malopre pomenuti primer.

U svakom slučaju, zbog ogromnog broja izračunavanja koje je neophodno izvršiti, zadatak pronalaženja čestih skupova nije trivijalan. Najpoznatiji i dovoljno efikasni algoritam koji se koristi za izdvajanje čestih skupova stavki iz velikog broja transakcija je *Apriori* algoritam.

1.2.3 Apriori algoritam i generisanje pravila pridruživanja

Apriori pronalazi sve skupove stavki sa podrškom ne manjom od zadatog praga *minsup*, odnosno pronalazi sve česte skupove stavki. Njega karakteriše algoritam obilaska grafa po širini koji pritom koristi *antimonotonost* podrške (pomenutu u prethodnom odeljku: „Ako skup stavki nije čest onda nijedan njegov nadskup nije čest.“). Apriori pravi nekoliko prolaza kroz skup transakcija. U prvom prolazu računa podršku pojedinačnih stavki (jednočlanih skupova) i određuje koje su od njih česte. Zatim, u svakom sledećem prolazu kolekcija čestih skupova pronađenih u prethodnom koraku koristi se za generisanje nove kolekcije potencijalno čestih skupova, zvanih *skupovi kandidati*, a u samom prolazu kroz transakcije izračunava se podrška svakog tog kandidata. Pred kraj svakog prolaza oni skupovi kandidati koji zadovoljavaju ograničenje zadato pragom *minsup* svrstavaju se u novu kolekciju čestih skupova koja se onda koristi u narednom prolazu. Algoritam se izvršava sve dok kolekcija čestih skupova ne bude prazna.

Po konvenciji, Apriori pretpostavlja da su stavke u svakoj transakciji sortirane u leksikografskom poretku. Skup sa k stavki se naziva k-skup, i kaže se da je veličine ili dužine k. Označimo kolekciju svih čestih skupova dužine k sa F_k, a kolekciju svih skupova kandidata dužine k sa C_k. Apriori algoritam je dat u prikazu 1.1.
U prvom prolazu kroz skup transakcija on samo prebrojava pojavljivanja pojedinačnih stavki, odnosno određuje česte 1-skupove. Dalje, k-ti ($k > 1$) prolaz se sastoji od dve faze.

Prvo, česti skupovi F_{k-1} pronađeni u $(k-1)$ prolazu se koriste da bi se generisala kolekcija C_k, a taj posao obavlja posebna generatorska funkcija *apriori-gen* (o kojoj će biti reči kasnije). Druga faza se sastoji od samog prolaza kroz transakcije kada se računa podrška za sve skupove kandidate iz kolekcije C_k. Pri računanju podrške koristi se funkcija *podskup*, koja za zadatu transakciju t određuje skupove kandidate koji jesu podskupovi od t.

Apriori algoritam

ulaz: skup transakcija D (sa leksikografski sortiranim stavkama),
minsup;
izlaz: skup svih čestih skupova stavki nad skupom transakcija D;

F_1 = {česti 1-skupovi};
for (k = 2; $F_{k-1} \neq \emptyset$; k ++) **do begin**
    C_k = apriori-gen(F_{k-1}); // generisanje skupova kandidata
    foreach $t \in D$ **do begin**
        C_t = podskup(C_k, t); // kandidati koji su podskup transakcije t
        foreach $c \in C_t$ **do**
            c.brojač ++;
    end
    F_k = { $c \in C_k$ | c.count $\geq minisup$ }
end
Na izlaz → $\bigcup_k F_k$;

Prikaz 1.1

Funkcija *apriori-gen* za argument uzima skup F_{k-1}, a vraća C_k – nadskup skupa F_k. Jedan deo ove funkcije je opisan u prikazu 1.2,

insert into C_k
select p[1], p[2], … , p[k-1], q[k-1]
from $F_{k-1}p$, $F_{k-1}q$
where p[1] = q[1], … , p[k-2] = q[k-2], p[k-1] < q[k-1]

Prikaz 1.2

gde je $F_k p$ skup svih k-čestih skupova stavki p, a $p[k]$ oznaka za k-tu stavku skupa p. Nakon izvršavanja aktivnosti navedenih u prikazu 1.2, funkcija *apriori-gen* dodatno filtrira kolekciju C_k. To čini tako što eliminiše one kandidate za koje postoji $(k-1)$-podskup koji nije u kolekciji F_{k-1}.

Funkcija *podskup* za datu transakciju određuje skupove iz C_k koje ta transakcija sadrži. Pri tome, da bi se smanjio broj upoređivanja i time povećala efikasnost ove funkcije skupovi kolekcije C_k čuvaju se u posebnoj strukturi – *heš-stablu*[1].

[1] Heš-stablo podrazumeva heš funkciju i ograničenje za veličinu listova odnosno maksimalni broj elemenata/objekata koje može da sadrži list stabla.

Tokom svakog prolaza Apriori algoritma, nakon generisanja kolekcije C_k, konstruiše se heš-stablo koje u svojim listovima čuva sve skupove kandidate, tako da u okviru funkcije *podskup* umesto upoređivanja transakcije sa svakim kandidatom, ona je upoređuje samo sa kandidatima raspoređenim u odgovarajuće heš grupe.

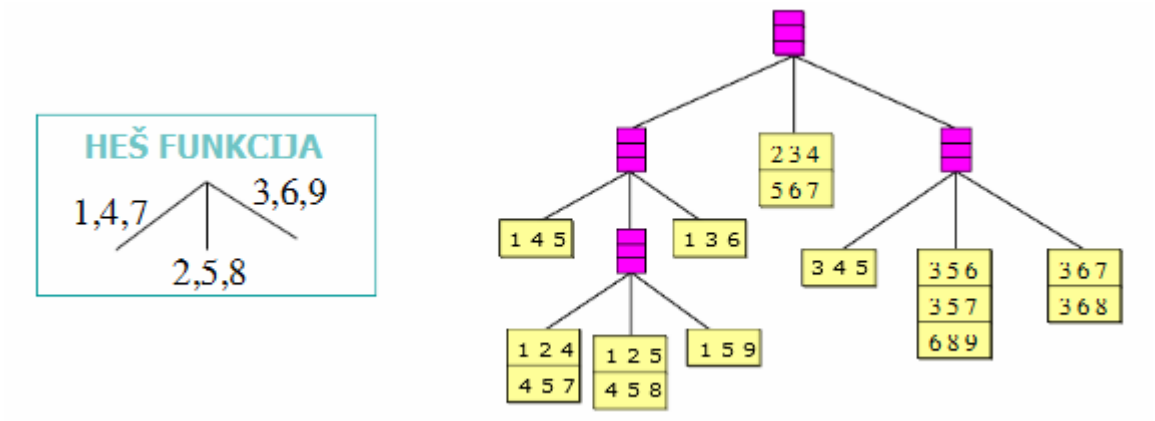
Naime, kreira se heš-stablo koje sadrži samo koren i k listova (koren stabla je na dubini $d = 1$). Zatim se skupovi kandidati $p \in C_k$ raspoređuju po listovima stabla, pri čemu je pripadnost svakog skupa p nekom listu određena vrednošću heš funkcije primenjene na d-tu stavku skupa p (d je dubina unutrašnjeg čvora stabla kod kojeg se skupovi raspoređuju). Ukoliko broj skupova kandidata u nekom listu prevazilazi zadato ograničenje, onda se taj list transformiše u unutrašnji čvor stabla, što podrazumeva da se na taj čvor nadovezuje novih k listova u koje se propagiraju njegovi skupovi kandidati.

Na ovako konstruisanom heš-stablu funkcija *podskup* locira kandidate koji se sadrže u datoj transakciji t. Počinje od korena drveta, gde svaku stavku transakcije t hešira i kreće se za jedan korak kroz drvo tamo gde je usmerava heš funkcija. Ako je dostignut list drveta onda se kandidati koji se nalaze u tom listu upoređuju sa transakcijom t i oni odgovarajući se dodaju na izlaz funkcije. Ukoliko je heširanjem stavke i dostignut unutrašnji čvor, rekurzivno se heširaju stavke koje dolaze nakon stavke i[2] sve dok se ne dostigne list drveta. Oni listovi koji ne budu posećeni zasigurno ne sadrže kandidate koji se sadrže u zadatoj transakciji t.

Primer - Konstrukcija (slika 1.2) i korišćenje (slika 1.3) jednog heš-stabla tokom jednog koraka Apriori algoritma. Neka je dat skup C_3 sa:

$$C_3 = \{\{1\ 4\ 5\}, \{1\ 2\ 4\}, \{4\ 5\ 7\}, \{1\ 2\ 5\}, \{4\ 5\ 8\}, \{1\ 5\ 9\}, \{1\ 3\ 6\}, \{2\ 3\ 4\},$$
$$\{5\ 6\ 7\}, \{3\ 4\ 5\}, \{3\ 5\ 6\}, \{3\ 5\ 7\}, \{6\ 8\ 9\}, \{3\ 6\ 7\}, \{3\ 6\ 8\}\}$$

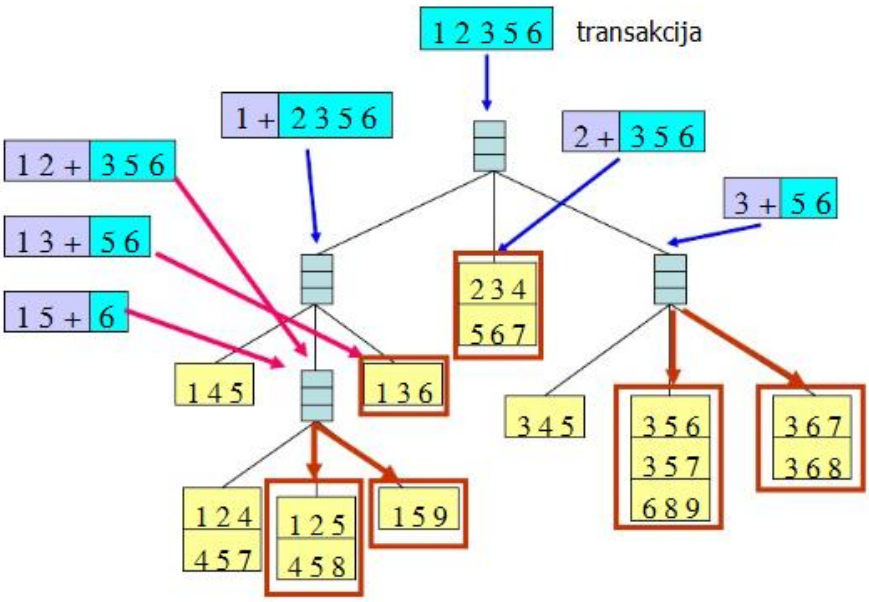
Za heš funkciju uzeta je funkcija ostatka po modulu k, tj: $h(x) = x\ mod\ k$. Na osnovu zadate heš funkcije konstruiše se heš-stablo (maksimalni broj elemenata smeštenih u jednom listu je 4). Na slici 1.2, ljubičastom bojom su označeni unutrašnji čvorovi stabla, dok su žutom markirani kandidati.



Slika 1.2

[2] U skladu sa leksikografskim poretkom

Na slici 1.3 se vidi kako funkcija *podskup* za zadatu transakciju $t = \{1\ 2\ 3\ 5\ 6\}$ pronalazi kandidate koji se u njoj sadrže. ■



Slika 1.3[3]

Nakon pronalaska svih čestih skupova stavki nad skupom transakcija D preostaje da se izgenerišu dobra odnosno jaka pravila pridruživanja. Za razliku od podrške, pouzdanost nema osobinu antimonotonosti, ali treba napomenuti njenu specifičnost kod pravila generisanih od istog čestog skupa f, koja je ključna za algoritam generisanja pravila pridruživanja.

Neka je a neprazan podskup čestog skupa f. Tada pravilo $a' \Rightarrow (f - a')$ ne može imati veću pouzdanost od pravila $a \Rightarrow (f - a)$, gde je $a' \subseteq a$. Dakle, da bi pravilo $(f - a) \Rightarrow a$ ispunjavalo uslov *minconf*, pravila $(f - a') \Rightarrow a'$ za svako $a' \subseteq a$ moraju da ispunjavaju uslov *minconf*.

Algoritam za generisanje pravila pridruživanja koristi upravo ovo svojstvo pouzdanosti. Naime, ovaj algoritam generiše pravila po grupama, gde prva grupa pravila sadrži samo jednu stavku u posledici, a svaka sledeća grupa se dobija spajanjem pravila iz prethodne tako da dobijena pravila imaju po jednu stavku više u posledici i jednu manje u premisi. Ukoliko neko pravilo ne zadovoljava uslov *minconf*, onda se za sva pravila koja bi se iz njega generisala zna da ne zadovoljavaju uslov *minconf*, shodno osobini pouzdanosti.

Pseudo-kod ovog algoritma je dat u prikazima 1.3 i 1.4.On u sebi koristi funkciju *apriori-gen* kao i informacije o podršci čestih skupova stavki izračunatih tokom njihovog pronalaženja.

[3] Slike korišćene u okviru primera ovog odeljka su preuzete sa http://www-users.cs.umn.edu/~kumar/dmbook/index.php

Algoritam – Generisanje pravila pridruživanja

ulaz: skup F – skup svih čestih skupova stavki nad podacima D,
minconf;

foreach (*čest* k*-skup* f_k, $k \geq 2$) **do begin**
  $H_1 = \emptyset$; // Inicijalizacija skupa posledica od po 1 stavke
  $A = \{a \mid a \subset f_k, |a| = (k-1)\}$;
  foreach $a_{k-1} \in A$ **do begin**
    conf = $s(f_k)/s(a_{k-1})$; // Podrška je već poznata
    if (conf ≥ *minconf*) **then begin**
      print a_{k-1} + „ ⇒ " + $(f_k - a_{k-1})$ +
        „; *pouzdanost =* " + conf +
        „; *podrška =* " + $s(f_k)$ + „;";
      $H_1 = H_1 + (f_k - a_{k-1})$; // Dodavanje skupu posledica
    end
  end
  call ap-genrules(f_k, H_1);
end

Prikaz 1.3

Procedure ap-genrules (f_k: čest k-skup, H_m: skup posledica od po m stavki)

  if ($k > m + 1$) **then begin**
    H_{m+1} = apriori-gen(H_m);
    foreach $h_{m+1} \in H_{m+1}$ **do begin**
      conf = $s(f_k)/s(f_k - h_{m+1})$; // Podrška je već poznata
      if (conf ≥ *minconf*) **then**
        print $(f_k - h_{m+1})$ + „ ⇒ " + h_{m+1} +
          „; *pouzdanost =* " + conf +
          „; *podrška =* " + $s(f_k)$ + „;";
      else
        $H_{m+1} = H_{m+1} - h_{m+1}$;
    end
    call ap-genrules(f_k, H_{m+1});
  end

Prikaz 1.4

Apriori, generalno, ima dobre performanse, međutim, u situacijama kada postoji veoma veliki broj čestih skupova, do čega može da dođe ukoliko je prag *minsup* dovoljno mali, stalno generisanje skupova kandidata i prolaženje kroz bazu transakcija postaje veoma skupo. U relativno bliskoj prošlosti, napravljeno je mnogo pokušaja da se optimizuje Apriori.

Neke od tehnika koje se implementiraju kako bi se Apriori poboljšao su: tehnika bazirana na heširanju (smanjuje broj skupova kandidata), particionisanje skupa transakcija (particioniše problem istraživanja na n manjih problema), uzorkovanje skupa transakcija itd.

