MATEMATIČKI FAKULTET
UNIVERZITET U BEOGRADU

Marija Jeličić 1125/2010

Povezanost dužine epitopa i uređenosti delova proteina

Master rad

Mentor: dr Nenad Mitić

Beograd, 2012.

Sadržaj

1. Uvod

U većini slučajeva funkcionalnost proteina zavisi od njegove jedinstvene tercijarne strukture. Međutim, postoje proteini koji nemaju definisanu tercijarnu strukturu (neuređeni proteini) i zahvaljujući tome obavljaju neke važne biološke funkcije. Dosadašnja istraživanja su pokazala da su neuređeni proteini povezani sa ljudskim bolestima kao što su kancer, autoimune bolesti, alergije, neuredegenerativne bolesti i druge. Da bi se bolje razumele ove bolesti potrebno ispitati da li postoji korelacija između neuređenih regiona i antigenih regiona (epitopa) u proteinu.

Preduslov za pokretanje ćelijskog imunog odgovora je da T ćelije prepoznaju peptide (epitope) koji su vezani za MHC molekule. U oblasti bioinformatike ulaže se veliki napor da se što preciznije identifikuju T ćelijski epitopi, što je preduslov za dalje proučavanje imunog odgovora. U ovom radu se kombinuju metode za predviđanje epitopa (na osnovu sekvence proteina) sa metodama za predviđanje uređenosti strukture proteina. Na ovaj način je pokušano da se dođe do odgovora na neka od značajnih imunoloških pitanja kao što su jačina vezivanja epitopa za molekule MHC klase I i II u zavisnosti od uređenosti regiona, učestalosti epitopa u proteinskim regionima sa uređenom odnosno neuređenom strukturom i druga.

Određivanje povezanosti između uređenosti delova proteina i antigenih delova može da doprinese određivanju delova proteina koji su ključni za određene biološke funkcije, a time i boljem tretmanu bolesti, razvoju novih terapija i vakcina.

2. Proteini i njihova struktura

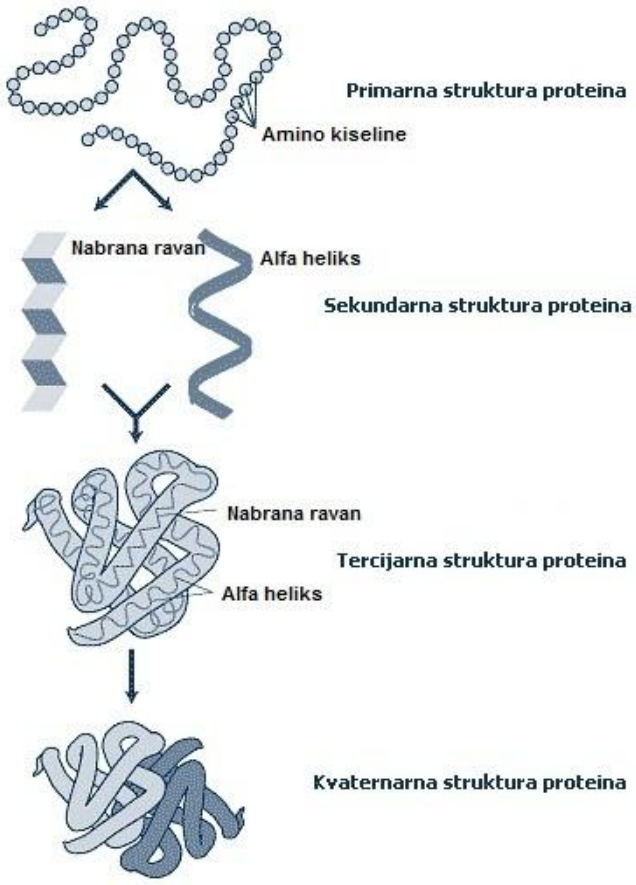
Peptidi i polipeptidi (proteini) su polimeri sačinjeni od aminokiselina koje su međusobno povezane peptidnim vezama. Peptidi su polimeri male ukupne molekulske mase, dok polipeptidi najčešće sadrže više od 100 aminokiselina i njihova molekulska masa je veća. Svaki proteinski polimer je sekvenca koja predstavlja niz sačinjen od 20 različitih L-α aminokiselina.[1] Povezivanje aminokiselina se ostvaruje kovalentnom peptidnom vezom koja nastaje povezivanjem α-karboksilne grupe jedne aminokiseline i α-amino grupe druge aminokiseline pri čemu se oslobađa molekul vode. Tako nastaje nerazgranati polipeptidni lanac (niska) izgrađen od pravilno ponavljane okosnice ili glavnog lanca i međusobno različitih ogranaka. Svaki protein ima jedinstvenu aminokiselinsku sekvencu koja je određena sekvencom nukleotida u odgovarajucem genu. "Standardna" grupa amino kiselina se može podeliti esecijalne i neesecijalne.

[1] Neki proteini u svom sastavu mogu da imaju 22 različite amino kiseline. Pored 20 "standardnih" amino kiselina, postoje i 2 "nestandardne" i to su Selenocistein (eng. *Selenocysteine*, simboli Sec, U) i Pirolizin (eng. *Pyrrolysine*, simboli Pyl, O). Ove dve amino kiseline se ređe javljaju.

1. Esencijalne: Arginin, Histidin, Leucin, Izoleucin, Lizin, Metionin, Fenilalanin, Treonin, Triptofan, Valin
2. Neesencijalne: Alanin, Asparagin, Asparaginska kiselina, Cistein, Glutaminska kiselina, Glutamin, Glicin, Prolin, Serin, Tirozin

Proteini su ključni gradivni elementi svakog živog organizma. Oni ucestvuju u svim ćelijskim i među-ćelijskim procesima. Struktura proteina je određena redosledom aminokiselina u polipeptidnom lancu i od nje direktno zavisi funkcija proteina.

Zbog specifičnog vezivanja lanaca aminokiselina, proteini imaju 4 strukturna nivoa koji određuju njihov izgled u prostoru (konformaciju). Struktura proteina može biti: primarna, sekundarna, tercijarna i kvaternarna.



Slika 1. Strukture proteina

Primarna struktura predstavlja redosled aminokiselina u polipeptidnom lancu. Redosled aminokiselina se održava kompaktnim pomoću kovalentne peptidne veze.

Sekundarna struktura se zasniva na vodoničnim vezama između amido i karboksilne grupe u sekvenci amino kiselina. Sekundarna struktura je lokalna 3D organizacija atoma okosnice polipeptidnog lanca i ne zavisi od konformacije bočnih lanaca amino kiselina. Opisuje prostorni raspored susednih aminokiselina u lancu definisan torzionim uglovima između α-C atoma i C atoma COOH grupe i N atoma NH_2 grupe. Osnovni oblici sekundarne strukture su α-heliks i β-(nabrana) ravan i β-zavoj.

Tercijarna (3D) struktura predstavlja celokupan oblik polipeptida, odnosno trodimenzionalni raspored atoma u jednom proteinu. Tericijarna struktura je zavisna od primarne strukture, jer je u velikoj meri

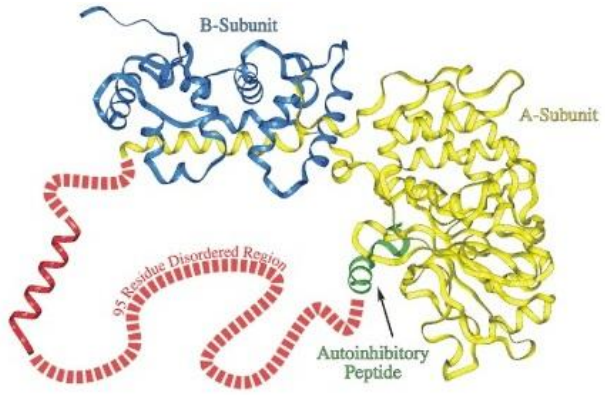
određena redosledom aminokiselina u polipeptidnom lancu. Interakcijom udaljenjih delova u polipeptidnom lancu primarne strukture nastaje tercijarna struktura.

Kvaternarna struktura je nivo organizacije proteina koji je sačinjen od više polipeptida. Svaki polipeptid (sa uređenom tercijarnom strukturom) predstavlja jednu subjedinicu. Kvaternarna struktura je prostorni raspored subjedinica u složenoj celini.

3. Uređenost proteina

Do sredine 20. veka naučnici su verovali da je uređena struktura (odnosno jedinstvenost tercijarne strukture) preduslov za funkcionisanje proteina. Međutim, tokom poslednjih decenija definisana je posebna vrsta proteina koja nije u skladu sa ovim pravilom. Mnogi biološki aktivni proteini nemaju stabilnu tercijarnu i/ili sekundarnu strukturu, što je pokazano na eksperimentalnim podacima u *in vitro* uslovima. Ovi proteini su nazvani suštinski neuređeni proteini (eng. *intrinsically disordered proteins*). Neki autori pri definisanju ovakvih proteina koriste i druge terimne: neuvijeni, delimično neuvijeni, neuređeni, urođeno neuređeni, urođeno denaturisani, pokretni proteini. Ono što ih razlikuje od potpuno uređenih proteina jeste činjenica da u svojoj strukturi sadrže određenu količinu neuređenosti (na određenim delovima ili celom proteinu, slika 2). Takvi proteini pokazuju različite konformacione izomere u kojima se pozicije atoma i torzionih uglova polipeptidne okosnice (kičme) menjaju u toku vremena [1]. Neuređena struktura ovim proteinima omogućava ostvarivanje ključnih funkcija, npr. zauzimanje uređene strukture u kontaktu sa određenim makromolekulom partnerom ili partnerima. Funkcionalni repertoar neuređenih proteina sličan je funkcijama uređenih proteina [2].

Grubom podelom, proteine možemo da svrstamo u dve grupe, proteini sa kompaktnom i proteini sa istegnutom neuređenom strukturom. Prema ovoj klasifikaciji proteini mogu biti manje ili više kompaktni i mogu da poseduju manju ili veću fleksibilnost 2D/3D strukture.



Slika 2. Primer neuređene strukture jednog proteina gde su neuređeni regioni predstavljeni crvenom isprekidanom linijom.

Ipak, neuređenost ne dolazi bez biološke cene. Smatra se da upravo ona ima ulogu u razvoju Alchajmerove i Parkinsonove bolesti, neurodegenerativnih bolesti, kancera kao i autoimunih bolesti od kojih boluju milioni ljudi širom sveta, pa su istraživanja u ovoj oblasti jako važna.

3.1. Korelacija redosleda aminokiselina u proteinu (sekvence aminokiselina) i neuređene strukture

Sposobnost proteina da se "uvija" ili "ne uvija", (eng. *folding*, *unfolding*) odnosno da poseduje uređenu ili neuređenu strukturu zavisi od redosleda aminokiselina u tom proteinu. Na osnovu istraživanja koje je sprovedeno nad 275 uređenih i 91 neuređenih proteina, zaključeno je da kombinacija niske srednje vrednosti hidrofobnosti (usled čega dolazi do lošeg pakovanja proteina) i velike količine neto naelektrisanja (koje dovodi da jakog elektrostatičkog odbijanja) predstavlja važan preduslov za odsustvo kompaktne strukture proteina [3].

Na osnovu osobina kao što su hidrofobnost, naelektrisanje i druge, i učestalosti pojavljivanja amino kiselina u neuređenim proteinima, amino kiseline možemo podeliti u dve grupe:

1. Amino kiseline koje "promovišu neuređenost" (eng. *disorder-promoting*) i to su E (glutaminska kiselina), K (lizin), R (arginin), G (glicin), Q (glutamin), S (serin), P (prolin), A (alanin). Neuređeni proteini su bogati ovim amino kiselinama.
2. Amino kiseline koje "promovišu uređenost" (eng. *order-promoting*) u neuređenim regionima se javljaju u manjem procentu. U ovu grupu spadaju L (leucin), V (valin), W (triptofan), I (izoleucin), Y (tirozin), C (cistein), F (fenilalanin), N (asparagin).

3.2. Prepoznavanje uređenih proteina (proteinskih regiona)

Poznato je 667 neuređenih proteina[2], od kojih su neki u potpunosti neuređeni (i takvo stanje im je prirodno, grupa urođeno-neuređenih proteina), dok preostali imaju neuređene delove različitih dužina. Za prepoznavanje neuređenih proteina, odnosno neuređenih regiona, koriste se različite eksperimentalne metode. Uređeni proteini imaju veliku gustinu i manji radijus okretanja. Suprotno od njih neuređene proteine odlikuje manja gustina, pa se neuređeni proteini mogu otkriti metodama koje su osetljive na molekulske veličine, molekulske gustine ili hidrodinamički otpor. Najčešće se koriste:

- Nuklearno magnetna rezonantna spektroskopija (NMR)
- Difrakciona kristalografija X – zracima
- Cirkularni dihroizam (CD).

Takođe se koriste i:

- Rasipanje X-zraka pri malim uglovima
- Ramanova spektroskopija
- Merenje koeficijenta difuzije

[2] Broj neuređenih proteina dostupnih u Disprot bazi (http://www.disprot.org), verzija 6.0. Proteini su svrstani u različite funkcionalne kategorije. Svi podaci su zasnovani na objavljenim eksperimentalnim podacima.

Više od 20 biofizičkih i biohemijskih metoda je fokusirano na prepoznavanje neuređenih proteinskih regiona, gde svaka metoda različitim informacijama doprinosi razumevanju nestruktuiranog stanja proteina. Ove metode su skupe i vremenski zahevne. Osim toga, često je korisno primeniti više od jedne metode da bi se neuređeni proteini potpuno okarakterisali, jer različite metode otkrivaju različite aspekte neuređenosti. Neuređeni regioni u mnogim proteinima dele neke zajedničke karakteristike, pa se na ovoj činjenici zasnivaju brojni prediktori za prepoznavanje neuređenih proteina. Razvijeno je više od 50 prediktora i veliki broj aplikacija koje koriste ove prediktore.

Prediktori i metode predviđanja se mogu podeliti na 3 pristupa:

1. Pristup "s početka" ili početni pristup (lat. *Ab-initio*) u kom se predviđanje zasniva na samoj sekvenci obično korišćenjem tehnika mašinskog učenja (kao sto je SVM), neuronskih mreža, Bajesovskog klasterovanja, itd. U ovu grupu spadaju RONN, DISOPRED, DisEMBL, VSL2.
2. Pristup zasnovan na obrascima (eng. *template*) u kom se ispituju slične sekvence sa sličnom strukturom. U ovu grupu spadaju prediktori koji su zasnovani na fizičko-hemijskim osobinama aminokiselina u proteinu kao što su PONDR, FoldUnFold, PreLINK, IUpred, FoldIndex.
3. Meta pristup u kom se predviđanje zasniva na kombinaciji nekoliko pristupa (algoritama). Neki od njih su MD, GeneSilico Metadisorder, PONDR-FIT, metaPrDOS.

Ova 3 pristupa je teško uporediti, jer je svaki namenjen i razvijan za specifičnu vrstu neuređene strukture.

Neuređene regione možemo podeliti po njihovoj dužini na 3 grupe [4]:

(a) kratke:1–30
(b) duge: 30–200
(c) veoma duge: duže od 200 amino kiselina.

Neki autori prave i preciznije podele, pa se kratki regioni mogu podeliti u 3 podgrupe:1–3, 4–15, 16–30. Dugi regioni mogu biti podeljeni u dve podgrupe: 30–100 i 101–200.