1.2.4 Diskretizacija kontinualnih atributa

Prethodno opisani postupak pretrage pravila pridruživanja pretpostavlja da su ulazni podaci u formi skupa transakcija, međutim, često podaci na kojima želimo da naučimo pravila nisu tog oblika. Pre nego što se upustimo u dalju diskusiju o rešavanju ovog problema, potrebno je da, barem ukratko, razjasnimo pojam atributa i tipa atributa.

Definicija

Atribut je karakteristika ili osobina objekta koja može da varira od objekta do objekta ili zavisno od vremenskog faktora.[1]

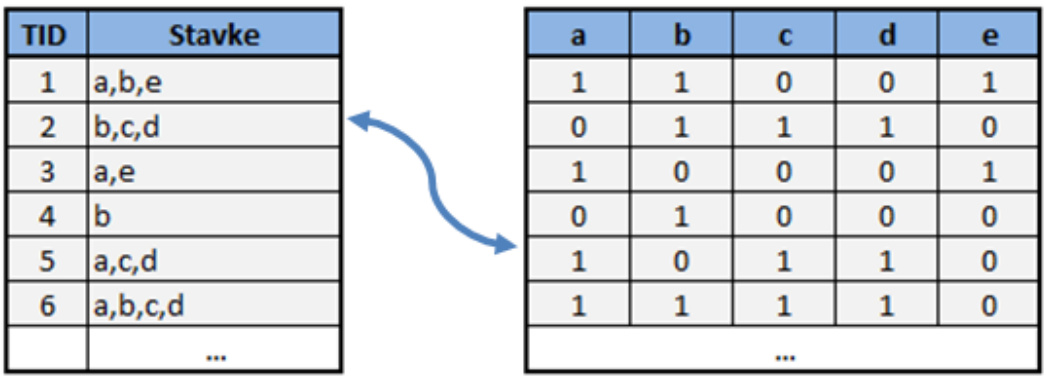
Na primer, za vođenje evidencije o zaposlenima, atributi jednog zaposlenog bi mogli da budu ID broj, godine radnog iskustva, trenutna pozicija, prosečna zarada u poslednjih 12 meseci, godine starosti itd. Očigledno je da su ovi atributi različiti, neki su numerički (celi ili realni brojevi), neki su tekstualni. Dakle, atributi mogu biti različitog tipa. Za početak pravi se razlika između **kategoričkih** i **numeričkih atributa**, odnosno, kvalitativnih i kvantitativnih. Kategorički atributi mogu imati numeričke ili tekstualne vrednosti, ali njihove vrednosti se pomatraju najviše kao simboli. Dve vrednosti kategoričkog atributa možemo međusobno porediti ($=, \neq$) ili u najboljem slučaju staviti ih u neki poredak ($<$, $>$), ali ne možemo nad njima da vršimo nikakve aritmetičke operacije. Sa druge strane, numerički atributi imaju numeričke vrednosti i uglavnom sve osobine brojeva. Vrednosti ovih atributa mogu biti celi ili realni brojevi.

Atribute takođe možemo razlikovati prema broju vrednosti koje oni mogu da imaju. Tu primećujemo dve osnovne grupe: **diskretne** i **kontinualne atribute**, gde je potpuno jasno šta koja grupa podrazumeva. Atributi čiji je domen vrednosti {0,1}, {tačno, netačno} ili {da, ne} i sl., nazivamo **binarnim atributima**, i oni su u okviru grupe diskretnih atributa. U nekim istraživanjima kod binarnih atributa je relevantno samo pojavljivanje jedne od ukupno dve vrednosti, tako da metod istraživanja potpuno ignoriše pojavu one druge vrednosti tog atributa. Takve binarne atribute nazivamo **asimetričnim binarnim atributima**.

Skup transakcija veoma lako može da se transformiše u skup podataka sa isključivo asimetričnim binarnim atributima, i obrnuto (slika 1.4). Naime, stavka se u nekoj transakciji ili pojavljuje ili ne, a tokom istraživanja pravila, prebrojavaju se samo pojavljivanja stavki.

Često postoji potreba da se istraže pravila pridruživanja iz podataka koji se sastoje od različitih tipova atributa, poput simetričnih binarnih, kategoričkih ili numeričkih itd. Da bismo istražili pravila, kategoričke atribute transformišemo u stavke, tj. za svaki različiti par (atribut, vrednost) kreiramo novu stavku. Na taj način dobijamo nešto poput desne tabele na slici 1.4, odakle lako dolazimo do skupa transakcija.

Sa druge strane, razvijene su razne tehnike koje se koriste u svrhu pripreme kontinualnih podataka za primenu analize pridruživanja – diskretizacijske, statističke i nediskretizacijske metode. Pravila pridruživanja koja u sebi sadrže kontinualne atribute nazivamo **kvantitativnim pravilima pridruživanja**.



TID	Stavke
1	a,b,e
2	b,c,d
3	a,e
4	b
5	a,c,d
6	a,b,c,d
	...

a	b	c	d	e
1	1	0	0	1
0	1	1	1	0
1	0	0	0	1
0	1	0	0	0
1	0	1	1	0
1	1	1	1	0
...				

Slika 1.4

Najčešće se koriste **diskretizacijske metode**. Ovim pristupom se vrednosti atributa grupišu u konačan broj intervala, tako da se od kontinualnog dobija kategorički atribut, iz čega lako dolazimo do asimetrično-binarnog i konačno željenog skupa transakcija.

Svaka diskretizacija podrazumeva:

- Sortiranje vrednosti atributa koji diskretizujemo;
- Određivanje $(n-1)$ podelaka koji dele skup vrednosti na n intervala;
- Mapiranje vrednosti jednog intervala u istu kategoričku vrednost.

Suštinski problemi diskretizacije su određivanje broja intervala, odnosno broja rezultujućih kategoričkih vrednosti i postavljanje podelaka na odgovarajuća mesta. Rezultat diskretizacije se može predstaviti skupom intervala $\{(x_0, x_1], (x_1, x_2], \dots, (x_{n-1}, x_n)\}$, gde su $x_0, x_1, \dots, x_n$ podeoci, a x_0 i x_n mogu da budu $+\infty$ ili $-\infty$.

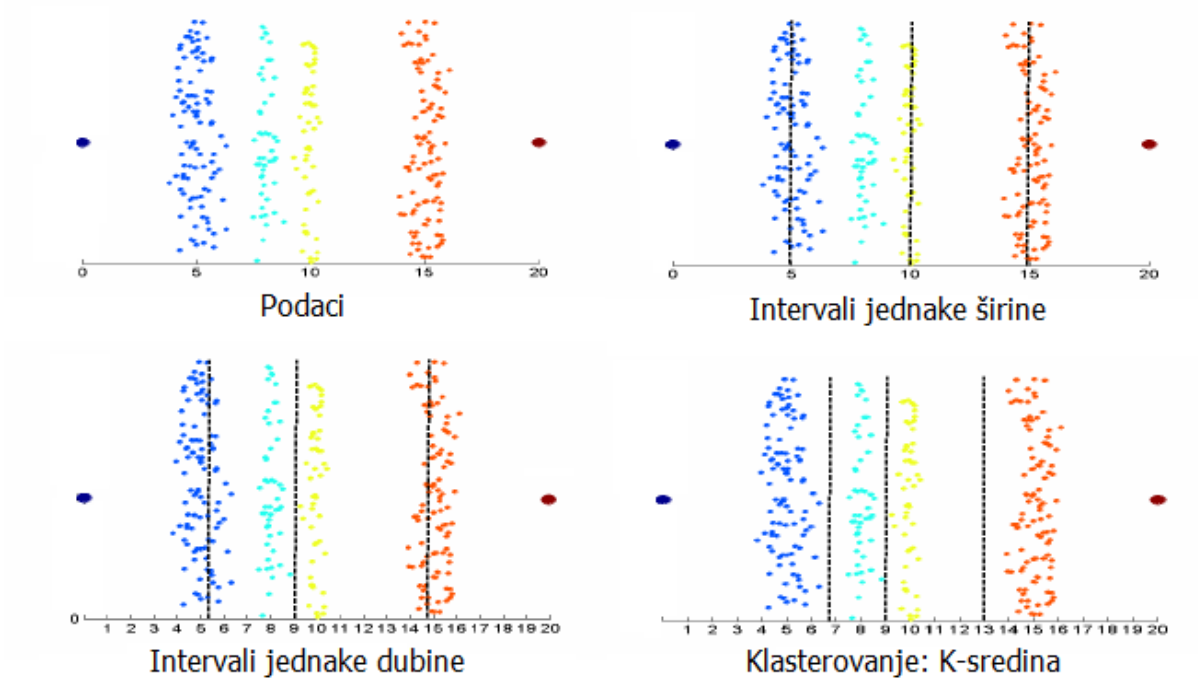
Neke od tehnika diskretizacije su:

- **Intervali jednake širine** – Postavlja podeljke na međusobno jednakoj udaljenosti, odnosno kreira željeni broj intervala iste širine. Ova metoda je izuzetno jednostavna, ali i veoma osetljiva na *elemente van granica*[4].
- **Intervali jednake dubine** – Za razliku od malopre opisane, ova metoda pokušava da grupiše vrednosti atributa u željeni broj intervala tako da svi intervali sadrže jednak broj vrednosti. Metoda je nezavisna od postojanja autlajera i često se koristi.
- **Klasterovanje** – Ova metoda podrazumeva primenu neke od tehnika klasterovanja nad vrednostima atributa koji želimo da diskretizujemo i da rezultujuće klastere tretiramo kao diskretizacijske grupe.

[4] Elementi van granica na dalje u tekstu – *autlajeri*. Autlajeri su anomalije. Po definiciji oni su neobične vrednosti u odnosu na većinu ostalih tipičnih vrednosti. Treba ih razlikovati od šuma, jer su autlajeri za razliku od šuma zapravo validne vrednosti.

Dobru diskretizaciju često možemo i sami da napravimo upoznavajući se sa podacima putem neke vizualizacije ili vršenjem raznih merenja itd.

Primer – Primena različitih tehnika diskretizacije nad jednim skupom vrednosti (slika 1.5). Potrebno je diskretizovati skup vrednosti na 4 grupe. Primenjujemo 3 tehnike: intervali jednake širine, intervali jednake dubine i klasterovanje *K*-sredina i posmatramo kako se koja ponaša nad istim skupom. Vrednosti su na slici predstavljene tačkama (primetiti da postoje dva autlajera – levo i desno), a pozicije podelaka su prikazane sa vertikalama.



Slika 1.5

Kao što se vidi na slici na datom skupu vrednosti veoma je lako izvršiti diskretizaciju prostim posmatranjem grafičke reprezentacije podataka, ali to nije automatizovan proces, te je cilj dostići isti rezultat nekom od navedenih tehnika. Od primenjenih, u ovom slučaju najbolje je klasterovanje, pa zatim intervali jednake dubine i na kraju intervali jednake širine. ■

Pored pomenutih tehnika, postoji još jedna podvrsta diskretizacija, tzv. **kontrolisane diskretizacije** (eng. *Supervised Discretization*) koje uglavnom daju bolje rezultate, a prilikom rada koriste dodatne informacije (obično informaciji o klasi). One postavljaju podeljke tako da rezultujući intervali zadovoljavaju unapred zadati prag čistoće i minimalnu veličinu intervala. Čistoća intervala se obično izražava entropijom[5] pri čemu se koristi informacija o klasi. Princip particionisanja kontinualnih atributa zasnovan je na deljenju skupa vrednosti na dva dela, tako da rezultujući intervali imaju minimalnu entropiju. Svaka vrednost atributa je potencijalni podelak. Particionisanje se rekurzivno primenjuje na

[5] Uglavnom se koristi Šenonova (eng. *Claude Shannon*, april 1916 – februar 2001) entropija, koja meri nepredvidljivost ili neuređenost u podacima. Konkretno za diskretizaciju, najčistiji intervali imaju entropiju 0 – sve vrednosti grupisane u taj interval su iste klase; kada je entropija maksimalna, intervali su najnehomogeniji – klase u tom intervalu su jednako distribuirane.

intervalu sa najvećom entropojim i taj proces se ponavlja sve dok se ne dostigne željeni broj intervala ili dok se ne zadovolji neki kriterijum zaustavljanja.

Odabir broja intervala ili broja klastera može značajno da utiče na otkrivanje pravila pridruživanja iz diskretizovanog skupa podataka. Ukoliko su intervali previše široki onda je moguće da u rezultatu nećemo dobiti neka zanimljiva pravila zbog neispunjavanja uslova pouzdanosti. Slično, ukoliko su intervali previše uski, onda možda izgubimo zanimljiva pravila zbog neispunjavanja uslova podrške.

1.3 Molekularna biologija – Osnovni pojmovi

Da bismo se bavili bioinformatičkim istraživanjem neophodno je da objasnimo neke osnovne pojmove molekularne biologije, te sledi tekst koji ih uvodi na neformalan način.

Ćelije su osnovni gradivni elementi živih bića. Svaka ćelija sadrži jedro, mitohondrije, ribozom, endoplazmatični retikulum, vakuolu, citoplazmu, itd. Od toga, jedro se izdvaja kao jako važna organela, jer u sebi sadrži hromozome koji uključuju DNK. Neophodne informacije za sintetisanje proteina – molekuli, koji obavljaju gotove sve poslove unutar ćelije – zapisane su upravo u genima unutar DNK. Naučnici bi želeli da što bolje razumeju biologiju živih bića i da to znanje primene na lečenje bolesti.

1.3.1 DNK

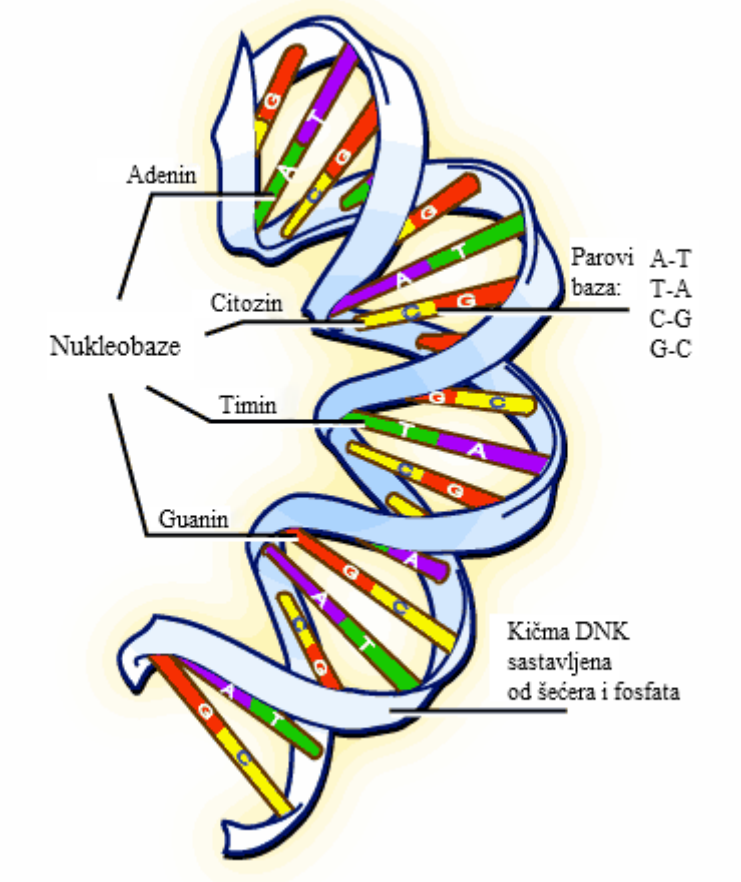
Nukleinske kiseline su lanci manjih molekula – *nukleotida*. **Nukleotid** se satoji od šećera, fosfata i baze (nukleobaze). Nukleinsku kiselinu kraće zapisujemo kao sekvencu baza nukleotida. Postoje dva različita tipa nukleinskih kiselina: *dezoksiribonukleinska kiselina* – **DNK** i *ribonukleinska kiselina* – **RNK**. Prefiksi *dezoksiribo* i *ribo* se redom odnose na šećere dezoksiribozu i ribozu koja se nalazi u sastavu nukleotida.