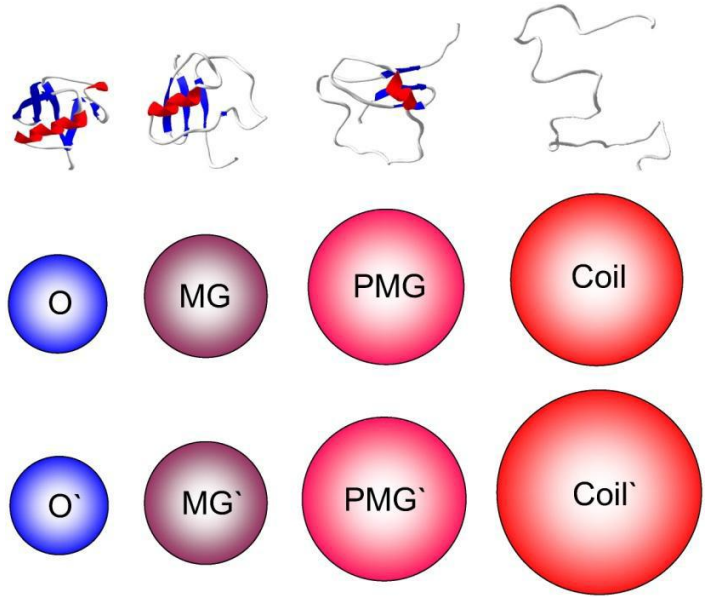
3.3. Funkcije neuređenih proteina

Proteini predstavljaju osnovne komponente u svakoj živoj ćeliji. Oni igraju ključnu ulogu u održavanju života i poznato je da disfunkcija proteina izaziva razvoj različitih bolesti. Proteini poseduju gotovo beskrajan skup bioloških funkcija. Međutim, standardna paradigma sekvenca-struktura-funkcija koja je zasnovana na činjenici da proteini zauzimaju jednu stabilnu 3D stukturi koja uslovljava funkciju proteina, je nevažeća kada su u pitanju neuređeni proteini.

Iako nemaju uređenu 3D strukturu, neuređeni proteini imaju velikog značaj u obavljanju bioloških funkcija. Veruje se da neke biološke funcije zaista zahtevaju fleksibilnost i dinamičnost konformacije koju uređeni proteini nemaju [4]. Struktura proteina može da se menja i javlja se u 4 oblika:

- Uređena struktura (eng. *ordered*)
- Topljiva globula (eng. *molten globule*)
- Pre-topljiva globula (eng. *pre-molten globule*)
- Nasumično klupko (eng. *random coil-like*)

Protein može da obavlja biološke funkcije zauzimajući bilo koju od ove 4 strukture. Neki neuređeni proteini mogu da prelaze iz neuređenog u uređeno stanje i obrnuto nakon interakcija sa drugim makromolekulima ili nakon promena u biohemijskim procesima, dok drugi mogu da obavljaju svoje funkcije i neuređenom obliku [2].



Slika 3. Primeri ilustrovane strukture proteina. Prvi red: uređena struktura, topljiva globula, pre-topljiva globula i nasumično klupko. Slika predstavlja model strukture polipeptidnog lanca dužine 100 amino kiselina. Srednji red: Relativna hidrodinamička sfera koju zauzimaju polipeptidni lanci dužine 100 amino kiselina u ove 4 strukture. Treći red: Relativna hidrodinamička sfera koju zauzimaju polipeptidni lanci dužine 500 amino kiselina u ove 4 strukture. Poređenjem drugog i trećeg reda uočava se uvećenje sfere u odnosu na uređenu strukturu.

Normalna fiziologija i funkcija svakog organizma je zasnovana na skupu visoko koordinisanih proteinskih interakcija. Koordinacija je kontrolisana prepoznavanjem jedinstvenog identifikacionog regiona koji se često nalaze unutar neuređenih proteina. Zato su mnogi neuređeni proteini često uključeni u regulaciju, prepoznavanje, signaliziranje i kontrolu raznih događaja u ćeliji. Neophodni preduslovi za obavljanje ovih funcija koje poseduju neuređeni proteini su visoko specifične interakcije niskog afiniteta (sa više partnera) i sposobnost vezivanja i interakcija sa više partnera.

Iako su poznate brojne funkcije neuređenih proteina, njihov funkcionalni potencijal je nedovoljno istražen. Neke od uloga neuređenih proteina su njihovo vezivanje sa drugim molekulima koji su uključeni u brojne procese u ćeliji: regulaciju kontrolnih DNK regiona, aktiviranje enzima, životni vek proteina [1], regulacija ćelijskog ciklusa, membranski transport, molekularno prepoznavanje i signalizacija [5].

Jedna od jedinstvenih funkcionalnih karakteristika neuređenih proteina je sposobnost jednog proteina da se veže sa više partnera (eng. *binding promiscuity*). U protekloj deceniji brojna istraživanja su proizvela rezultate koje potvrđuju značaj protein-protein interakcija i njihov značaj u ćelijskoj signalizaciji. Neuređenost može da omogući da se jedan protein poveže sa više partnera (jedan-više signalizacija) ili da se više partnera poveže sa jednim proteinom (više-jedan signalizacija). Proteini koji mogu da obavljaju više zadataka istovremeno imaju ulogu čvorova (tzv. *habova*, eng. *hubs*) u mreži proteinskih interakcija. Habovi su od ključne važnosti za funkcionisanje i stabilnost protein-protein mreže interakcija u svakom organizmu [3].

Prepoznavanje funkcija neuređenih proteina može biti realizovano preko nekoliko molekulskih mehanizama, koji su često u vezi sa promenom strukture (tranzicijom od neuređene ka uređenoj) izazvanom vezivanjem za svoje partnere. Struktura koju neuređeni protein usvoji u vezanoj formi može biti podstaknuta od strane molekula sa kojim je vezan ili da odražava njegovu prirodnu konformacionu sklonost [6].

3.4. Neuređenost proteina i ljudske bolesti, D^2 koncept

Kao što je ranije spomenuto, neuređeni proteini imaju centralnu ulogu u mreži preteinskih interakcija. Na taj način su uključeni u važne biološke funkcije, pa su mnogi povezani sa raznim ljudskim bolestima. Grupa autora je definisala koncept koji se bavi ovom problematikom i nazvala ga D^2 koncept (eng. *D^2 concept – disorder in diseases*) [7]. Oni su analizirali patološke uloge nekoliko pojedinačnih neuređenih proteina. Prikupili su specifične skupove proteina koji su povezani sa datom bolešću i izvršili računarsku/bioinformatičku analizu tih skupova podataka kotisteći brojne prediktore. Neuređeni proteini su se pojavljivali u grupama proteina koji su bili povezani sa kancerom, kardiovaskularnim bolestima, neurodegenerativnim bolestima, amiloidozom, dijabetesom i nekoliko drugih bolesti. Na osnovu ovog, ali i mnogih drugih istraživanja razvijen je D^2 koncept prema kojem neuređeni proteini imaju ključnu ulogu u različitim ljudskim bolestima. Neuređeni proteini koji su u korelaciji sa ljudskim bolestima su raznovrsni, brojni, dinamični i od vitalnog značaja i baš zato su istraživanja u ovoj oblasti česta i značajna.

U radu [8] grupa autora je sprovela istraživanje nad grupom od 228 kancer-testis antigena (CTA) u cilju predviđanja uređenosti kancer-asociranih proteina. Za predviđanje neuređenosti

primenjeni su Foldindex (zasnovan na prosečnoj hidrofobnosti) i RONN (zasnovan na neuronskim mrežama) algoritmi i u nekim slučajevima metaPrDOS. Na osnovu rezultata predviđanja, svi CTA su klasifikovani u jednu od tri klase:

1. Veoma uređeni (0-10% neuređenosti)
2. Umereno neuređeni (11%-30% neuređenosti)
3. Veoma neuređeni (31%-100% neuređenosti)

Rezultati ovog istraživanja pokazuju da većina CTA (više od 90%) pripada klasi veoma neuređenih proteina bez obzira na metod predviđanja. Uzimajući u obzir prirodu neuređenih proteina, proverene su neke njihove osobine na istim podacima. Tako je potvrđena pretpostavka da se neki CTA vezuju za DNK, pa su tako uključeni u transkripcionu regulaciju ili druge procese kao što su oštećenje/popravka DNK (eng. *demage/repair DNA*). Potvrđeno je i da CTA zauzimaju hab pozicije u proteinskoj mreži interakcija. Ovo predstavlja značajan doprinos, jer daljim istraživanjem i razumevanjem funkcija takvih CTA omogućava se razvoj novih pristupa i terapija za lečenje kancera.