U prirodi postoji samo 5 tipova nukleobaza: *adenin* (A), *citozin* (C), *guanin* (G), *timin* (T) i *uracil* (U). U DNK se mogu naći nukleotidi[6] A,C,G i T, dok u RNK umesto T nalazimo U. Tokom 1950. godine na *Univerzitetu Kolumbija* pronađeno je da se u DNK jednako frekventno pojavljuje adenin i timin, kao i par citozin i guanin. Ovo zapažanje je kasnije postalo poznato kao Čargafovo[7] pravilo.

Najpoznatije naučno otkriće o strukturi DNK napravljeno je 1953. godine. Naučnici Frensis Krik (eng. *Francis Crick*, jun 1916 – jul 2004) i Džejms Votson (eng. *James Watson*, april 1928) zaključili su da je DNK sastavljena od dva polinukleotidna lanca koji su spiralno uvijeni jedan oko drugog, odnosno formira se tzv. **dvostruka heliks struktura** (slika 1.6). Baze duž ova dva lanca se međusobno uparuju (hibridizacija) i to prema pravilima vezivanja – adenin se vezuje sa timinom i citozin sa guaninom. Kaže se da su ovi parovi baza *komplementarni*, a dva polinukleotidna lanca su međusobno *antiparalelna*. Na osnovu sekvence baza na jednom lancu lako zaključujemo odgovorajuću sekvencu na drugom lancu. Ova dvostruka struktura je ključna, jer DNK molekulu omogućava jako jednostavan mehanizam za samoreprodukciju. Molekul RNK za razliku od DNK, obično formira jednostruku strukturu.

[6] Nukleotide označavamo simbolima za baze koje u sebi sadrže.

[7] Ervin Čargaf (eng. *Erwin Chargaff,* avgust 1905 – jun 2002) – austrijski biohemičar, najpoznatiji po svojim istraživanjima DNK. Pokazao je da je u prirodnoj DNK broj jedinica guanina jednak broju jedinica citozina, kao i za timin i adenin. Godine 1952. sa svojim nalazima je upoznao dvojicu kembridžiskih naučnika, što je pomoglo da se 1953. otkriju veoma značajne činjenice.



Slika 1.6[8]

1.3.2 Proteini

Proteini su veoma važni molekuli u organizmu. Neki primeri proteina su: antitela, receptori, enzimi, neurotransmiteri, neki hormoni itd.

Proteini su lanci manjih molekularnih entiteta – *aminokiselina*. One se sastoje od centralnog ugljenikovog atoma, amino i karboksilne grupe i posebnog bočnog lanca specifičnog za svaku aminokiselinu ponaosob. Vezivanjem karboksilne grupe jedne aminokiseline i amino grupe druge aminokiseline nastaje *peptidna veza* i na taj način se formiraju lanci aminokiselina, odnosno *polipeptidni lanci*. Proteine posmatramo kao jedan polipeptidni lanac[9] i kraće ih zapisujemo kao sekvencu aminokiselina. U prirodi postoji puno poznatih aminokiselina, ali samo 20 koje formiraju proteine – pobrojane su u tabeli 1.1.

Struktura proteina je mnogo komplikovanija od DNK molekula, i definisana je sledećom hijerarhijom.

[8] Originalna slika je objavljena na http://www.scq.ubc.ca/a-monks-flourishing-garden-the-basics-of-molecular-biology-explained/

[9] Neki proteini se mogu sastojati i od nekoliko polipeptidnih lanaca, ali s obzirom da to nije značajno za temu ovog rada, ovde je izvršena generalizacija.

Ime	eng.	Troslovna oznaka	Jednoslovna oznaka
Alanin	*Alanine*	Ala	A
Arginin	*Arginine*	Arg	R
Asparagin	*Asparagine*	Asn	N
Asparaginska kiselina	*Aspartic acid*	Asp	D
Cistein	*Cysteine*	Cys	C
Glutaminska kiselina	*Glutamic acid*	Glu	E
Glutamin	*Glutamine*	Gln	Q
Glicin	*Glycine*	Gly	G
Histidin	*Histadine*	His	H
Izoleucin	*Isoleucine*	Ile	I
Leucin	*Leucine*	Leu	L
Lizin	*Lysine*	Lys	K
Metionin	*Methionine*	Met	M
Fenilalanin	*Phenylalanine*	Phe	F
Prolin	*Proline*	Pro	P
Serin	*Serine*	Ser	S
Treonin	*Threonine*	Thr	T
Triptofan	*Tryptophan*	Trp	W
Tirozin	*Tyrosine*	Tyr	Y
Valin	*Valine*	Val	V

Tabela 1.1

Definicija

Razlikuju se četiri nivoa strukture proteina:

Primarna struktura – Primarna struktura je sekvenca aminokiselina duž polipeptidnog lanca.

Sekundarna struktura – Sekundarna struktura opisuje interakcije između atoma na kičmi polipeptidnog lanca koje formiraju podstrukture kao što su α-heliksi, β-trake i petlje.

Tercijarna struktura – Tercijarna struktura se odnosi na prostorni raspored svih atoma polipeptidnog lanca. Elementi sekundarne strukture se grupišu u motive i u funkcionalne jedinice koje se nazivaju domeni.

Kvartarna struktura – Kvartarna struktura opisuje spoj polipeptidnih podjedinica ili potencijalno drugih molekula u proteinu.[5]

Uloga ili **funkcija proteina** je usko povezana sa njegovom prostornom strukturom, dok se postorna, odnosno, trodimenzionalna struktura povezuje sa sekvencom aminokiselina u proteinu. Samo predviđanje strukture proteina na osnovu sekvence aminokiselina nije u potpunosti rešen problem i jedan je od najbitnijih zadataka bioinformatike.

1.3.3 Sintetisanje proteina

Još početkom 20. veka proučavanjem bolesti uzrokovane ćelijama u kojima se ne izvršavaju određene biohemijske reakcije, npr. zbog nepostojanja nekog enzima, a koje su u isto vreme i nasledne, uočena je veza između gena i proteina. Do 1950. godine naučnici su čvrsto verovali da je veza između određene sekvence nukleotida u DNK molekulu i sekvence aminokiselina u proteinu linearna, ali im je bilo nepoznato kako raspored nukleotida utiče na proizvodnju proteina. Kasnije je eksperimentalno pokazano da trojke nukleotida – **kodoni –** kodiraju jednu aminokiselinu u polipeptidnom lancu. Npr. kodon od tri *uracila* (UUU) kodira jedan *fenilalanin* (F) u sekvenci aminokiselina. Ubrzo zatim su otkriveni i ostali kodoni (postoji $4^3 = 64$ kodona, a samo 20 aminokiselina, te jednu aminokiselinu može da kodira i više različitih kodona). Skup parova kodona i aminokiselina je **genetički kod**. Treba imati u vidu da postoji nekoliko genetičkih kodova zavisno od porekla DNK molekula.

Dakle, proteinske sekvence aminokiselina su kodirane određenim sekvencama nukleotida u DNK molekulima. Te sekvence nukleotida u DNK nazivamo **genima**. Svaki gen počinje nekim standardnim *start kodonom* i završava se nekim standardnim *stop kodonom*. Jedan gen sadrži informacije za proizvodnju jednog proteina. Mehanizam transformisanja informacija sadržanih unutar gena u sekvence aminokiselina, odnosno proteine, veoma je kompleksan i sastoji se iz dve faze (slika 1.7). Ovde se taj mehanizam samo površno opisuje.

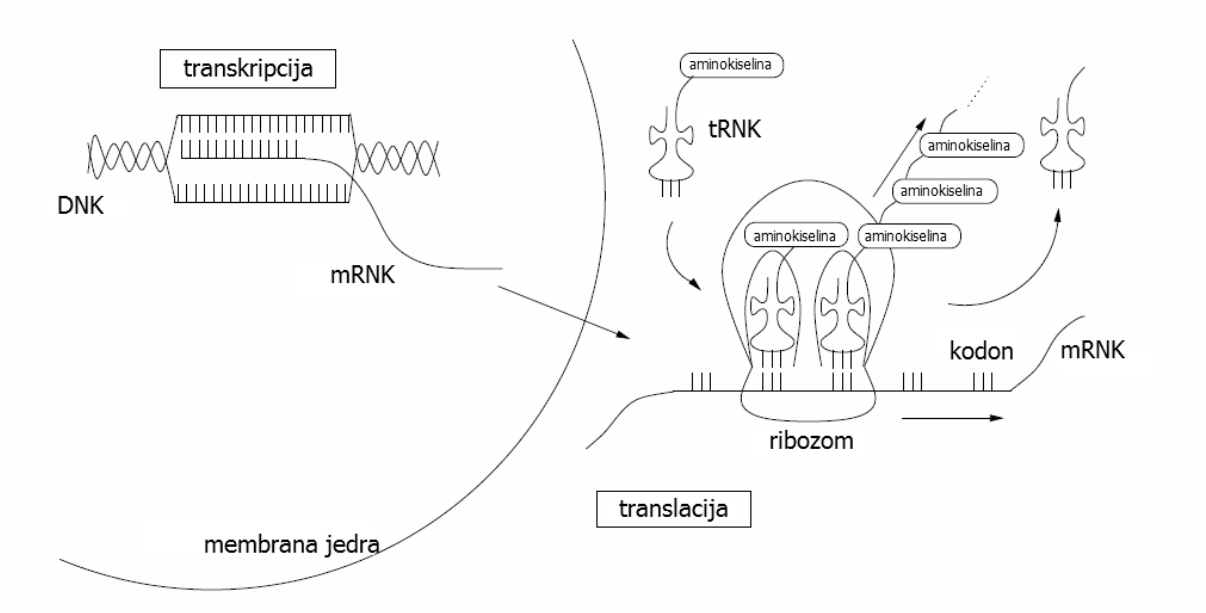
Prvo se neophodne informacije sadržane u datom genu preslikavaju u sekvencu nukleotida koja može da napusti jedro ćelije – proces poznat pod nazivom **transkripcija**. Ovde se dva polinukleotidna lanca DNK odvijaju i razdvajaju jedan od drugog i to samo u regionu gde se nalazi konkretan gen, zatim se kreira RNK kopija gena i to sastavljena samo od nukleotida koji sudeluju u kodiranju aminokiselina[10]. Ova kopija se označava sa *mRNK* (eng. *messenger RNA*). mRNK napušta jedro ćelije prenoseći neophodne informacije za proizvodnju proteina u ribozom[11] i time započinje drugu fazu.

U ribozomu se informacije sadržane u mRNK transformišu u sekvencu aminokiselina – proces poznat pod nazivom **translacija**. Kodoni iz mRNK se jedan po jedan pomeraju kroz ribozom. Za svaki kodon se kreira tRNK (eng. *transfer RNA*) molekul koji sadrži odgovarajuću aminokiselinu i tri nukleotida komplementarna trenutnom kodonu iz mRNK – ta grupa komplementarnih nukleotida se naziva *antikodon*.

tRNK sa antikodonom se vezuje za trenutni kodon iz mRNK i pritom oslobađa aminokiselinu koja se vezuje za lanac aminokiselina koji je do tada konstruisan. Ovaj proces se nastavlja sve dok antikodon u tRNK ne bude STOP kodon.

[10] U okviru gena eukariota (organizama čije ćelije sadrže jedro) postoje manje sekvence nukleotida koje ne kodiraju aminokiseline, nazivaju se *introni*, a preostale grupe nukleotida nazivaju se *egzoni*.

[11] Ribozomi su ćelijske organele koje sintetišu proteine. Oni spajaju aminokiseline u rasporedu koji diktira sekvenca nukleotida u mRNK, čime se dobija određeni protein.



Slika 1.7 [5]

1.3.4 Neuređeni proteini

Centralno uverenje strukturalne biologije - da je stabilna prostorna struktura proteina neophodna da bi ispravno izvršavao svoju biološku funkciju kao i da struktura definiše funkciju proteina, ne može se primeniti na sve proteine. Prostorna struktura proteina i njegova biološka funkcija su nesumnjivo usko povezane, ali rastući skup eksperimentalno sakupljenih podataka pokazuje da postoji veliki broj funkcionalnih proteina koji ne zauzimaju jedinstvenu, uravnoteženu tercijarnu strukturu, već strukturu u kojoj se pozicije atoma i sam polipeptidni lanac vremenom menjaju. Za takve proteine kažemo da su **suštinski nestruktuirani** (eng. *intrinsically unstructured proteins*) ili **neuređeni** (eng. *disordered*). Svaki protein koji u sebi sadrži bar jedan neuređen region nazivamo neuređenim proteinom.[8]

2 POSTAVKA PROBLEMA, PODACI I METODE

Istraživanjima u protekloj deceniji potvrđeno je da regulatorni proteini, zatim oni koji su umešani u procese signalizacije i molekulskog prepoznavanja, češće sadrže neuređene regione. Funkcionisanje nekih neuređenih proteina, koji obavljaju upravo ove biološke procese, povezuje se sa bolestima poput raka, dijabetesa, kardiovaskularnim i neurodegenerativnim bolestima. Pronalažanje detaljnih objašnjenja na koji način neuređeni proteini funkcionišu, kao i shvatanje same prirode neuređenosti je, takoreći, preduslov iliti osnova za konstruisanje izglednih terapija i lekova.

Ovo poglavlje posvećeno je postavkama problema bioinformatičke karakterizacije neuređenosti aminokiselina njihovim fizičkohemijskim svojstvima, zasnovane na metodama istraživanja podataka – motivaciji, formulaciji zadatka i dosadašnjim rezultatima, zatim izvorima podataka i metodama izračunavanja.

2.1 Formulacija zadatka i dosadašnji rezultati

Primarna struktura u potpunosti određuje prostornu strukturu, pa i njenu nestabilnost, odnosno neuređenost. Male izmene u rasporedu aminokiselina mogu da izazovu velike posledice. Naime, variranje samo jednog egzonskog nukleotida može da prouzrokuje promenu u strukturi odgovarajućeg proteina, pa i da promeni tendenciju nekog regiona da bude stabilno struktuiran ili neuređen (sve to utiče na funkcionalnost samog proteina i konačno na rizik od oboljevanja organizma). Nije poznato kako lanac aminokiselina zauzima konkretnu prostornu strukturu u ćeliji. Može se postaviti pitanje: „Šta kod lanaca aminokiselina dovodi do pojave neuređenih regiona?“. Ovaj rad je motivisan tim pitanjem, a cilj je da se metodama istraživanja podataka potvrde i potencijalno otkriju nove zavisnosti između karakteristika aminokiselina i njihove tendencije da se pojavljuju u neuređenim regionima. Rezultati bi mogli da doprinesu razumevanju sastava i uopšte osobina neuređenih regiona u proteinima.