4. Imuni odgovor

Imunost je sposobnost organizma da prepozna patogene (štetne organizme ili molekule) i da ih eliminiše na različite načine. Imuni sistem se aktivira svaki put kada se bilo koje strano telo ili organizam nađu u ljudskom telu. To su najčešće mikroorganizmi (bakterije, virusi, gljivice) ili toksini i takve činioce nazivamo *antigenima* (eng. *antigene*, od prvobitnog *anti*body *gen*erator). Postoje milioni antigena koji među sobom imaju i neznatne razlike, a koje imuni sistem prepoznaje i protiv kojih započinje reakciju zvanu imuni odgovor. Imuni sistem se može podeliti na:

- Nespecifičan – urođeni imuni sistem
- Specifičan – stečeni imuni sistem

Urođeni imuni sistem deluje bez prethodnog susreta organizma sa stranim mikroorganizmima, nema imunološku memoriju i predstavlja prvu liniju odbrane organizma. Razlikuje strane od vlastitih materija, ali ne prepoznaje vrstu stranog agensa.

Stečeni imuni sistem se razvija postepeno nakon rođenja. Podrazumeva da za svaki antigen postoji specifičan imuni odgovor. Aktivan je i dinamičan i ima sposobnost da pamti svaki prethodni kontakt sa određenim antigenom, odnosno da prepoznaje stranu materiju čak i mnogo godina nakon prvog kontakta, pa da pokrene imuni odgovor sa ciljem da je uništi. Imuni odgovor stečenog imunog sistema može biti humoralni (posredstvom *antitela*, odnosno *imunoglobulina*) i ćelijski. Humoralni je usmeren uglavnom protiv bakterija, dok je ćelijski usmeren uglavnom

protiv virusa ili malignih tumora. Ovi imuni odgovori se zasnivaju na različitim komponentama imunog sistema.

- Ključne ćelije imunog sistema su *limfociti*, *antigen-prezentujuće ćelije* (eng. *antigen-presenting* cells, APC) i izvršne (efektorne) ćelije. Limfociti su ćelije koje imaju sposobnost da prepoznaju i odgovore stranom telu, odnosno antigenu i kao takve su posrednici humoralnog i ćelijskog imuniteta. Postoje različite podvrste limfocita koji se razlikuju po funkciji i po tome kako razlikuju antigene. Razlikujemo B limfocite ili B ćelije i T limfocite ili T ćelije. Za prepoznavanje antigena odgovore su dve različite vrste molekula: *B-ćelijski receptori* (eng. *B cell receptor*, BCR) koji su antitela vezana za ćelijsku membranu B limfocita i *T-ćelijski receptori* (eng. *T cell receptor*, TCR) - receptorni molekuli koji se nalaze na ćelijskoj membrani T limfocita.
- Molekul antitela ili receptorski molekuli na B i T ćelijama se obično ne vezuju za ceo molekul antigena već, specifično samo na deo molekula antigena, koji se naziva epitop. Epitop je kratka amino-kiselinska niska, spečifično mesto za koje se vezuje TCR ili određeni imunoglobin. Jedan molekul antigena može imati, i obično ima, više različitih epitopa koji reaguju sa različitim antitelima različite specifičnosti i afiniteta [9]. Na osnovu strukture dele se na linearne i prostorne (konformacione) epitope.
- BCR i TCR imaju mnogo zajedničkih osobina, ali se razlikuju po svojoj strukturi i tipu epitopa za koje se vezuju. Obično T limfociti prepoznaju linearne epitope, a B limfociti konfomacione epitope. Osim toga, načini na koje B limfociti i T limfociti prepoznaju antigene se znatno razlikuju pa se razlikuje i imuni odgovor.

Humoralni imuni odgovor je mehanizam odbrane od mikroorganizama (i njihovih toksina) koji napadaju ćelije spolja da bi se sprečio njihov ulazak u ćeliju. Zasniva se na stvaranju specifičnih antitela protiv određenog antigena. Antitela luče B limfociti koji na svojoj povrsini takođe imaju antitela (B ćelijske receptore). B limfociti prepoznaju uglavnom nelinearne (konformacione) epitope koji su uglavnom locirani na površini molekula i pretežno su hidrofilni.

Ćelijski imunitet je usmeren je na linearne epitope antigena. T limfociti su zaduženi za prepoznavanje proteina poreklom iz sopstvenih ćelija ili stranih izvora, koji su proteolitički obrađeni do epitopa u ćelijama. Ćelijski imutet je značajan je kod virusnih i gljivičnih infekcija, malignih tumora i transplantacije organa. Ovaj imuni odgovor se još naziva i "ćelijama posredovani imunitet" i zasniva se na:

1. pomažućim T limfocitima (eng. *T helper*, Th ili CD4) koji luče citokine da deluju na B limfocite, na druge T limfocite i na zapaljenske ćelije (makrofagi, neutrofili i dr.) i
2. citotoksičnim T limfocitima (eng. *Cytotoxic T*, Tc ili CD8) koji napadaju i uništavaju strane ćelije ili ćelije koje ispoljavaju strane epitope na svojoj površini, kao što su ćelije inficirane virusima ili tumorske ćelije.

Središnja uloga u regulaciji ćelijskog imunog odgovora pripada regulatornim/pomažućim (eng. *regulatory/helper*, Tr ili Th) T limfocitima CD4+ (ili T4). Oni određuju vrstu interakcije među imunokompetentnim ćelijama lučeći različite vrste citokina, određujući tako i specifičnost i mehanizam ćelijske imunološke reakcije:

- određuju specifičnost odgovora, tj. određuju koji antigen ili koji epitop će se prepoznati,
- uključuju se u odabir izvršnog mehanizma koji će se pokrenuti protiv odabranog antigena,
- pomažu proliferaciju odabranih izvršnih ćelija,
- postiču funkcije fagocita i ostalih nespecifičnih ćelija,
- mogu da zaustave imuni odgovor.

Nakon što odrede koji epitop će biti predmet imunološkog prepoznavanja i cilj imunološkog odgovora, limfociti Th "odlučuju" koji će izvršni mehanizam biti najprikladniji za reakciju protiv odabranog antigena.

Bez obzira na vrstu, imuni odgovor se odvija u 3 faze: prepoznavanje antigena, aktivacija limfocita i izvršna (efektorna) faza eliminacije antigena. Imuni odgovor (kod adaptivnog imunog sistema) započinje prepoznavanjem specifičnih antigena. Imuni sistem sisara je evoluirao tako da izlaže delove proteinskih antigena, koji potiču od mikrobnih patogena, efektornim ćelijama imunog sistema [10]. Prepoznavanje nastaje ukoliko je antigen (epitop) obrađen proteolitičkim enzimima i predstavljen pomoću antigen-prezentujućih ćelija. Obrada se odvija tako što APC obuhvate antigen i uvuku ga, zatim se u specijalizovanim organelama vrši obrada antigena koja podrazumeva njegovu denaturaciju, tj. razvijanje i cepanje na kratke peptide. Neki od tih peptida se nekovalentno vezuju za proteine glavnog histokompatibilnog kompleksa (eng. *major histocompatibility complex*, MHC) i predstavljaju na površini APC. T ćelije pomoću svojih receptora, TCR, mogu da prepoznaju antigene, odnosno komplekse (CD3/TCR) formirane pomoću peptida vezanih za molekule proteina glavnog histokompatibilnog kompleksa na površinama ćelija domaćina.