Zadatak: Pronaći korelacije i izdvojiti pravilnosti između koeficijenta neuređenosti aminokiselina i njihovih fizičkohemijskih i biohemijskih svojstava, takozvanih aminokiselinskih indeksa, ili *AAindeksa*. Istraživanje sprovesti nad skupom proteina datih u *DisProt*[12] bazi i vrednostima indeksa aminokiselina datih u bazi *AAindex*[13]. Za pronalaženje pravilnosti koristiti tehniku istraživanja podataka – *analizu pridruživanja*.

Do sada je izvedeno dosta bioinformatičkih istraživanja koja su neposredno ili barem posredno za cilj imala da pronađu veze između osobina aminokiselina i neuređenih proteina. Zna se da u odnosu na uređene sekvence, neuređene karakteriše manje aromatična struktura, veća naelektrisanja, veća hidropatija i još neka svojstva.

Neophodno je izdvojiti rad *Jovane Kovačević* [7] u okviru kojeg su izvršeni detaljni proračuni nad proteinima smeštenim u *DisProt* bazi. Izračunat je koeficijent neuređenosti (zasebno za N i C terminale i središnji deo proteinskih sekvenci) svake aminokiseline ponaosob i upoređen je sa njihovim osobinama – *hidropatijom* i *masom*. Takođe, formirana je nova skala aminokiselina prema njihovom koeficijentu neuređenosti u proteinima datim u DisProt bazi. Uočena je visoka korelacija između koeficijenta neuređenosti i hidrofobnosti aminokiselina, što potvrđuje rezultate poznate od ranije (dobijene nad drugim skupom proteina).

Postavlja se pitanje: Da li postoji još neki *AAindeks*, poput hidropatije, koji je povezan sa skalom neuređenosti?. Pronalaženje odgovora na to pitanje ključni je zadatak ovog istraživanja.

[12] http://www.disprot.org - Baza neuređenih proteina. Neuređeni regioni proteina koji su smešteni u DisProt bazu su eksperimentalno otkriveni.

[13] http://www.genome.jp/aaindex - Baza numeričkih indeksa koji predstavljaju fizičkohemijska i biološka svojstva aminokiselina i parova aminokiselina.

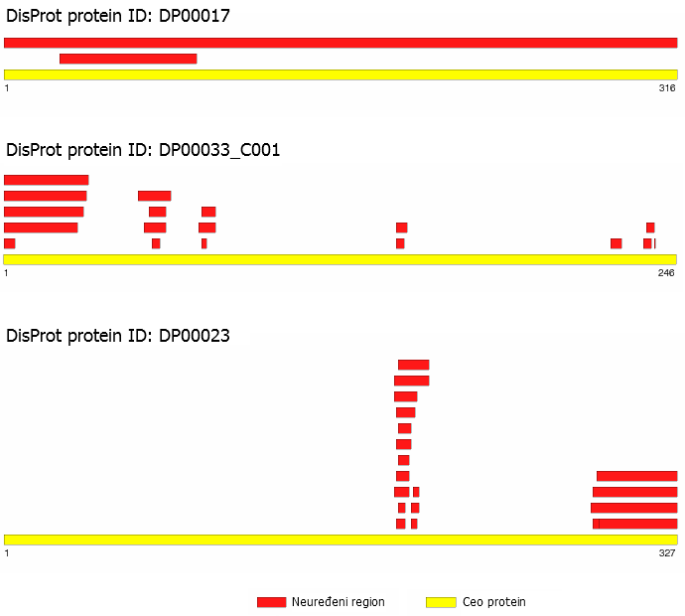
2.2 Izvori podataka

Kao što je navedeno u formulaciji zadatka, skup podataka koji se istražuje treba preuzeti iz dve onlajn baze podataka: *Disprot* i *AAindex* (podaci ovih baza javno su dostupni u tekstualnom formatu). Za potrebe ovog rada napravljen je modularan i lako proširiv program koji preuzima sve dostupne podatke o proteinima i indeksima aminokiselina i smešta ih u za to posebno dizajniranu bazu podataka. Pre prikaza samog istraživanja u ovom odeljku date su neophodne informacije o oba izvora.

2.2.1 DisProt

DisProt je rastuća baza podataka koja čuva informacije o proteinima koji nemaju stabilnu prostornu strukturu, a njihova neuređenost se utvrđuje eksperimentalnim putem. Ova baza je formirana da bi olakšala istraživanje neuređenosti organizovanjem i sakupljanjem informacija o eksperimentalnim karakteristikama i funkcionalnostima neuređenih proteina. Svi podaci su javno dostupni na http://www.disprot.org.

Izdanje baze (verzija 5.8) koje je korišćeno u ovom radu[14], sadrži 644 proteina, 1378 neuređenih i 70 uređenih regiona. Unos svakog proteina sadrži jedinstveni identifikacioni kod, sekvencu celog proteina, početnu i završnu lokaciju njegovih regiona sa naznakom da li je taj region uređen ili neuređen, i to je ono što je najpotrebnije za ovo istraživanje[15]. Datoteka sa podacima o proteinima se može preuzeti u XML ili FASTA formatu pojedinačno po proteinu ili za sve zajedno.



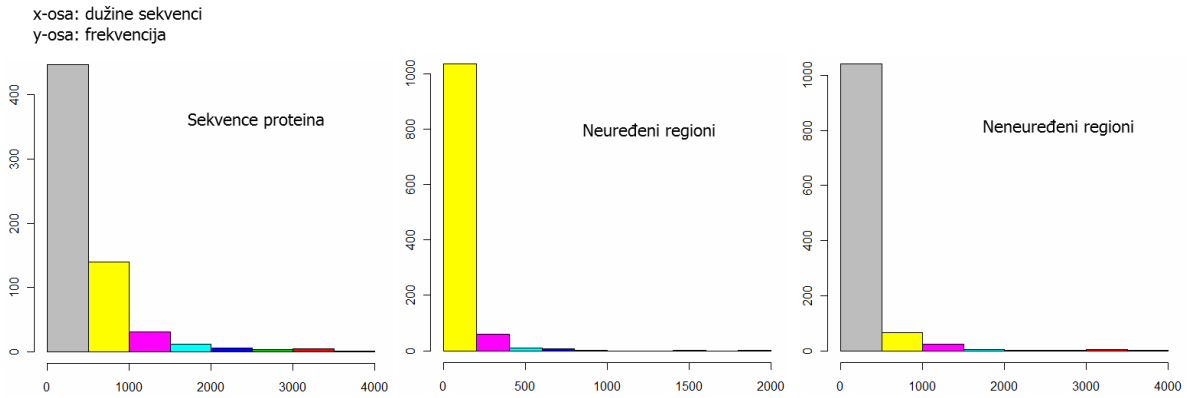
Prikaz 2.1

Kod nekih proteina neuređeni regioni se preklapaju (prikaz 2.1). Tada se u razmatranje uzima **unija** tih regiona. Postoje i preklapanja uređenih regiona (veoma retko), pa se i tu primenjuje navedeni princip, s tim da se prednost uvek daje uniranju neuređenih regiona. Takođe, u *DisProt* unosima postoje delovi sekvence proteina koji nisu svrstani niti u uređene niti u neuređene regione, te se uzima da je njihova struktura **nepoznata**.

[14] Nadalje u tekstu svi proračuni se odnose na podatke koji su bili dostupni u 5.8 verziji *DisProt* baze.
[15] Detaljni opis *DisProt* formata unosa o proteinu dat je na adresi http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1751543/.

Kako su za ovaj rad najvažniji neuređeni regioni, uspostavlja se sledeće pravilo: *Jedna aminokiselina u okviru sekvence proteina ekskluzivno pripada neuređenom ili neneuređenom regionu*. Pri tome, **neneuređen** region u okviru jednog proteina je unija uzastopnih uređenih regiona i onih delova sekvence za koje je nepoznata struktura. Uniranje neuređenih regiona uvek ima prednost nad neneuređenim regionima.

U skladu sa ovim pravilima sekvence proteina iz *DisProt* baze su u potpunosti izdeljene na 2262 regiona, od čega 1145 neneuređenih i 1117 neuređenih. Jedan protein u proseku sadrži 3.5 regiona. Na prikazu 2.2 dati su histogrami dužina sekvenci celih proteina, neuređenih i neneuređenih regiona[16]. Primećuje se da su neneuređeni regioni uglavnom duži od neuređenih, što je kasnije utvrđeno prebrojavanjem, dužina svih neuređenih regiona je 73688, a neneuređenih 241904.



Prikaz 2.2

2.2.2 AAindex

AAindex je baza numeričkih indeksa raznih fizičkohemijskih i biohemijskih osobina aminokiselina i parova aminokiselina izmerenih eksperimentalnim putem. Sadrži tri odvojene sekcije:

- *AAindex1* – skup indeksa aminokiselina, gde jedan indeks podrazumeva kolekciju od 20 numeričkih vrednosti (za svaku aminokiselinu po jednu vrednost).
- *AAindex2* – skup mutacionih matrica aminokiselina.
- *AAindex3* – skup kontanktnih matrica aminokiselina.

U ovom radu koriste se podaci iz prve sekcije, tj. *AAindex1*. Podaci su dati u tekstualnom obliku (eng. *flat file*), po jedna datoteka za svaku sekciju. Svi podaci su javno dostupni na http://www.genome.jp/aaindex/. U okviru sekcije *AAindex1* date su vrednosti za 544 različitih indeksa (decembar 2011). Svaki unos jednog indeksa sadrži njegov jedinstveni identifikacioni kod, kratak opis, spisak referenci na kojima se detaljno može pročitati o konkretnom eksperimentu i samom indeksu i 20 numeričkih vrednosti (prikaz 2.3).Takođe, svaki indeks sadrži spisak indeksa sa kojima je visoko korelisan (apsolutna korelacija veća od 0.8).

[16] Na histogramima je zbog preglednosti ignorisan jedan protein, dužina sekvence tog proteina iznosi 18534, a sadrži 3 neuređena i 4 neneuređena regiona sa dužinama redom {250, 3886, 30} i {3545, 831, 813, 9179}.

Svaki unos je sledećeg formata:

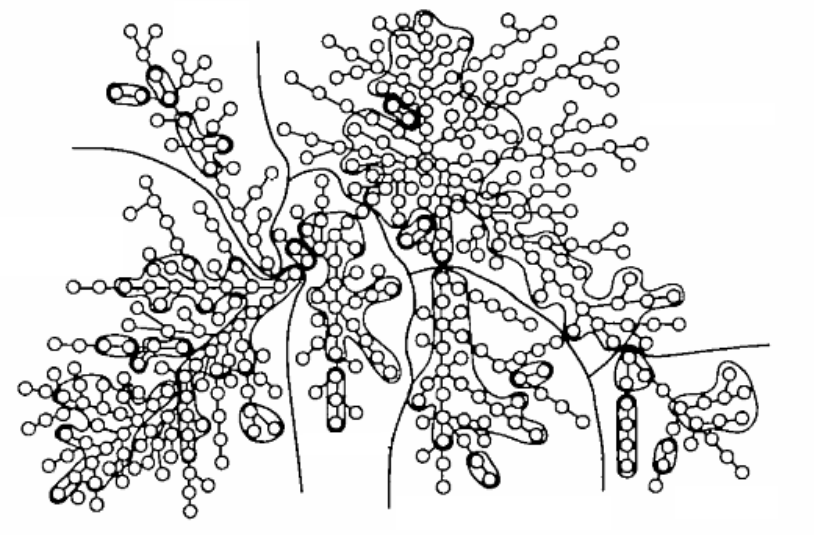
H	Identifikacioni kod
D	Opis indeksa
R	Spisak referenci u LITDB formatu
A	Autori
T	Naslov članka
J	Referenca na časopis koji sadrži članak
*	Komentari (iako rezervisano mesto, nigde se ne pojavljuje)
C	Spisak identifikacionih brojeva indeksa koji su slični aktuelnom indeksu
I	Numeričke vrednosti indeksa. Uvek su date prema sledećem redosledu

Ala Arg Asn Asp Cys Gln Glu Gly His Ile
Leu Lys Met Phe Pro Ser Thr Trp Tyr Val

Prikaz 2.3

Spisak svih identifikacionih brojeva i kratkih opisa se mogu pronaći na http://www.genome.jp/aaindex/AAindex/list_of_indices.

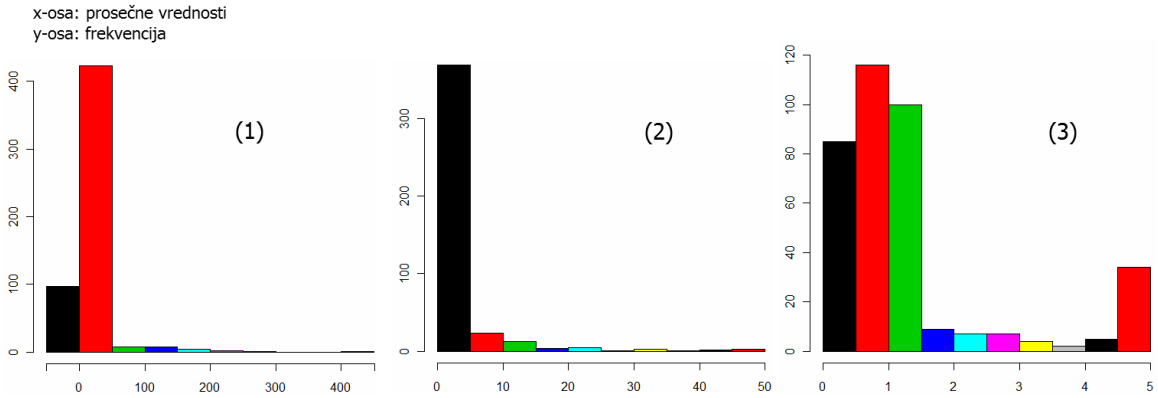
Da bi se vizualizovale veze između indeksa, u [9] je konstruisano minimalno-razgranato drvo (slika 2.1 – slika je informativnog tipa, a trenutno aktuelno drvo je dato u [10] i dostupno na http://www.genome.jp/aaindex/AAindex/Figure1.jpg). Drvo sadrži 402 indeksa[17], gde je svaki indeks jedan objekat predstavljen kružićem. Objekti su međusobno povezani koristeći *single-link* hijerarhijsko klasterovanje (mera rastojanja je korelacija). Dodatno su zaokružene oblasti unutar kojih su objekti na rastojanju najviše 0.1, odnosno apsolutna korelacija između bilo kojeg para objekata u toj oblasti je veća od 0.9. Drvo je zbog preglednosti podeljeno na šest regija[18].



Slika 2.1

[17] Kada je drvo kreirano, 1996. godine, u bazi je bilo dostupno 402 indeksa.
[18] Dakle dva indeksa iz iste regije ne moraju da budu u visokoj korelativnoj vezi.

Vrednosti indeksa su date kao realne vrednosti. Opseg vrednosti kao i distribucija vrednosti u nekom opsegu je različita od indeksa do indeksa, što dodatno otežava automatizaciju u istraživanju. Na prikazu 2.4 dati su histogrami prosečnih vrednosti indeksa. Na prvom histogramu su prisutni svi indeksi, dok su na drugom i trećem u obzir uzeti samo oni čija prosečna vrednost pripada intervalu redom (0,50) i (0,5). Prvi i drugi histogram ukazuju da većina indeksa, oko 330 od ukupno 544, ima prosečnu vrednost u okviru intervala (0,5). Situacija je dosta drugačija na trećem histogramu – više korpi je popunjeno te ne možemo da izdvojimo neku korpu gde se nalazi većina (prosečnih) vrednosti. Standardna devijacija više od 400 indeksa nije veća od 2.



Prikaz 2.4

2.3 Metode

Do sada je pominjan termin *koeficijent neuređenosti* aminokiseline kao i *korelacija*, a da to nije i precizno definisano. Ovaj odeljak je upravo tome posvećen, ali će uz to biti prezentovani i rezultati nekih merenja nad podacima.

2.3.1 Koeficijent neuređenosti

Koeficijent neuređenosti jedne aminokiseline je numerička vrednost koja treba da izrazi njenu tendenciju da se pojavi u neuređenom regionu proteina. Može se računati na razne načine. U ovom slučaju, akcenat se stavlja na odnos frekventnosti aminokiseline u neuređenim i neneuređenim regionima proteina.

Neka i jedinstveno određuje sekvencu aminokiselina (i-ta sekvenca) u nekom skupu sekvenci X i neka j jedinstveno određuje aminokiselinu (j-ta aminokiselina), $1 \leq j \leq 20$. Sa F_{ji} označavamo relativnu frekvenciju j-te aminokiseline u i-toj sekvenci.

Frekvenciju j-te aminokiseline u skupu sekvenci X računamo na sledeći način (u [7] ova frekvencija se naziva eng. *Mole Fraction*):

$$F_j(X) = \frac{\sum_i n_i * F_{ji}}{\sum_i n_i}$$

gde je n_i dužina i-te sekvence.

Neka je A skup sekvenci neuređenih regiona, a B skup sekvenci neneuređenih regiona. Sada, **koeficijent neuređenosti** j-te aminokiseline računamo na sledeći način (u [7] ovaj odnos frekvencija se naziva eng. *Fractional difference*):

$$FD_j = \frac{F_j(A) - F_j(B)}{F_j(B)}$$

Za $FD_j > 0$ j-ta aminokiselina ima veću frekvenciju u neuređenim nego u neneuređenim regionima celog skupa proteina, i obrnuto za $FD_j < 0$. Takođe, kroz ovaj koeficijent se može posmatrati i jačina tendencije da se aminokiselina pojavljuje ili ne u neuređenim regionima. Naime, za veće vrednosti koeficijenta FD_j veća je i neuređenost j-te aminokiseline, odnosno raste njena tendencija da se pojavljuje u neuređenim regionima.

Nad modifikovanim[19] podacima izvučenim iz $DisProt$ baze, izvršena su potrebna merenja da bi se izračunao koeficijent neuređenosti za svaku aminokiselinu – rezultati tih merenja dati su u tabeli 2.1 (C – broj pojavljivanja nad svim sekvencama; C(A) – broj pojavljivanja nad sekvencama neuređenih regiona).

Proračuni su vršeni nad svim proteinima i sortirani su prema koeficijentu neuređenosti.

[19] Shodno pravilima o regionima, modifikovani su neuređeni i kreirani neneuređeni regioni.

AA	C	C(A)	F(A)	F(B)	FD
C	4598	586	0.00795	0.01659	$\mathbf{-0.521}$
W	3206	468	0.00635	0.01132	$\mathbf{-0.439}$
I	14319	2292	0.0311	0.04972	$\mathbf{-0.374}$
F	9976	1721	0.02335	0.03412	$\mathbf{-0.316}$
Y	8330	1488	0.02019	0.02828	$\mathbf{-0.286}$
L	26225	4698	0.06375	0.08899	$\mathbf{-0.284}$
M	6942	1322	0.01794	0.02323	$\mathbf{-0.228}$
V	19660	3967	0.05383	0.06487	$\mathbf{-0.170}$
N	12919	2665	0.03616	0.04239	$\mathbf{-0.147}$
R	16036	3398	0.04611	0.05224	$\mathbf{-0.117}$
H	6729	1439	0.01953	0.02187	$\mathbf{-0.107}$
T	18231	4246	0.05761	0.05781	$\mathbf{-0.003}$
A	23925	5939	0.08059	0.07435	$\mathbf{0.084}$
G	21811	5460	0.07409	0.06759	$\mathbf{0.096}$
Q	15248	3924	0.05324	0.04681	$\mathbf{0.137}$
S	24800	6595	0.08949	0.07526	$\mathbf{0.189}$
D	17816	4739	0.0643	0.05406	$\mathbf{0.190}$
K	21160	6028	0.08179	0.06255	$\mathbf{0.308}$
P	18942	5477	0.07432	0.05566	$\mathbf{0.335}$
E	24719	7236	0.09818	0.07227	$\mathbf{0.359}$
ukupno:	**315592**	**73688**			

Tabela 2.1

Na ovaj način dobijena je **skala aminokiselina** u odnosu na njihovu neuređenost (CWIFYLMVNRHTAGQSDKPE). Ova skala se ne poklapa u potpunosti sa onom dobijenom u [7][20], ali ipak korelacija između njih je veća od 0.98, te se može reći da su u skladu jedna sa drugom.

2.3.2 Korelacija

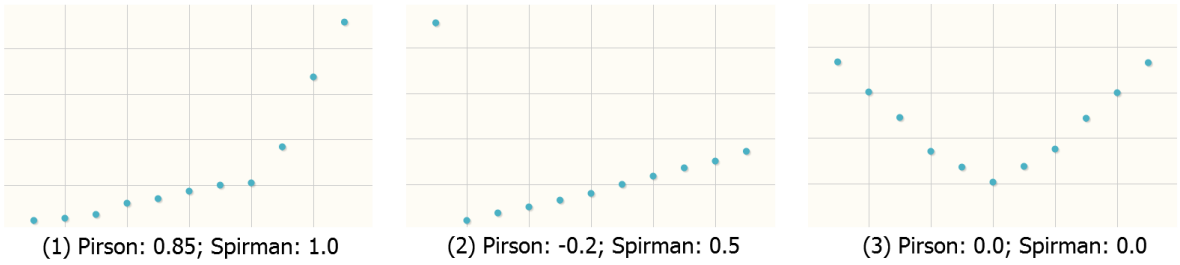
U statistici povezanost između slučajnih promenljivih se može razlikovati po smeru i po jačini. Najjača ili najuža veza između dve promenljive je *funkcionalna veza*, tj. takva veza da svaka vrednost jedne promenljive funkcionalno zavisi od tačno određene vrednosti druge. Labavija veza između promenljivih, koja je podložna manjim ili većim odstupanjima, naziva se **korelativna veza**.

Naime, **kovarijansa** je mera zajedničke varijabilnosti dve slučajne promenljive. Ako veće i manje vrednosti jedne promenljive odgovaraju redom većim i manjim vrednostima druge promenljive, onda se te dve promenljive slično ponašaju, pa je njihova kovarijansa pozitivna. I suprotno, ako veće i manje vrednosti jedne promenljive odgovoraju redom manjim i većim vrednostima druge promenljive, onda je njihova kovarijansa negativna. Posebno, ako su promenljive nezavisne onda je njihova kovarijansa 0.

[20] Skala aminokiselina u [7] je računata samo nad središnjim delovima proteina, a neki proteini nisu ni bili uzimani u razmatranje. Zbog ovoga postoje mala neslaganja te i skale dobijene u ovom radu.

Dakle, kovarijansa dve promenljive jasno daje smer njihove povezanosti, ako uopšte postoji, ali iz nje se ipak ne vidi i jačina povezanosti, te je za merenje povezanosti dve promenljive bolje koristiti normalizovanu verziju kovarijanse – **koeficijent korelacije** – čije se vrednosti kreću u intervalu $[-1, 1]$. Znak koeficijenta određuje smer povezanosti, a njegova apsolutna vrednost određuje jačinu, što je bliža jedinici - to je povezanost jača.

Ipak, priroda kovarijanse pa i koeficijenta korelacije je takva da loše meri **nelinerane veze** između promenljivih, a pored toga na proračun značajno utiču i autlajeri. Zato su konstruisane malo robusnije mere, manje osetljive na postojanje autlajera i malo bolje prate nelinearne veze, tzv. **korelacije ranga**. Na prikazu 2.5 date su neke karakteristične forme veza između dve promenljive. Na (1) vidi se kako korelacija ranga (*Spirman*) bolje meri nelinearnu vezu, na (2) koeficijent korelacije (*Pirson*) zbog autlajera potpuno pogrešno procenjuje smer povezanosti i na (3) vidi se kako ipak neke nelinearne veze ne može da prati ni korelacija ranga ni koeficjent korelacije.



Prikaz 2.5

U ovom radu za pronalaženje jako korelisanih indeksa sa koeficijentom neuređenosti paralelno se koriste Pirsonov koeficijent korelacije i Spirmanova korelacija ranga. Slede konkretne formule po kojim je vršeno izračunavanje.

Neka su iz nekog skupa podataka izdvojeni objekti $x = (x_1, x_2, \dots, x_n)$ i $y = (y_1, y_2, \dots, y_n)$, gde su vrednosti atributa $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}$. Koeficijent korelacije objekata x i y, poznat kao i *Pirsonov koeficijent* ili *Pirsonova korelacija*, dat je sa:

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}}$$

gde su $\bar{x}$ i $\bar{y}$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Računanje *Spirmanove korelacije ranga* se obavlja u dve etape. Prvo se numeričke vrednosti atributa objekata x i y sortiraju po veličini i određuje se njihova pozicija iliti rang u tom rasporedu. Ovim se gube stvarne razlike između numeričkih vrednosti. Ukoliko su neke vrednosti $v_i, v_{i+1}, \dots, v_{i+k}$ jednog objekta međusobno jednake, onda se svi njihovi rangovi $r_i, r_{i+1}, \dots, r_{i+k}$ zamenjuju sa $r_i + r_{i+1} + \cdots + r_{i+k}/k+1$, $(1 \le i < n, 1 \le k < n - i)$.

U drugoj etapi izračunava se koeficijent korelacije nad određenim rangovima vrednosti i rezultat toga je upravo Spirmanova korelacija ranga.

Izračunate su obe korelacije između svakog od indeksa i koeficijenta neuređenosti. U tabeli 2.2 izlistana su dva spiska od po 20 najkorelisanijih indeksa sa neuređenošću. Radi preglednosti pored samih korelacija naveden je samo identifikacioni kod indeksa i kratki komentar koji se odnosi na regiju kojoj pripada[21]. Rezultati su sortirani prema apsolutnoj vrednosti Pirsonove odnosno Spirmanove korelacije.

ID	Pirs	Regija	ID	Spir	Regija
BAEK050101	-0.914	NA (*linker*)	BAEK050101	-0.941	NA (linker)
VINM940101	0.912	NA (*flexibility*)	VINM940101	0.926	NA (*flexibility*)
CASG920101	-0.91	NA (*hydrophobicity*)	NISK860101	-0.926	*hydrophobicity*
MIYS990104	0.907	NA (*energies*)	CASG920101	-0.925	NA (*hydrophobicity*)
NISK860101	-0.897	*hydrophobicity*	MIYS990104	0.923	NA (*energies*)
GUYH850102	0.893	NA (*energies*)	MIYS990105	0.907	NA (*energies*)
MIYS990103	0.892	NA (*energies*)	WERD780101	-0.896	*hydrophobicity*
MIYS990105	0.891	NA (*energies*)	PONP930101	-0.892	NA (*hydrophobicity*)
WERD780101	-0.884	*hydrophobicity*	MIYS990103	0.891	NA (*energies*)
VINM940103	0.871	NA (*flexibility*)	GUYH850102	0.888	NA (*energies*)
VINM940102	0.869	NA (*flexibility*)	OOBM770103	0.881	*hydrophobicity*
CORJ870101	-0.867	NA (*hydrophobicity*)	CORJ870101	-0.878	NA (*hydrophobicity*)
PONP930101	-0.863	NA (*hydrophobicity*)	QIAN880121	-0.867	*beta propensity*
BASU050102	-0.862	NA (*hydrophobicity*)	MEIH800101	0.864	*hydrophobicity*
OOBM770103	0.855	*hydrophobicity*	VINM940103	0.86	NA (*flexibility*)
KARP850102	0.852	*hydrophobicity*	KARP850102	0.856	*hydrophobicity*
FASG890101	0.852	NA (*hydrophobicity*)	WIMW960101	-0.854	NA (*energies*)
ZHOH040103	-0.85	NA (*hydrophobicity*)	BASU050102	-0.854	NA (*hydrophobicity*)
ROSG850102	-0.85	*hydrophobicity*	CHOP780202	-0.854	*beta propensity*
NISK800101	-0.848	*hydrophobicity*	BIOV880101	-0.853	*hydrophobicity*

Tabela 2.2

Ako bismo pravili neki zaključak samo na osnovu ovih spiskova, mogli bismo da pretpostavimo da postoji neka zavisnost između hidrofobnosti i koeficijenta neuređenosti. Pored hidrofobnosti, kategorije/regije najkorelisanijih indeksa su: *linker*, *flexibility*,*energies* i *beta propensity*.

[21] Ukoliko indeks prema [9] nije kategorisan u neku od 6 regija, onda se beleži neki naziv zasnovan na kratkom opisu indeksa koji je dostupan u *AAindex1*, i to u formi NA(*naziv*).

3 ISTRAŽIVANJE ZAVISNOSTI AMINOKISELINSKIH INDEKSA I KOEFICIJENTA NEUREĐENOSTI

Cilj - otkrivanje zavisnosti između karakteristika aminokiselina i njihove tendencije da se pojavljuju u neuređenim regionima. Pre svega, potrebno je da odredimo šta su uopšte karakterstike tj. osobine aminokiselina u kontekstu ulaznih podataka. Na primer, posmatrajući *AAindex* bazu, jedan indeks možemo da uzmemo kao meru jedne specifične osobine tj. *jedan indeks = jedna osobina*, ili kako je to urađeno u [9], gde su svi indeksi prema svojoj prirodi razvrstani u samo 6 grupa, možemo celu grupu indeksa da poistovetimo sa jednom osobinom, tj. indekse jedne grupe posmatramo kao različita merenja jedne iste osobine. Za ovaj rad je odabran ovaj drugi princip, tj. *grupa indeksa = jedna osobina*.

Dakle, potrebno je rasporediti indekse u neke prirodne grupe i istražiti da li postoji ikakva veza između tako dobijenih grupa i koeficijenta neuređenosti. Ta podela treba da obuhvati što više indeksa koji se odnose na poznate fizičkohemijske osobine. Nakon razmatranja celokupnog skupa indeksa izdvojene su sledeće **grupe** (namerno se koriste nazivi na engleskom jeziku da bi se lakše povezivali sa izvorom podataka): **hydro**, **energ**, **propensity**, **weight**, **volume**, **charge**, **polar**, **frequency**. Nazivi grupa su zapravo ključne reči kojima raspoređujemo indekse. Naime, indeks pripada nekoj grupi ukoliko njegov opisni tekst (sekcija *D* u *AAindex1* formatu) uključuje naziv te grupe. Na primer - indeks sa opisom: *Hydrophobicity index* pripada grupi *hydro.* Treba napomenuti da je bez dobrog poznavanja domena odabir samih grupa težak zadatak.