Postoje 2 klase molekula glavnog histokompatibilnog kompleksa, klase I i II (skr. MHC I i MHC II). Kod čoveka ove klase se nazivaju *antigeni ljudskih leukocita*, HLA I i HLA II (eng. *human leukocyte antigens*). Njihove kombinacije predstavljaju individualnu tkivnu i imunološku specifičnost organizma, koja je genetski definisana (genskim alelima klase MHC I i II) [1]. Kod ljudi postoje 5 vrsta HLA I molekula: HLA-A, HLA-B, HLA-C, HLA-E, HLA-G i 3 vrste HLA II molekula: HLA-DP, HLA-DQ, HLA-DR.

Izvršne (efektorne) ćelije, odnosno T limfociti CD4+ i CD8+ prepoznaju jednu od klasa MHC i to:

- CD8+ T limfociti prepoznaju epitope vezane u kompleksu sa molekulima MHC I klase
- CD4+ T limfociti prepoznaju epitope kao polipeptide vezane u kompleksu sa molekulima MHC II klase.

U zavisnosti od porekla antigena, obrada antigena može se odvijati kroz jedan od dva glavna puta (endocitozni i citosolični put), koji određuju imunski odgovor.

Antigen	Endocitozni put	Citosolični put
Glavni izvori antigena	Endocitozovani ekstracelularni proteini (domaći i strani) Membranski proteini (domaći i strani)	Proteini iz citosola domaćina ili intracelularnih patogena (virusa, bakterija, parazita) Signalni peptidi (domaći i strani)
Obrada antigena	Lizozomalni enzimi	Proteozomi
Aktivirani tip ćelije	Profesionalne APC	Sve ćelije sa jedrom
Mesto vezivanja antigen-MHC	Endocitozne vezikule, prelizozomi	Granuralni endoplazmatski retikulum
Korišćeni MHC	Klasa II	Klasa I
Prezentovan za	CD4+ (pomažuće) T ćelije	CD8+ (citotoksičke) T ćelije

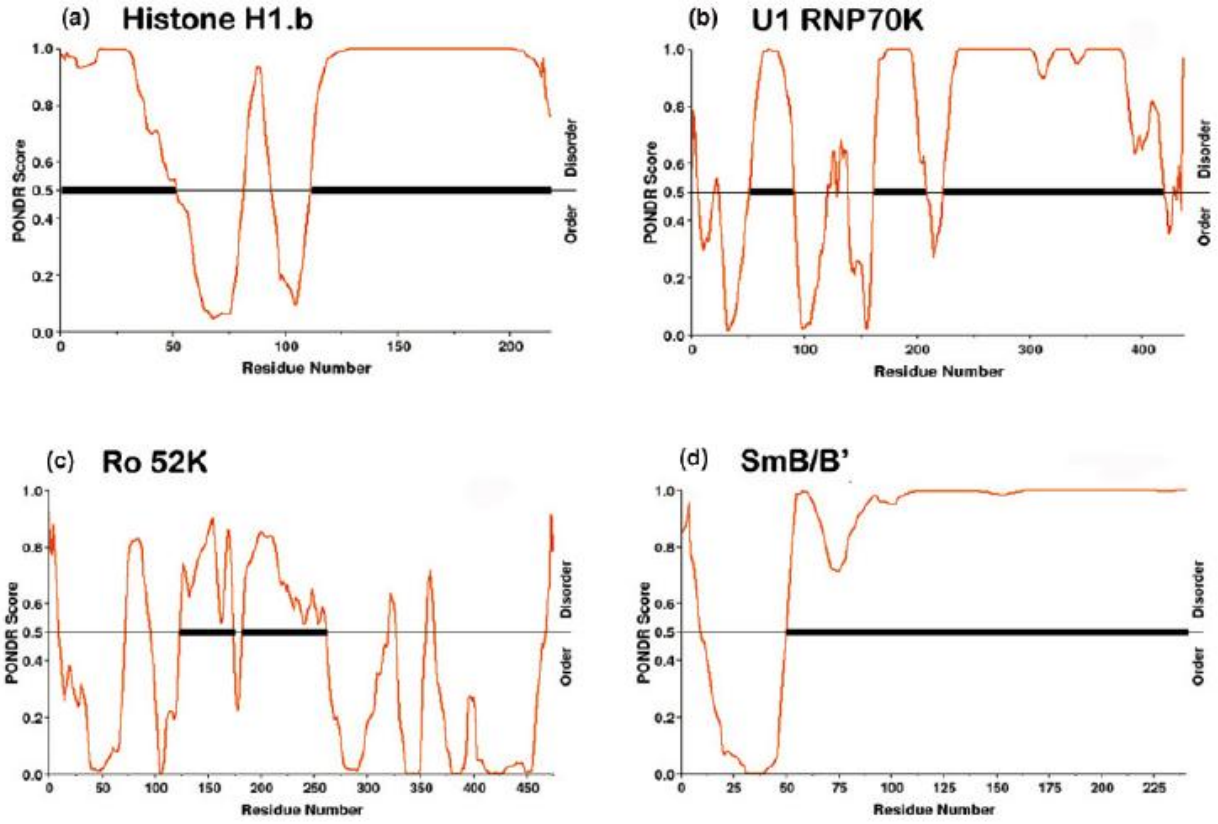
Tabela1. Putevi obrade antigena

4.1. Neuređeni proteini, autoantigeni i tumor-asocirani antigeni (TAA)

Za neuređene regione proteina se može predpostaviti da predstavljaju slabe antigene. Oni mogu da zauzmu više konformacija i fleksibilni su, pa je teško da će izazvati stvaranje B-ćelijskog imunogog odgovora (antitela) specifičnog za određenu konformaciju (odnosno region proteina koji ima određen prostorni raspored). Neuređeni regioni su, stoga, loši aktivatori B ćelija, jer većina B ćelija prepoznaje diskontinualne epitope u uređenoj 3D strukturi antigena. Neuređeni regioni su, takođe, osetljivi na dejstvo proteaza i pretpostavlja se da bi epitopi iz ovih regiona imali slab afinitet za MHC II molekule i da bi zbog toga bili loše prikazani kao T-ćelijski epitopi [11]. Neuređeni proteini se vezuju za druge proteine ili za nukleinske kiseline, pa mogu biti maskirani i fizički nedostupni za imuni sistem, odnosno za T ćelije i B ćelije. U određenim slučajevima ovakvi regioni postaju "vidljivi" i tada nastaje autoimuna bolest.

Autoimune bolesti nastaju kao posledica gubitka imunološke tolerancije organizma na sopstvene antigene (tzv. *autoantigene*) i tada se proizvodi povišen nivo antitela na ove proteine ili komplekse proteina i drugih makromolekula [11]. Još uvek nije poznato zašto neki proteini postaju autoantigeni. Veruje se da na to utiču razni faktori koje možemo svrstati u četiri grupe: strukturne osobine, katabolizam i sudbina ćelije nakon smrti, koncentracija i mikrookruženje, i imunološka i inflamatorna svojstva. U radu [9] autori su se bavili strukturnim osobinama među kojima su visoko šaržirani elementi, ponavljajući površinski elementi, vezane nukleinske kiseline i struktura uvrnutog klupka (eng. *coiled coil*). Ispitujući strukturne osobine autoantigena, došli su

do zaključka da je većina sistemskih nuklearnih (lociranih u jedru ćelije) autoantigena neuređena. Predviđanje je sprovedeno nad grupom autoantigena, ukupno 51, gde je 76% ekstremno neuređeno. Za predviđanje neuređenosti autori su koristili PONDR (VL-XT) prediktor i pokazali su da većina sistemskih nuklearnih autoantigena sadrži dugačke ekstremno neuređene regione i da oni mogu da se preklapaju sa epitopima, odnosno delovima autoantigena (slika 4). Za nekoliko primera je dokazano da u tako dugačkim neuređenim regionima skoro da nema (ili uopšte nema) epitopa koji se vezuju za molekule MHC-II klase. B ćelije koje su usmerene na epitope u neuređenim regionima i T limfociti koji kooperiraju sa njima, su slabo zastupljeni u ukupnom repertoaru imunog odgovora, i zbog slabog afiniteta za ove epitope, mogu da izbegnu imuno isecanje (eng. *deletion*) [9].



Slika 4. Grafički predstavljenji PONDR rezultati za 4 autoantigena: (a) histone H1b; (b) U1 RNP70K; (c) Ro 52K; (d) SmB/B. Tamno crne horizontalne linije ukazuju na regione koji sadrže 39 ili više uzastopnih aminokiselina koje su predviđanjem klasifikovane kao neuređene sa rezultatom većim od praga 0.5 (eng. *treshold*).

Ključni događaj za izazivanje autoimune bolesti je širenje epitopa čijoj pojavi doprinose i T i B ćelije, a odnosi se na razvoj imunog odgovora protiv drugih epitopa na autoantigenu ili makromolekularnom kompleksu. Kod zdravog imunog sistema širenje epitopa je važan funkcionalni element i ima ulogu u zaštiti protiv infektivnih agenasa. U autoimunitetu širenje epitopa najčešće započinje molekularnom mimikrijom ili unakrsnom-reaktivnošću, između

određenog mikrobnog epitopa i epitopa proteina domaćina [9]. Ono predstavlja proširenje imune reaktivnosti od početnog regiona jake antigenetičnosti duž polipeptida na drugi epitop, ili sa epitopa jednog polipetida na drugi (najčešće susedni) polipeptid. Širenje epitopa može da dovede do bržeg i intenzivnijeg sekundarnog odgovora kao i do dugotrajnije imunološke memorije. U radu [9] Carl i sar. su pretpostavili da širenje epitopa započinje u uređenim regionima (epitopima) i može da se proširi na neuređene regione i tako izazove autoimunu bolest.

Antigeni povezani sa tumorima (tumor-pridruženi antigeni, eng. *tumor associated antigens*, TAA) su, kao i autoantigeni, proteini istog organizma, na koje je moguć imuni odgovor usled spontanog (u patogenim uslovima) ili namerno izazvanog (u svrhu lečenja) gubitka imunološke tolerancije organizma na ove antigene. Eksperimentalne i bioinformatičke studije su pokazale da su neuređeni proteini značajno zastupljeni kod TAA (poglavlje 3.4). Neki od njih, kao što je protein p53, koji reguliše popravke DNK, zaustavljanje ćelijskog ciklusa i ćelijsku smrt, su veoma intenzivno proučeni, i nađeno je da postoji povezanost između pojedinih regiona i multifunkcionalnosti proteina [3] [5]. Dokazano je da je protein p53 imunogen kako u svojim mutiranim oblicima, tako i u nemutiranom obliku, zbog toga što dolazi do njegove povišene ekspresije (ispoljavanja) kod tumora [12]. Povišena ekspresija proteina povezana je sa njihovom neuređenošću, jer su neuređeni regioni proteina skloni promiskuitetnim interakcijama pri povšenoj koncentraciji, što može dovesti do patoloških promena [8]. Patogeni efekat CTA, povezan je sa njihovom povišenom ekspresijom u uznapredovalim stadijumima kancera, što je i dovelo do otkrića da ovi antigeni većinom proteini sa visokom procentom neuređenosti [8]. CTA su potencijalni ciljni molekuli u imunoterapiji tumora, zbog toga što se, kao što je gore navedeno, ispoljavaju na nepravilan način u tumorima, dok je njihovo ispoljavanje u normalnom tkivu vezano samo za imunoprivilegovana tkiva, kao što su testisi, ovarijumi i u nekim slučajevima, placenta. Bioinformatička analiza pojedinih familija visoko imunogenih CTA, kao što su antigeni koji pripadaju MAGE-A i NY-ESO familijama je pokazala slaganje između predviđanja epitopa i uređenih struktura proteina, kao i visoki procenat slaganja sa eksperimentalno potvrđenim epitopima koji pripadaju antigenima ovih grupa. Jedan od zadataka ovog rada biće i utvrđivanje odnosa između predviđenih uređenih i neuređenih regiona svih do sada poznatih CTA i predviđanje epitopa.

5. Predviđanje T-ćelijskih epitopa

Kao što je ranije objašnjeno T limfociti na svojoj površini poseduju receptore koji služe za prepoznavanje proteina MHC kompleksa. T-ćelijski epitopi su peptidi koji su vezani za MHC molekul i zajedno čine kompleks koji se nalazi na površini antigen-prezentujućih ćelija, a koji prepoznaju T limfociti. Prepoznavanje epitopa je ključni događaj u ćelijskom imunom odgovoru. Ovi peptidi mogu da potiču od unutarćelijskih (nalaze se u kompleksu sa MHC klase I) i vanćelijskih proteina (nalaze se u kompleksu sa MHC klase II). U prvom slučaju se aktiviraju

citotoksični T limfociti (CD8) što dovodi do ubijanja zaraženih ćelija, dok se u drugom slučaju aktiviraju pomažući T limfociti (CD4), koje nemaju citotoksičnu i fagocitnu aktivnost, već imaju ulogu u pokretanju kontrole imunog odgovora i usmeravanju drugih ćelija imunog sistema. Identifikacija T-ćelijskih epitopa je važna za razumevanje bolesti i predstavlja važnu komponentu u razvoju vakcina i imunoterapija. Računarske metode za predviđanje koje su razvijene u te svrhe daju značajne rezultate i doprinose u dijagnostici i poboljšanju lečenja zaraznih bolesti, alergija, autoimunih bolesti i kancera. Do danas je razvijeno mnogo računarskih metoda i modela koji se razlikuju u tehnikama i algoritmima koje primenjuju, kao i po prirodi podataka za koje se vrše izračunavanja. Sve metode za predviđanje epitopa se mogu podeliti u dve grupe:

1) Metode zasnovane na sekvencama proteina su uglavnom usmerene na kontinualne epitope. Ovakve metode se zasnivaju na identifikaciji obrazaca u sekvenci za koju je poznat afinitet za određeni tip MHC i da bi ovaj metod dobro radio, neophodna je velika količina podataka za trening. Npr. NetMHCpan metode, koje se koriste za predviđanje kontinualnih epitopa, sadrže informaciju o strukturnim karakteristikama MHC molekula u vidu pseudosekvence alela (nepovezane aminokiseline iz različitih delova MHC molekula koje mogu da se vezu za epitop)
2) Metode zasnovane na strukturi proteina, uglavnom su usmerene su na diskontinualne epitope. Ovakve metode su opštije, ali su dosta sporije i ograničene zbog nedostatka podataka o 3D strukturi proteina.

U prvu grupu spadaju procedure koje su zasnovane na izdvajanju motiva, kvantitativnim matricama, stablima odlučivanja, veštačkim neuronskim mrežama (eng. *artificial neural networks*, ANNs), skrivenim Markovljevim modelima (eng. *Hidden Markov models*, HMM) i metodama podržavajućih vektora (eng. *Support vector machine,* SVM), dok u metode druge grupe spadaju *Protein threading, Homology modeling, Docking*. Neke karakteristike procedura na kojma se zasnivaju prediktori koji se koriste u radu su:

Izdvajanje motiva sekvence. Otkrićem da su peptidi koji se vezuju za određene MHC molekule funkcionalno srodni i dele amino kiseline sa sličnim osobinama na različitim pozicijama primarne sekvence, otpočela su najranija predviđana koja su zasnivana na izdvajanju motiva iz proteinskih sekvenci. Primer programa za predviđanje epitopa koji koristi ovu metodu je SYFPEITHI koji pronalazi peptide koji zadovoljaju osobine motiva (smeštenih u bazi podatala) koji se vezuju za neku od MHC klasa. Ovakav pristup ima dosta nedostataka, a najveći problem je to što je ovaj metod deterministički: peptid je ili "vezujući" ili nije "vezujući". Upoređivanjem sa (nedovoljno odgovarajućim) motivom mogu da se dobiju ishodi: *False Positive* i *False Negative*, a ovakvi ishodi nisu pokriveni u metodi izdvajanja motiva sekvence.