U okviru ovog rada sprovedena su dva istraživanja, opisno imenovana – *istraživanje unifikacijom* i *istraživanje najkorelisanijih indeksa*. To su zapravo dva pristupa rešavanja istog problema, idejno se razlikuju, ali imaju isti cilj. Jednim se ispituju veze posebno definisanih grupa indeksa sa koeficijentom neuređenosti, a u drugom umesto grupa posmatraju se pojedinačni reprezentativni indeksi. Željeni rezultat oba pristupa je skup pravila koja detaljno opisuju otkrivene povezanosti. Za automatsko generisanje pravila koristi se algoritam *pretrage pravila pridruživanja*[22]. Ovaj algoritam na ulazu očekuje podatke u formi skupa transakcija, pa je neophodno transformisati, odnosno diskretizovati podatke, što pored grupisanja indeksa predstavlja još jedan od ključnih elemenata u oba pristupa.

Dalje u tekstu detaljno su opisani ovi pristupi, uključujući metode korišćene za transformaciju podataka i rezultate.

[22] Za generisanje pravila i njihovu vizualizaciju korišćen je alat za istraživanje podataka – *IBM Intelligent Miner*.

3.1 Istraživanje unifikacijom

U ovom pristupu istražuju se veze između posebno definisanih grupa indeksa sa koeficijentom neuređenosti[23]. Pretražuju se pravila za svaku grupu ponaosob. Pre pretrage neophodno je diskretizovati indekse datih grupa i sam FD. U skladu sa principom *grupa indeksa = jedna osobina*, potrebno je da se diskretizacijom u jednoj grupi postigne tzv. **unifikacija indeksa**. To znači da nakon diskretizacije, skup diskretnih vrednosti bude isti za svaki indeks, npr. {*malo*, *srednje*, *veliko*}. Ukoliko je ovo ispunjeno, onda možemo da formiramo skup transakcija, gde svaka transakcija ima tačno dve stavke, jedna koja sadrži diskretnu vrednost nekog indeksa grupe i druga koja sadrži odgovarajuću diskretnu vrednost FD (tabela 3.1 – skup transakcija za jednu grupu kreiran od *N* diskretizovanih indeksa sa vrednostima {*malo*, *srednje*, *veliko*} i diskretnog FD sa vrednostima {*negativno*, *pozitivno*}).

AA	Indeks	TID	Osobina	FD
A	indeks1	1	malo	negativno
	indeks2	2	veliko	negativno
	indeks3	3	veliko	negativno
	...		...	
	indeksN	N	srednje	negativno
R	indeks1	N+1	veliko	pozitivno
	indeks2	N+2	malo	pozitivno
	...		...	
V	indeksN	20*N	veliko	negativno

Tabela 3.1

Očekuje se da ukoliko neka osobina aminokiselina jeste povezana sa neuređenošću, onda će se ovim pristupom uočiti pravilnost između odgovarajuće grupe indeksa i FD, odnosno biće generisana neka pravila koja opisuju tu vezu.

3.1.1 Pretpostavka i grupe

Pretpostavka ovog pristupa jeste da su grupe dobro određene. Unifikacija indeksa jedne grupe može da dovede do velikog gubitka informacija kao generalna posledica diskretizacije, ali još važnije, može da dovede do velike konfuzije u podacima. Naime, ukoliko u jednoj grupi imamo indekse koji su na različite načine povezani sa FD, npr. neki su u pozitivnoj, a neki u negativnoj korelativnoj vezi sa FD, onda se unifikovanjem njihovi uticaji potiru i onemogućavaju generisanje bilo kakvih pravila, barem onih potencijalno interesantnih.

U ovom radu pri određivanju grupa, cilj je bio da one obuhvate indekse koji se odnose na istu fizičkohemijsku odnosno biohemijsku osobinu ne obraćajući pažnju na čistoću grupe, te da poistovećujući ih sa osobinama, potražimo njihove veze sa neuređenošću. Stoga, ako ne pronađemo pravilnosti za jednu grupu, onda možemo da zaključimo da je grupa preširoka ili da nije povezana sa neuređenošću.

[23] Dalje u tekstu koeficijent neuređenosti se označava sa FD.

Ipak, lako je zamisliti da u nekoj grupi postoje i indeksi koji su pozitivno i oni koji su negativno korelisani sa FD, te je preventivno svaka grupa podeljena na dve:

$grupa \in$ {hydro, energ, propensity, weight, volume, charge, polar, frequency}

$$grupa = grupa+ \cup grupa-$$

gde oznaka $+$ stoji za pozitivno, a $-$ za negativno korelisane. Tako $grupa+$ sadrži samo pozitivno korelisane indekse grupe $grupa$ i obrnuto za $grupa-$. Sada, za svaku ovu podgrupu vršimo nezavisno istraživanje.

3.1.2 Diskretizacija i autlajeri

Unifikacija indeksa jedne grupe podrazumeva diskretizaciju njenih indeksa koja rezultuje uvek istim skupom diskretnih vrednosti, pritom diskretne vrednosti treba da imaju svojstveno značenje za svaki indeks posebno.

Pravila koja želimo da dobijemo na kraju analize su dužine 2 i oblika:

$$(osobina = v_1) \Longrightarrow (FD = v_2)$$

gde je v_1 iz skupa diskretnih vrednosti indeksa, a v_2 iz skupa diskretnih vrednosti FD. Pritom želimo da diskretne vrednosti v_1 i v_2 budu opisnog karaktera. Npr. $v_1 \in$ {*malo*, *srednje*, *veliko*} i $v_2 \in$ {*neuređeno, uređeno*}.

Za proces diskretizovanja indeksa koristi se metod – **intervali jednake širine**, opisan u uvodnom poglavlju ovog teksta. Bira se fiksni broj diskretnih vrednosti n, odnosno broj diskretizacijskih intervala, važeći za sve indekse date grupe. Pored toga, za svaki indeks i određuje se širina intervala d_i na osnovu raspona njegovih vrednosti i zadatog broja n. Odabrana je baš ova diskretizacija jer je jednostavna, odgovara potrebama unifikacije, ali i zato što rezultujuće diskretne vrednosti opisuju raspon vrednosti svakog indeksa ponaosob.

Ipak, ova metoda diskretizacije je jako osetljiva na autlajere. Autlajeri indeksa i utiču na pogrešno izračunavanje širine intervala d_i, što dovodi do nejednakog raspoređivanja vrednosti indeksa u diskretizacijske intervale, pa između ostalog rezultat gubi opisni karakter.

U podacima su uočeni autlajeri, te da bi primenili ovu diskretizacijsku metodu, neophodno je da na neki način eliminišemo njihov uticaj. U tu svrhu osmišljen je jednostavan i uglavnom uspešan metod za otkrivanje autlajera tokom diskretizacije.

U prikazu 3.1 dat je algoritam unifikacije indeksa koji podrazumeva primenu diskretizacije *intervala jednake širine* nad svim indeksima jedne grupe uz razmatranje autlajera. Rezultat unifikacije indeksa je skup diskretizacijskih skala, za svaki indeks jedna skala. Naime, diskretizacijska skala je kolekcija intervala za diskretizaciju određenog indeksa. Dakle, skala potpuno određuje diskretizaciju jednog indeksa.

Algoritam – Unifikacija indeksa

ulaz: n – broj intervala, skup I_g – skup indeksa grupe g,
out_tol – koeficijent tolerancije;
izlaz: skup *scales* – skup diskretizacijskih skala indeksa iz I_g;

if ($n < 2$) **then begin**
error „*Broj intervala mora biti veći od 1.*";
exit();
end
scales = ∅; // Inicijalizacija skupa skala
foreach ($i \in I_g$) **do begin**
i_aut_free = eliminate_autliers(i, *out_tol*); // eliminacija autlajera

i_min = min(i_aut_free);
i_max = max(i_aut_free);
d = abs(i_max – i_min)/n; // određivanje širine intervala

scale = ∅; // inicijalizacija skupa diskretizacijskih intervala
split = i_min + d;
scale += ($-\infty$, split); // dodavanje prvog diskretizacijskog intervala
for (counter = 1; counter < (n–1); counter++) **do begin**
scale += [split, split+d);
split = split + d;
end
scale += [split, $+\infty$); // dodavanje poslednjeg diskretizacijskog intervala

scales += (id(i), scale);
end
Na izlaz → *scales*;

Prikaz 3.1

Eliminacija autlajera je koncipirana tako da uvek prvo sumnjiči globalne ekstremne vrednosti, tj. mininum i maksimum. Vrednosti indeksa se razvrstavaju u dva skupa, one koje su bliže minimumu nego maksimumu idu u skup I_{min}, a ostale u skup I_{max}. Ukoliko I_{min} sadrži skoro sve vrednosti indeksa, onda je maksimum dosta odvojen od ostalih vrednosti i smatra se autlajerom. U prikazu 3.2 data je funkcija koja je zadužena za eliminaciju autlajera. Ona na ulazu prihvata koeficijent tolerancije, tj. procentualno izraženu donju granicu za količinu vrednosti u jednom od skupova I_{max} odnosno I_{min} da bi odgovarajuća ekstremna vrednost bila proglašena autlajerom. Prepoznati autlajeri se izbacuju iz skupa vrednosti odgovarajućeg indeksa i takav modifikovan indeks se stavlja na izlaz funkcije.

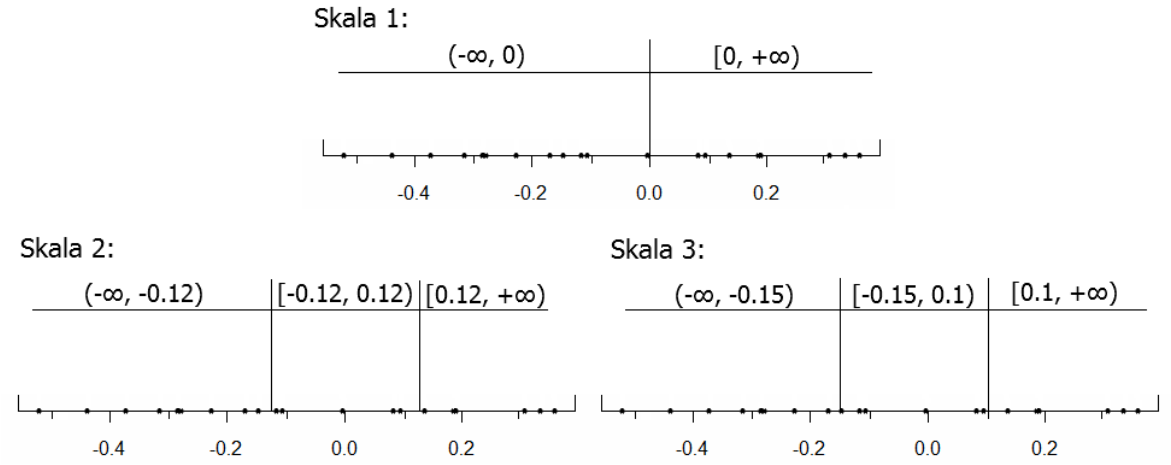
Primenom ovog algoritma u 91 od 544 dostupna indeksa su pronađeni autlajeri, ukupno 111 autlajer vrednosti, za koeficijent tolerancije 85%. Tokom istraživanja tolerancija od 85% generalno je prihvaćena za otkrivanje autlajera.

```
eliminate_autliers (i: indeks, out_tol: koeficijent tolerancije)

    status = true;
    i_tmp = i;

    while (status) do begin
        if (broj različitih vrednosti u i_tmp < 3) then
            status = false;
        else begin
            c = count(i_tmp);      // broj vrednosti u i_tmp
            min = min(i_tmp);
            max = max(i_tmp);
            // broj vrednosti koje su bliže maksimumu nego minimumu i obrnuto
            c_min = count(subset(i_tmp, abs(i_tmp−min) ≥ abs(max−i_tmp)));
            c_max = count(subset(i_tmp, abs(max−i_tmp) ≥ abs(i_tmp−min)));

            max_stat = 100*c_max/c;    // procenti
            min_stat = 100*c_min/c;

            // diskusija za otkrivanje autlajera
            if (max_stat ≥ out_tol) then
                outlier = max;
            else if (min_stat ≥ out_tol) then
                outlier = min;
            else
                outlier = ∅;

            if (outlier == ∅) then
                status = false;
            else   // brišu se sve vrednosti jednake otkrivenom autlajeru
                i_tmp = i_tmp – outlier;
        end
    end
```

Prikaz 3.2

Ovaj metod diskretizacije za parametar ima broj intervala n. Ovde je za sve grupe izvršena diskretizacija na 3, a za neke i na 4 intervala.

Pored indeksa, potrebno je diskretizovati i FD. Za određivanje diskretizacijskih intervala FD ne postoji tačno determinisan pristup. U ovom istraživanju uglavnom se koriste tri skale (prikaz 3.3). Najjednostavnija obuhvata samo dve diskretne vrednosti – *negativno* i *pozitivno*, odnosno *uređeno* (tačnije *neneuređeno*) i *neuređeno*. U drugoj i trećoj skali, koje obuhvataju po tri diskretne vrednosti, oko nule dodaje se mali interval tolerancije, tako da prvi i treći interval budu bolje razdvojeni. Ideja je da se na ovaj način obezbede čistija pravila koja se odnose na prvi i treći interval, tj. *jaku neneuređenost* i *jaku neuređenost*.



Prikaz 3.3

3.1.3 Analiza i rezultati

Indeksi su razvrstani po grupama $grupa+$ i $grupa-$, $grupa \in \{hydro, energ, propensity, weight, volume, charge, polar, frequency\}$. Nad svakom grupom je izvršena unifikacija indeksa i formirani su skupovi transakcija. Za sve grupe je izvršena unifikacija sa tri, a za neke i 4 diskretne vrednosti[24]. Takođe, primenjivane su različite diskretizacijske skale za FD.

Nakon velikog broja izvršenih analiza, primećen je obrazac – *grupe koje sadrže relativno veći broj indeksa ne daju rezultate*, pa čak i kod grupa *hydro+* i *hydro-* za koje se prema dosadašnjim saznanjima očekuje da postoji nekakva veza koja bi se manifestovala i u formi pravila. U tabeli 3.2 su date kardinalnosti svih grupa.

Grupa	|Grupa|	Grupa	|Grupa|
hydro+	13	*hydro-*	40
energ+	28	*energ-*	23
propensity+	14	*propensity-*	7
weight+	22	*weight-*	27
volume+	0	*volume-*	12
charge+	4	*charge-*	2
polar+	3	*polar-*	2
frequency+	67	*frequency-*	32

Tabela 3.2

Ovaj obrazac je posledica načina na koji su grupisani indeksi. Naime, pri grupisanju se čistoća grupe u odnosu na tip povezanosti koju indeks može da ima sa FD, nije uzimala u obzir. Pored toga, u grupama postoji i značajan procenat indeksa za koje se ne može uočiti da su u bilo kakvoj vezi sa FD, te igraju ulogu šuma i negativno utiču na tok analize.

Iako u njima nema puno indeksa, grupe *charge+* i *charge-*, takođe ne daju interesantna pravila. Može se reći da nemaju veze sa neuređenošću, ali ipak treba imati u vidu da vrednosti indeksa koji pripadaju ovim grupama uzimaju samo tri različite vrednosti -1, 0 i 1. Jedino preostaju grupe *polar+* i *polar-*.