Kvantitativne matrice. Matrice povezanosti (eng. *Binding matrices*) predstavljaju unapređenje prethodnog pristupa gde se konstruišu matrice dimanzija l x 20, gde l predstavlja veličinu

peptida dok je druga dimenzija veličine 20 da bi se predstavili simboli svih amino kiselina. Matrice se konstruišu sabiranjem broja pojavljivanja svake amino kiseline na različitim pozicijama u peptidima već poznatim kao epitopi. Primeri programa zasnovanih na ovoj metodologiji su EpiMatrix i BIMAS. Na osnovu kvantitativnih matrica razvijeni su i složeniji modeli koji otkrivaju slabe "vezujuće obrasce" (motive) i daju izveštaj o šumu i kolinearnim podacima.

Stabla odlučivanja. Ovi modeli su zasnovani na pravilima koja klasifikuju obrasce koristeći sekvence sa već poznatim, dobro ustanovljnim, pravilima. Motivi sa specifičnom pozicijom se konvertuju u pravila koja se postavljaju u čvorove stabla. Struktura rezultujućeg stabla je takva da ukazuje na osobine aminokiselina koje su u čvrstoj vezi sa fizičko-hemijskim osobinama vezujućih peptida. Shodno pravilima pretrage u stablu odlučivanja, predviđanje za zadati peptid se odvija polazeći od korena staba naniže-kroz čvorove (pravila), a kao rezultat daju predviđanje za ceo peptid u listu. Stabla odlučivanja mogu da se primene i na linearne i nelinearne podatke, pa se na ovoj metodologiji zasniva veliki broj programa za predviđanje epitopa. Primer programa koji predstavlja inplementaciju ovog metoda je BONSAI.

Veštačke neuronske mreže. Modeli zasnovani na veštačkim neuronskim mrežama su pogodni za klasifikaciju i prepoznavanje složenih obrazaca. ANNs mogu da kodiraju nelinearne podatke i koriste se za predviđanje peptida koji se povezuju i za MHC klasu I i II. Karakteristike peptida su predstavljene pomoću deskriptora amino kiselina, kao što su: kompozicija, hidrofobnost, naelektrisanje. Deskriptori se koriste za treniranje ANN za klasifikovanje peptida na "vezujuće" i "nevezujuće" (eng. *binders, nonbinders*) tj. one koji se vezuju i one koji se ne vezuju sa nekom od MHC klasa. ANN metod je pokazao znatno bolje rezultate od ostalih metoda. Glavni nedostatak ovog pristupa je što zahteva ulaz fiksne dužine, pa određen ANN model može da vrši predviđanje peptida koji su iste dužine kao peptidi kojima je treniran model. Programi koji su korišćeni u ovom radu, NetMHCpan i NetMHCIIpan, su zasnovani na veštačkim neuronskim mrežama.

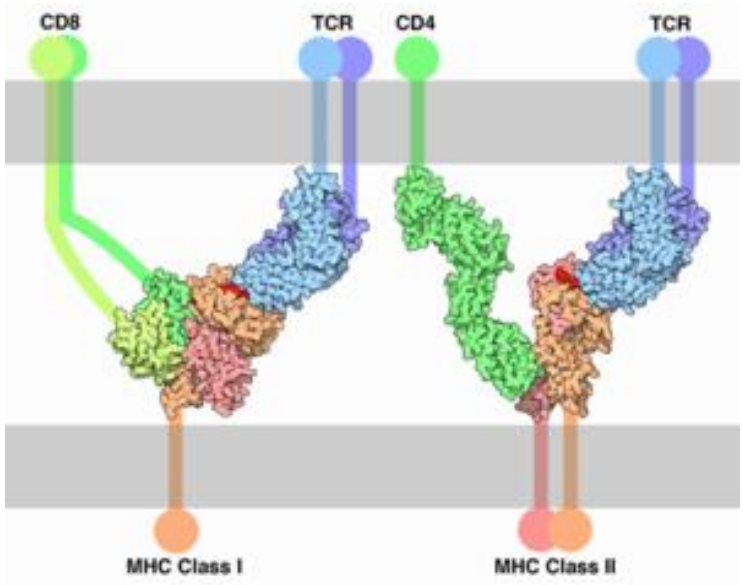
HMM. Predstavlja verovatnosni, grafički model, koji se često primenjuje u aplikacijama čiji je zadatak da sa velikom tačnošću prepoznaju statičke obrasce i klasifikuju statičke podatke. Slično stablima odlučivanja i ANN, HMM dobro radi i sa nelinearnim podacima. Prednost ovog modela je to što je pogodan i za ulaze različitih dužina. HMM modeli su razvijeni u cilju prevazilaženja nedostatka metoda zasnovanih na veštačkim neuronskim mrežama.

SVM. SVM su metode statističkog učenja zasnovane na principu minimizovanja strukturalnog rizika. Slično prethodnim, pogodne su i za linearne i nelinearne podatke. Svaki peptid predstavlja vektor specifičnih karakteristika, kao što su: kompozicija amino kiselina, hidrofobnost, polarnost, naelektrisanje itd. Parametri se treniraju preslikavanjem ulaznih vektora u više-dimenzioni prostor (zavisi od broja karakteristika), zatim se maksimizira granica između

“vezujućih” i “nevezujućih” peptida sa optimalnom razdvajajućom hiper ravni. SVM modeli imaju bolje performanse od ANN i stabala odlučivanja kada su podaci za trening manji.

5.1. Predviđanje vezivanja peptida za molekule MHC klase I i II

Vakcine su najefikasnije sredstvo za borbu protiv zaraznih bolesti. Takođe predstavljaju perspektivne terapije za lečenje raka, alergija i autoimunih bolesti. Cilj vakcinacije je da podstakne imuni sistem u borbi protiv patogena i ćelija raka. Zato je važno da se analizira veza između domaćina i patogena, odnosno uticaj patogena na imuni sistem. Ranije je objašnjeno da se MHC molekuli vezuju za kratke peptide iz antigena i prikazuju ih na površini T-ćelijama. Mehanizam vezivanja je za sada najselektivniji korak u identifikaciji T-ćelija. Razvijeni su razni softveri za predviđanje vezivanja sa MHC klasom I i MHC klasom II. Predviđanje vezivanja sa MHC klasom I je do sada dobro proučeno, tako da su metode koje predviđaju vezivanje sa klasom I velike tačnosti, čak do 95%. Predviđanje vezivanja antigenih epitopa sa MHC klasom II je nešto slabije tačnosti i još uvek nedovoljno istraženo. Razlog tome je što se MHC klase razlikuju po strukturi i po načinu predstavljanja peptida. Antigeni prezentovani klasom II su duži, obično imeđu 15 i 24 amino kiseline, dok je za klasu I dužina peptida između 8 i 11.