Grupa *polar+*:

P_1: $(polar+ = malo) \Longrightarrow (FD = neuređeno)$; $s(P_1) = 45\%$ $c(P_1) = 77.14\%$

Diskretizacija FD – {*neuređeno*, *neuređeno*}
Unifikacija indeksa – {*malo*, *srednje*, *veliko*}

[24] Za neke konkretne grupe npr. polar+ i polar-, hydro+ i hydro-, volume+ i volume- vršena su istraživanja i za veće n (5, 7, 9, 11, 13), koristeći pritom i razne diskretizacijske skale koeficijenta neuređenosti odnosno FD.

Povećanje broja intervala unifikacije i upotreba drugih diskretizacijskih skala za FD, nije dovelo do pronalaženja još nekih zanimljivih pravila koja bi se izdvojila svojim kvalitetom.

Grupa *polar*- ima dva indeksa te daje skup od samo 40 transakcija, i ni u jednoj kombinaciji parametara za unifikaciju indeksa i diskretizaciju FD ne nudi neka zanimljiva pravila.

Napredak prvenstveno treba potražiti u boljem grupisanju indeksa. Bolje formirane grupe gde su indeksi pažljivo filtrirani, uz dobro poznavanje domena, u ovom slučaju molekularne biologije, može da dovede do željenih rezultata. Problem grupisanja indeksa rešavan je klasterovanjem u [9], gde su 402 indeksa kategorisana u 6 grupa. Ipak, tako dobijene grupe takođe imaju, slično kao i grupe kreirane u ovom radu, veliku različitost u načinu na koji su indeksi povezani sa koeficijentom neuređenosti, kao i značajnu količinu šuma. Tako da je malo verovatno da se korišćenjem ove podele može doći do značajno boljih rezultata.

3.2 Istraživanje najkorelisanijih indeksa

Da bismo zaobišli posledice koje nastaju unifikacijom suviše raznovrsnih grupa, pristupamo istraživanju na drugi način. Unutar svake grupe pronalazimo visoko korelisane reprezentativne indekse za koje se onda pojedinačno analiziraju detalji njihove povezanosti sa FD. Na ovaj način se otkrivaju različite povezanosti sa koeficijentom neuređenosti u okviru jedne grupe.

Isto kao i u prethodnom slučaju, ovde se koristi algoritam za pretragu pravila pridruživanja, ali sada u cilju ispitivanja veze između pojedinačnih, posebno odabranih indeksa i FD. Opet, potrebno je transformisati ulazne podatke u skup transakcija, pri čemu svaka transakcija ima uvek isti broj stavki, a najviše onoliko koliko data grupa ima reprezentativnih indeksa[25] i plus jednu za diskretnu vrednost FD. Naravno, da bismo oformili takav skup neophodno je da se diskretizuju svi indeksi i sam FD. Pritom, ovde se ne izvršava unifikacija indeksa, pa će skup transakcija uvek imati samo 20 redova.

3.2.1 Reprezentativni indeksi

U istraživanju zasnovanom na unifikaciji pronađeno je da su grupe suviše raznovrsne, odnosno da sadrže indekse koji su na različite načine povezani sa FD. Stoga, bilo bi dobro razbiti indekse na podgrupe, tako da indeksi jedne podgrupe budu na isti način povezani sa FD. Ovde se u tu svrhu primenjuje hijerarhijsko klasterovanje, sa metodom *single link*, gde se za meru rastojanja koristi Pirsonova korelacija ($d(I_1, I_2) = abs(pirs(I_1, I_2))$, rastojanje između indeksa I_1, I_2). Pritom, minimalno rastojanje između dva klastera iznosi 0.2. Rezultati ovog klasterovanja se slažu sa podacima prikazanim u bazi *AAindex*, gde je za svaki indeks dat spisak indeksa koji su u visokoj apsolutnoj korelativnoj vezi (> 0.8) sa njim.

Ovim klasterovanjem indeksi su razgraničeni na male homogene podgrupe. Iz svake podgrupe se bira indeks koji je u najvećoj korelativnoj vezi sa FD, pri čemu se za merenje korelacije koristi Spirmanova i Pirsonova mera. Na taj način iz jedne grupe se može izdvojiti do dva indeksa – jedan najkorelisaniji sa FD prema Pirsonovoj, a drugi prema Spirmanovoj meri. Te najkorelisanije indekse nazivamo **reprezentativnim indeksima**.

Grupa	Ukupno	R	F
hydro	53	13	3
energ	46	24	4
propensity	20	15	5
weight	47	26	3
volume	12	5	2
charge	5	5	2
polar	5	5	2
frequency	97	55	6

Tabela 3.3

Da bismo osigurali da veza između reprezentativnog indeksa i FD postoji i oslobodili se šuma, indekse dodatno filtriramo. Izdvajaju se oni čija je apsolutna korelacija sa FD veća od 0.7 (prema Spirmanovoj ili Pirsonovoj meri). U tabeli 3.3 za svaku grupu dat je ukupan broj indeksa, broj reprezentativnih (R) i broj filtriranih (F).

[25] Skup reprezentativnih indeksa jedne grupe se dodatno filtrira pre nego što se oformi skup transakcija, te je broj stavki jedne transakcije uvek $n + 1$, gde je n broj indeksa u filtriranom skupu reprezentativnih indeksa.

3.2.2 Diskretizacija

Pre formiranja skupa transakcija, neophodno je diskretizovati određene indekse i sam FD. Za diskretizaciju FD, koriste se već pomenute skale, date u prikazu 3.3, ali za diskretizaciju indeksa se koristi drugačija metoda od one korišćene u pristupu zasnovanom na unifikaciji.

Pošto ne postoji potreba za unifikacijom, skup diskretnih vrednosti može da se razlikuje od indeksa do indeksa. Moguće je koristiti neku od kontrolisanih metoda diskretizacije koje zavise od informacija o klasi i uglavnom daju dobre rezultate. U ovom kontekstu, diskretne vrednosti FD predstavljaju klasu (npr. klase neuređeno / neneuređeno).

Pravila koja želimo da dobijemo na kraju analize su dužine 2 i oblika:

$$(I = v_1) \Longrightarrow (FD = v_2)$$

gde je v_1 iz skupa diskretnih vrednosti indeksa I, a v_2 iz skupa diskretnih vrednosti FD. Čistoća diskretizacijskih intervala nekog indeksa u odnosu na diskretne vrednosti FD ima veliki uticaj na kvalitet pravila koja uključuju taj indeks. Čistiji intervali znače veću pouzdanost pravila, a širi intervali mogu da doprinesu boljoj podršci. Stoga, da bi dobili kvalitetnija pravila traži se da diskretizacija proizvede što manji broj diskretnih vrednosti sa što čistijim intervalima.

Ovde je osmišljena prosta metoda za diskretizovanje indeksa zavisno od diskretnog FD, koja je zasnovana na maksimizovanju čistoće intervala. U narednom primeru opisan je rad ove kontrolisane diskretizacijske metode.

Primer – Neka je diskretizovan FD tako da je skup diskretnih vrednosti {*neneuređeno*, *neuređeno*}. Diskretizovati indeks I. Vrednosti indeksa I (levo) i odgovarajuće diskretne vrednosti FD (desno) date su u tabeli 3.4. Tabela je sortirana prema vrednostima I.

I	FD
−0.4	*neuređeno*
−0.3	*neuređeno*
−0.2	*neneuređeno*
−0.19	*neuređeno*
−0.05	*neuređeno*
0.3	*neneuređeno*
...	
0.9	*neuređeno*
1.2	*neneuređeno*
1.4	*neneuređeno*

Tabela 3.4

Kaže se da je diskretna vrednost FD klasa odgovarajuće vrednosti indeksa I, npr. -0.3 je klase *neuređeno*.

Proces diskretizacije indeksa započinje čitanjem prvog reda tabele. Klasa prve vrednosti, u ovom primeru klasa *neuređeno* postaje ujedno i klasa intervala koji formiramo. Čitaju se redovi tabele sve dok klasa tekućeg reda ne bude različita od klase intervala. Dakle, poslednji pročitani red sadrži vrednost -0.2 klase *neneuređeno*. Tada se definiše prvi diskretizacijski interval $(-\infty, -0.2)$ i rekurzivno se kreiraju ostali intervali. Prvi sledeći interval je $[-0.2, -0.19)$, zatim $[-0.19, 0.3)$ i tako dalje do poslednjeg $[1.2, +\infty)$.

Za diskretne vrednosti indeksa uzimaju se diskretizacijski intervali, pa je skup diskretnih vrednosti indeksa I jednak sledećem:

$$\{„(-\infty, -0.2)“, „[-0.2, -0.19)“, „[-0.19, 0.3)“, \ldots, „[1.2, +\infty)“\}. \blacksquare$$

Sve vrednosti koje pripadaju jednom diskretizacijskom intervalu su iste klase, što znači da je čistoća tih intervala maksimalna. Pritom, ako je indeks I jako korelisan sa FD onda je broj intervala skale diskretizacije mali.

Dakle, uz pretpostavku da je indeks koji se diskretizuje jako korelisan sa FD ova diskretizacija daje mali broj i to maksimalno čistih intervala i time ispunjava uslove za generisanje kvalitetnijih pravila.

3.2.3 Analiza i rezultati

Indeksi su razvrstani po grupama $grupa \in \{hydro, energ, propensity, weight, volume, charge, polar, frequency\}$ i nad svakom grupom su odabrani reprezentativni. Pre diskretizacije, reprezentativni indeksi su filtrirani uslovom da je apsolutna korelacija sa FD veća od 0.7. FD je diskretizovan na tri različita načina, prikaz 3.3. Za različite diskretizacije FD, izvršena je različita diskretizacija filtriranih indeksa i formiran je skup transakcija nad kojim su pretraživana pravila. Cilj je da pravilima povežemo diskretne vrednosti indeksa sa diskretnim vrednostima FD.

Generisana su pravila dužine 2, oblika:

$$(I = v_1) \Longrightarrow (FD = v_2)$$

gde su v_1 i v_2 odgovarajuće diskretne vrednosti. Zbog karakteristične diskretizacije, sva pravila su 100% pouzdana[26], a podrška se kreće od 5% pa do podrške posledice pravila. Nezavisno tumačimo pravila svakog indeksa (pravila indeksa I su pravila koja u premisi imaju diskretne vrednosti indeksa I) i izdvajamo ona kvalitetna. Kvalitet pravila posmatramo kroz podršku – što veća podrška to bolje odnosno kvalitetnije pravilo. S obzirom na to da se istraživanje sprovodi nad samo 20 transakcija, najzanimljivija pravila su ona sa veoma visokom podrškom. Ovde je puna pažnja posvećena pravilima sa podrškom do 30% manjom od podrške posledice.

Pravila	Podrška	Pouzdanost
[CASG920101=[0.400, +b)] ==> [FD=(-b, 0)]	45.0000%	100.0000%
[FASG890101=(-b, -0.209)] ==> [FD=(-b, 0)]	45.0000%	100.0000%
[NADH010106=[79.000, +b)] ==> [FD=(-b, 0)]	40.0000%	100.0000%
[FASG890101=[0.779, 1.360)] ==> [FD=(-b, 0)]	10.0000%	100.0000%
[NADH010106=[-57.000, -47.000)] ==> [FD=(-b, 0)]	10.0000%	100.0000%
[CASG920101=[-0.500, -0.100)] ==> [FD=(-b, 0)]	10.0000%	100.0000%
[NADH010106=[-77.000, -67.000)] ==> [FD=(-b, 0)]	5.0000%	100.0000%
[FASG890101=[2.109, 2.299)] ==> [FD=(-b, 0)]	5.0000%	100.0000%
[NADH010106=[27.000, 45.000)] ==> [FD=(-b, 0)]	5.0000%	100.0000%
[NADH010106=[-47.000, 27.000)] ==> [FD=[0, +b)]	25.0000%	100.0000%
[FASG890101=[1.360, 2.109)] ==> [FD=[0, +b)]	15.0000%	100.0000%
[FASG890101=[-0.209, 0.779)] ==> [FD=[0, +b)]	15.0000%	100.0000%
[CASG920101=[-0.100, 0.400)] ==> [FD=[0, +b)]	10.0000%	100.0000%
[FASG890101=[2.299, +b)] ==> [FD=[0, +b)]	10.0000%	100.0000%
[NADH010106=[-67.000, -57.000)] ==> [FD=[0, +b)]	5.0000%	100.0000%
[NADH010106=(-b, -77.000)] ==> [FD=[0, +b)]	5.0000%	100.0000%
[NADH010106=[45.000, 79.000)] ==> [FD=[0, +b)]	5.0000%	100.0000%

Tabela 3.5

U tabeli 3.5 prikazana su pravila generisana nad indeksima iz grupe *hydro*.

Diskretizacija FD - $\{„(-\infty, 0)“, „[0, +\infty)“\}$,

$s\left(\left(FD = (-\infty, 0)\right)\right) = 60\%$

$s\left(\left(FD = [0, +\infty)\right)\right) = 40\%$

[26] Pouzdanost može da bude i manja od 100%, kao posledica diskretizacije kada neki indeks ima dve ili više jednakih vrednosti različitih klasa. Ipak ovakvi slučajevi su retki i nemaju značajan uticaj na istraživanje.

Na osnovu prva tri pravila jasno se skreće pažnja na indekse – CASG920101, FASG890101 i NADH010106. Analizom su pronađena kvalitetna pravila koja povezuju ove indekse sa neneuređenošću (FD=$(-\infty, 0)$), ali idealno bi bilo kada bi se na spisku nalazila i dobra pravila koja za posledicu daju neuređenost (FD=$[0, +\infty)$), odnosno da postoje dobra pravila za svaku diksretnu vrednost FD. Međutim, ovde to nije tako. Ipak, u nekim slučajevima nadovezivanjem uzastopnih diskretizacijskih intervala može se na račun pouzdanosti dobiti i kompletniji i pregledniji skup pravila. Posmatrajmo parove diskretnih vrednosti indeksa CASG920101 i FD na prikazu 3.4:

FD	(-∞,0)	(-∞,0)	(-∞,0)	(-∞,0)	(-∞,0)	(-∞,0)	(-∞,0)	(-∞,0)	(-∞,0)	(-∞,0)	(-∞,0)	(-∞,0)
AA	C	W	I	F	Y	L	M	V	H	N	T	R
CASG920101	[0.4,+∞)	[0.4,+∞)	[0.4,+∞)	[0.4,+∞)	[0.4,+∞)	[0.4,+∞)	[0.4,+∞)	[0.4,+∞)	[0.4,+∞)	[-0.5,-0.1)	[-0.5,-0.1)	(-∞,-0.5)

FD	[0,+∞)	[0,+∞)	[0,+∞)	[0,+∞)	[0,+∞)	[0,+∞)	[0,+∞)	[0,+∞)
AA	A	G	Q	S	D	K	P	E
CASG920101	[-0.1,0.4)	[-0.1,0.4)	(-∞,-0.5)	(-∞,-0.5)	(-∞,-0.5)	(-∞,-0.5)	(-∞,-0.5)	(-∞,-0.5)

Prikaz 3.4

Nadovežimo intervale $[-0.5, -0.1)$, $[-0.1, 0.4)$ i $[0.4, +\infty)$ u jedan zajednički $[-0.5, +\infty)$. Sada je ceo raspon vrednosti indeksa CASG920101 diskretizovan na dve kategoričke vrednosti. Mogu da se ekstraktuju pravila:

P_1: $\big(CASG920101 = (-\infty, -0.5)\big) \Longrightarrow \big(FD = [0, +\infty)\big)$; $\quad s(P_1) = 30\% \quad c(P_1) = 85.7\%$
P_2: $\big(CASG920101 = [-0.5, +\infty)\big) \Longrightarrow \big(FD = (-\infty, 0)\big)$; $\quad s(P_2) = 55\% \quad c(P_2) = 84.6\%$

Ova dva pravila imaju visoku i pozudanost i podršku i značajno su preglednija nego skup pravila indeksa CASG920101 dat u tabeli 3.5.

U sledećim tabelama (tabela 3.6-3.8) izdvojena su najinteresantnija pravila za svaku grupu i svaku upotrebljenu diskretizaciju FD. Kratak opis izdvojenih indeksa dat je u tabeli 3.9.

FD Diskretizacija: {„$(-\infty, 0)$“, „$[0, +\infty)$“}			
$s\left(\left(FD = (-\infty, 0)\right)\right) = 60\%$		$s\left(\left(FD = [0, +\infty)\right)\right) = 40\%$	
Indeks	Pravilo	Podrška	Pouzdanost
grupa: ***hydro***			
CASG920101	$\left(CASG920101 = (-\infty, -0.5)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	30%	85.7%
	$\left(CASG920101 = [-0.5, +\infty)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	55%	84.6%
grupa: ***energ***			
MUNV940104	$\left(MUNV940104 = (-\infty, 0.783)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	50%	100%
	$\left(MUNV940104 = [0.783, +\infty)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	40%	80%
MIYS990105	$\left(MIYS990105 = (-\infty, -0.019)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	60%	85.7%
	$\left(MIYS990105 = [-0.019, +\infty)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	30%	100%
grupa: ***propensity***			
KIMC930101	$\left(KIMC930101 = (-\infty, -0.409)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	55%	100%
	$\left(KIMC930101 = [-0.409, +\infty)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	40%	88.9%
FODM020101	$\left(FODM020101 = (-\infty, 0.949)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	35%	77.8%
	$\left(FODM020101 = [0.949, +\infty)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	50%	90.9%
WERD780101	$\left(WERD780101 = (-\infty, 0.419)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	30%	85.7%
	$\left(WERD780101 = [0.419, +\infty)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	55%	84.6%
grupa: ***weight***			
QIAN880121	$\left(QIAN880121 = (-\infty, -0.09)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	40%	88.9%
	$\left(QIAN880121 = [-0.09, +\infty)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	55%	100%
grupa: ***volume*** – rasplinuta pravila			
grupa: ***charge*** – rasplinuta pravila			
grupa: ***polar*** – rasplinuta pravila			
grupa: ***frequency***			
CHOP780209	$\left(CHOP780209 = (-\infty, 0.889)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	40%	88.9%
	$\left(CHOP780209 = [0.889, +\infty)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	55%	100%
CHOP780202	$\left(CHOP780202 = (-\infty, 0.87)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	35%	100%
	$\left(CHOP780202 = [0.87, +\infty)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	60%	92.3%
TANS770104	$\left(TANS770104 = (-\infty, 0.927)\right) \Longrightarrow \left(FD = (-\infty, 0)\right)$	55%	100%
	$\left(TANS770104 = [0.927, +\infty)\right) \Longrightarrow \left(FD = [0, +\infty)\right)$	40%	88.9%

Tabela 3.6

FD Diskretizacija: {„$(-\infty, -0.12)$“, „$[-0.12, 0.12)$“, „$[0.12, +\infty)$“} $s\left((FD = (-\infty, -0.12))\right) = 45\%$ $s\left((FD = [-0.12, 0.12))\right) = 25\%$ $s\left((FD = [0.12, +\infty))\right) = 30\%$			
grupa: ***hydro***			
CASG920101	$(CASG920101 = (-\infty, -0.5)) \Longrightarrow (FD = [0.12, +\infty))$	30%	85.7%
	$(CASG920101 = [-0.5, 0.5)) \Longrightarrow (FD = [-0.12, 0.12))$	20%	80%
	$(CASG920101 = [0.5, +\infty)) \Longrightarrow (FD = (-\infty, -0.12))$	40%	100%
grupa: ***energ***			
MIYS990104	$(MIYS990104 = (-\infty, -0.039)) \Longrightarrow (FD = (-\infty, -0.12))$	40%	100%
	$(MIYS990104 = [-0.039, 0.119)) \Longrightarrow (FD = [-0.12, 0.12))$	25%	100%
	$(MIYS990104 = [0.119, +\infty)) \Longrightarrow (FD = [0.12, +\infty))$	30%	85.7%
MIYS990105	$(MIYS990105 = (-\infty, -0.019)) \Longrightarrow (FD = (-\infty, -0.12))$	40%	100%
	$(MIYS990105 = [-0.019, 0.1)) \Longrightarrow (FD = [-0.12, 0.12))$	25%	100%
	$(MIYS990105 = [0.1, +\infty)) \Longrightarrow (FD = [0.12, +\infty))$	30%	85.7%
grupa: ***propensity*** – rasplinuta pravila			
grupa: ***weight*** – rasplinuta pravila			
grupa: ***volume*** – rasplinuta pravila			
grupa: ***charge*** – rasplinuta pravila			
grupa: ***polar*** – rasplinuta pravila			
grupa: ***frequency*** – rasplinuta pravila			

Tabela 3.7

FD Diskretizacija: {„$(-\infty, -0.15)$“, „$[-0.15, 0.1)$“, „$[0.1, +\infty)$“} $s\left(\left(FD = (-\infty, -0.15)\right)\right) = 40\%$ $s\left(\left(FD = [-0.15, 0.1)\right)\right) = 30\%$ $s\left(\left(FD = [0.1, +\infty)\right)\right) = 30\%$			
grupa: ***hydro***			
CASG920101	$\left(CASG920101 = (-\infty, -0.5)\right) \Rightarrow \left(FD = [0.1, +\infty)\right)$	30%	85.7%
	$\left(CASG920101 = [-0.5, 0.5)\right) \Rightarrow \left(FD = [-0.15, 0.1)\right)$	25%	100%
	$\left(CASG920101 = [0.5, +\infty)\right) \Rightarrow \left(FD = (-\infty, -0.15)\right)$	40%	100%
FASG890101	$\left(FASG890101 = (-\infty, -0.209)\right) \Rightarrow \left(FD = (-\infty, -0.15)\right)$	40%	88.9%
	$\left(FASG890101 = [-0.209, 1.36)\right) \Rightarrow \left(FD = [-0.15, 0.1)\right)$	20%	80%
	$\left(FASG890101 = [1.36, +\infty)\right) \Rightarrow \left(FD = [0.1, +\infty)\right)$	25%	83.3%
grupa: ***energ***			
MIYS990105	$\left(MIYS990105 = (-\infty, -0.019)\right) \Rightarrow \left(FD = (-\infty, -0.15)\right)$	40%	100%
	$\left(MIYS990105 = [-0.019, 0.109)\right) \Rightarrow \left(FD = [-0.15, 0.1)\right)$	30%	100%
	$\left(MIYS990105 = [0.109, +\infty)\right) \Rightarrow \left(FD = [0.1, +\infty)\right)$	30%	100%
MIYS990104	$\left(MIYS990104 = (-\infty, -0.039)\right) \Rightarrow \left(FD = (-\infty, -0.15)\right)$	40%	100%
	$\left(MIYS990104 = [-0.039, 0.14)\right) \Rightarrow \left(FD = [-0.15, 0.1)\right)$	30%	85.7%
	$\left(MIYS990104 = [0.14, +\infty)\right) \Rightarrow \left(FD = [0.1, +\infty)\right)$	25%	100%
WIMW960101	$\left(WIMW960101 = (-\infty, 3.829)\right) \Rightarrow \left(FD = [0.1, +\infty)\right)$	25%	100%
	$\left(WIMW960101 = [3.829, 4.119)\right) \Rightarrow \left(FD = [-0.15, 0.1)\right)$	25%	100%
	$\left(WIMW960101 = [4.119, +\infty)\right) \Rightarrow \left(FD = (-\infty, -0.15)\right)$	40%	80%
grupa: ***propensity***			
WERD780101	$\left(WERD780101 = (-\infty, 0.409)\right) \Rightarrow \left(FD = [0.1, +\infty)\right)$	25%	83.3%
	$\left(WERD780101 = [0.409, 0.639)\right) \Rightarrow \left(FD = [-0.15, 0.1)\right)$	20%	80%
	$\left(WERD780101 = [0.639, +\infty)\right) \Rightarrow \left(FD = (-\infty, -0.15)\right)$	40%	88.9%
grupa: ***weight*** – rasplinuta pravila			
grupa: ***volume*** – rasplinuta pravila			
grupa: ***charge*** – rasplinuta pravila			
grupa: ***polar***			
RADA880108	$\left(RADA880108 = (-\infty, -0.479)\right) \Rightarrow \left(FD = [0.1, +\infty)\right)$	25%	83.3%
	$\left(RADA880108 = [-0.479, 0.879)\right) \Rightarrow \left(FD = [-0.15, 0.1)\right)$	25%	71.4%
	$\left(RADA880108 = [0.879, +\infty)\right) \Rightarrow \left(FD = (-\infty, -0.15)\right)$	35%	100%
grupa: ***frequency***			
CHOP780202	$\left(CHOP780202 = (-\infty, 0.829)\right) \Rightarrow \left(FD = [0.1, +\infty)\right)$	25%	83.3%
	$\left(CHOP780202 = [0.829, 1.049)\right) \Rightarrow \left(FD = [-0.15, 0.1)\right)$	20%	100%
	$\left(CHOP780202 = [1.049, +\infty)\right) \Rightarrow \left(FD = (-\infty, -0.15)\right)$	40%	80%

Tabela 3.8

Indeks	Opis	Grupa
CASG920101	*Hydrophobicity scale from native protein structures*	***hydro***
FASG890101	*Hydrophobicity index*	
MUNV940104	*Free energy in beta-strand region*	***energ***
MIYS990104	*Optimized relative partition energies - method C*	
MIYS990105	*Optimized relative partition energies - method D*	
WIMW960101	*Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water*	
KIMC930101	*Thermodynamic beta sheet propensity*	***propensity***
FODM020101	*Propensity of amino acids within pi-helices*	
WERD780101	*Propensity to be buried inside*	
QIAN880121	*Weights for beta-sheet at the window position of 1*	***weight***
RADA880108	*Mean polarity*	***polar***
CHOP780209	*Normalized frequency of C-terminal beta-sheet*	***frequency***
CHOP780202	*Normalized frequency of beta-sheet*	
TANS770104	*Normalized frequency of chain reversal R*	

Tabela 3.9

Kroz indekse CASG920101 i FASG890101 otkrivena je očekivana veza između hidrofobnosti i neuređenosti aminokiselina.

Takođe, uočena je potencijalna uzročna veza između grupe *energ* i neuređenosti, gde postoje čak četiri različita indeksa sa kvalitetnim pravilima. Posebno se ističu 3 pravila indeksa MIYS990105 u tabeli 3.8, sva su 100% pouzdana i maksimalne su podrške.

Zanimljiva su i pravila grupe *frequency* indeksa CHOP780202 i CHOP780209 koji možda ukazuju na uzročnu vezu β-traka i neuređenosti aminokiselina.

Za grupe *volume* i *charge* nisu pronađena pravila koja bi na očigledan način ukazala na povezanost sa neuređenošću aminokiselina.

4 ZAKLJUČAK

Od ranije poznata je veza između hidrofobnosti i neuređenosti aminokiselina. Cilj ovog rada bilo je pronalaženje još nekih karakteristika aminokiselina koje se potencijalno mogu staviti u uzročno-posledičnu vezu sa neuređenošću, ali i uopšte da doprinese razumevanju sastava i osobina neuređenih regiona u proteinima. Stručnjaci molekularne biologije bi mogli da daju konačan sud o valjanosti rezultata dobijenih ovim istraživanjem, kao i eventualnim uzročno-posledičnim vezama između *AAindeksa* i neuređenosti aminokiselina.

Na osnovu informacija iz *DisProt* baze u proteinima su označeni prostorno neuređeni lanci aminokiselina – *neuređeni* regioni i izračunat je koeficijent neuređenosti za svaku aminokiselinu ponaosob. Ovaj koeficijent je računat tako da se akcenat stavlja na odnos frekventnosti aminokiseline u neuređenim regionima i ostalim delovima proteina, te on numerički opisuje tendenciju aminokiseline da se nađe u sastavu neuređenih regiona. Sa druge strane, numeričke vrednosti koje opisuju fizičkohemijska i biohemijska svojstva aminokiselina – *AAindeksi*, raspoređeni su u posebno određene apstraktne grupe, tako da one obuhvataju srodne indekse. Usledio je pokušaj povezivanja vrednosti grupe, dobijenih unifikacijom pripadajućih indeksa, i koeficijenta neuređenosti.

Iako pripadaju istoj grupi, dva *AAindeksa* mogu na različite načine da budu povezani sa koeficijentom neuređenosti i na taj način da kreiraju konfuziju u podacima. Pored ovoga, u jednoj grupi često postoje indeksi koji poput šuma negativno utiču na uočavanje pravilnosti. Iz ovih razloga, izostao je rezultat pristupa zasnovan na unifikaciji indeksa.

U drugom pristupu, grupe su dodatno razbijene, tako da jedna podgrupa sadrži indekse koji su na sličan način povezani sa koeficijentom neuređenosti. Zatim, za svaku podgrupu je odabran po jedan *AAindeks* i izdvojeni su oni najkorelisaniji sa neuređenošću, nakon čega su pravilima povezivane diskretne vrednosti *AAindeksa* i diskretne vrednosti koeficijenta neuređenosti.

Ovaj pristup daje opipljive rezultate. Posebno su se izdvojili indeksi grupa *hydro* i *energ*, a u okviru grupe *frequency* primećena je jaka povezanost frekvencije aminokiselina u β-trakama sa koeficijentom neuređenosti, dok za grupe *volume* i *charge* nisu otkrivena zanimljiva pravila.

Kvalitativna poboljšanja rezultata mogla bi da se dobiju, eventualno, razvojem i primenom drugih, sveobuhvatnijih mera korelisanosti odnosno pridruženosti slučajnih promenljivih određenog skupa, u našem slučaju – aminokiselinskih indeksa i koeficijenta neuređenosti aminokiselina. Ovo otvara prostor za istraživanje i razvog novih algoritamskih i informatičkih metoda i tehnika.

Literatura

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, *Introduction to Data Mining,* Addison-Wesley, 2006

[2] Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2006

[3] Xindon Wu, Vipin Kumar, *The Top Ten Algorithms in Data Mining*, CRC Press, 2009

[4] Alexander Isaev, *Introduction to Mathematical Methods in Bioinformathics*, Springer, 2006

[5] Hans-Joachim Böckenhauer, Dirk Bongartz, *Algorithmic Aspects of Bioinformathics*, Springer, 2007

[6] http://www.disprot.org, decembar 2011.

[7] Jovana J. Kovačević, *Computational analysis of experimentally determined disorder proteins*

[8] Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradović Z, Dunker AK, *Disprot: the Database of Disordered Proteins*, 2006

[9] Kentaro Tomii, Minoru Kanehisa, *Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins*, Protein Engineering (1996), Volume: 9, Issue: 1, stranice: 27-36

[10] Schuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, Minoru Kanehisa, *AAIndex: amino acid index database, progress report 2008*, Nucleic Acids Research (2008), Volume: 36, Database issue, stranice: 202-205