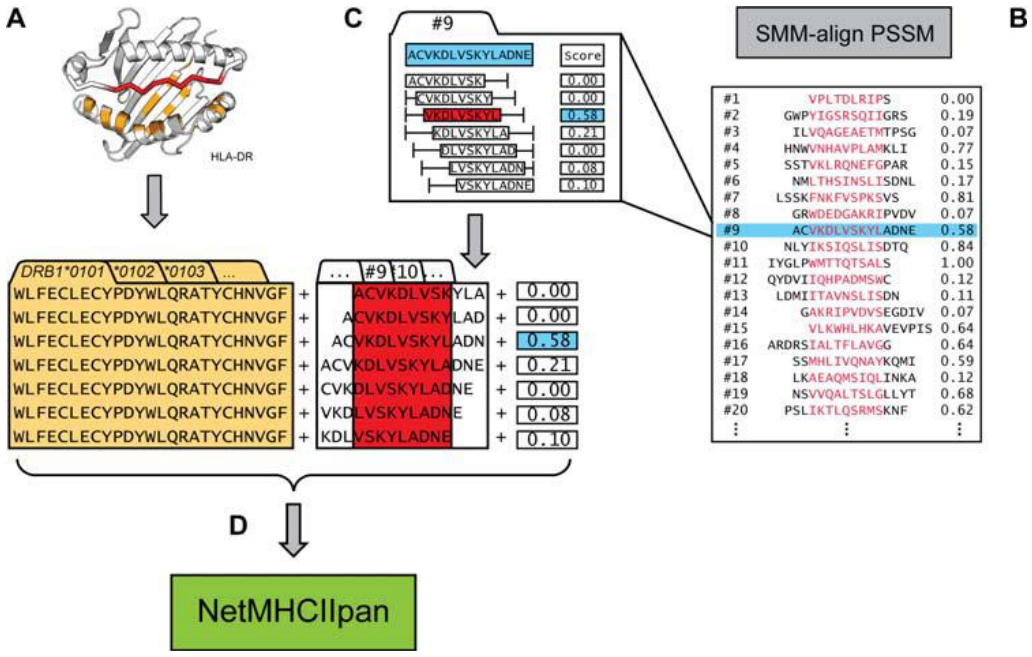
Slika 5. Razlike u strukturi molekula MHC klase I (levo) i MHC klase II (desno)

U ovom radu za predviđanje vezivanja epitopa za molekule MHC klase I je korišćen program NetMHCpan 2.4 [13], a molekule MHC klase II NetMHCIIpan 2.0 [14]. Programe je razvila grupa naučnika iz CBS grupe (eng. *Center of Biological Sequence Analysis* Grupa Tehničkog

Univerziteta u Kopenhagenu, Danska) i zasnivaju se na veštačkim neuronskim mrežama. Ovi programi predviđaju epitope za sve postojeće poznate ljudske alele i daju odličnu tačnost predviđanja epitopa za različite grupe proteina. Predviđanje epitopa koji se vezuju za molekule MHC klase I sa visokom tačnošću je moguće jer je dosupan dovoljno veliki skup ekperimentakno karakterisanih motiva koji predstavljaju podatke za trening. Za klasu II nije moguć pristup koji se se oslanja na specifičnost sekvenci jer su kvantitativni podaci o vezivanju dostupni samo za mali broj alela, pa je teško prepoznati neke pravilnosti. Takođe, MHC molekuli klase II mogu da se vezuju za peptide veoma različitih dužina. Preduslov za predviđanje vezivanja za klasu II sa visokom tačnošću je precizno poravnanje jezgra vezujućeg peptida sa HLA vezujućom "pukotinom" i ovakav pristup se koristi u programu NetMHCIIpan 2.0.

NetMHCpan 2.4 metoda je trenirana na više od 115 000 kvantitativnih MHC vezujućih podataka i pokriva više od 120 različitih MHC molekula. Predviđanje je moguće za HLA-A, HLA-B, HLA-C, HLA-G i HLA-E alele kao i za ne-ljudske primate, miša, svinju, majmuna. Predviđanje je moguće za peptide dužine 8-11. Većina HLA molekula ima snažnu sklonost da se vezuju za peptide dužine 9, pa se predviđanja za peptide ostalih dužina dobijaju aproksimacijom vrednosti dobijene za peptid dužine 9.

NetMHCIIpan 2.0 je metoda koja predviđa vezivanje peptida sa više od 500 HLA-DR alela. Predviđanje je moguće za peptide dužine 9-15 u okviru kojih se nalazi jezgro dužine 9. Početna pozicija jezgra ne mora da se poklapa sa početnom pozicijom peptida (osim za peptid dužine 9). Šematski prikaz NetMHCIIpan metoda je prikazan na slici 6.



Slika 6. Šematski prikaz NetMHCIIpan metoda

Oba programa se primenjuju na proteinsku sekvencu u FASTA formatu. Zadati protein se ”skenira” pomoću “kliznog” prozora veličine 9, počevši od prvih 9 amino kiselina i pomerajući se sa leva na desno. Mogu se zadati različiti parametetri:

- Dužina peptida
- Alela za koju se traži vezivanje (program radi za pojedinačnu, ali i za više alela odjednom)
- Prag vrednosti za “slabo vezujuće” i ”jako vezujuće” peptide (eng. *Weak Binders, Strong Binders*, WB, SB)
- Drugi parametri vezani za prikaz rezultata dobijenih programom

Za svaki peptid se daje kvantitavna ocena afiniteta vezivanja za određeni MHC molekul. Program je vremenski zahtevan, jer ispituje afinitet za sve peptide određene dužine (dobijene pomeranjem prozora za jedno mesto u svakom koraku). Samo mali broj peptida se vezuje za molekul MHC klase i takav peptid je epitop. Peptidi se klasifikuju u jednu od tri kategorije: jaki epitopi, slabi epitopi i peptidi koji nisu epitopi (ne vezuju se za MHC molekule). Takođe se određuje i mera predviđanja koja se računa uz pomoć afiniteta pomoću formule 1-logk(aff), čime se skalira vrednost afiniteta na intervalu [0,1]. Na slikama 7 i 8 su prikazani primeri izlaza programa NetMHCpan 2.4. i NetMHCIIpan 2.0.

pos	HLA	peptide	Identity	1-log50k(aff)	Affinity(nM)	%Rank	BindLevel
0	HLA-A*01:20	MSATTACW	gi_156547151_re	0.373	880.12	3.00	
1	HLA-A*01:20	SATTACWP	gi_156547151_re	0.034	34480.66	50.00	
2	HLA-A*01:20	ATTACWPA	gi_156547151_re	0.173	7659.20	15.00	
3	HLA-A*01:20	TTACWPAF	gi_156547151_re	0.508	204.83	1.00	<= WB
4	HLA-A*01:20	TACWPAFT	gi_156547151_re	0.074	22538.21	50.00	
5	HLA-A*01:20	ACWPAFTV	gi_156547151_re	0.082	20557.29	32.00	
6	HLA-A*01:20	CWPAFTVL	gi_156547151_re	0.090	18896.73	32.00	
7	HLA-A*01:20	WPAFTVLG	gi_156547151_re	0.063	25241.95	50.00	
8	HLA-A*01:20	PAFTVLGE	gi_156547151_re	0.020	40354.32	50.00	
9	HLA-A*01:20	AFTVLGEA	gi_156547151_re	0.062	25607.80	50.00	
10	HLA-A*01:20	FTVLGEAR	gi_156547151_re	0.138	11198.23	16.00	
11	HLA-A*01:20	TVLGEARG	gi_156547151_re	0.022	39270.43	50.00	
12	HLA-A*01:20	VLGEARGD	gi_156547151_re	0.013	43454.09	50.00	
13	HLA-A*01:20	LGEARGDQ	gi_156547151_re	0.023	39122.32	50.00	
14	HLA-A*01:20	GEARGDQV	gi_156547151_re	0.047	30163.32	50.00	
15	HLA-A*01:20	EARGDQVD	gi_156547151_re	0.012	44121.99	50.00	
16	HLA-A*01:20	ARGDQVDW	gi_156547151_re	0.026	37596.71	50.00	
17	HLA-A*01:20	RGDQVDWS	gi_156547151_re	0.082	20479.87	32.00	
18	HLA-A*01:20	GDQVDWSR	gi_156547151_re	0.032	35546.57	50.00	
19	HLA-A*01:20	DQVDWSRL	gi_156547151_re	0.086	19806.95	32.00	
20	HLA-A*01:20	QVDWSRLY	gi_156547151_re	0.702	25.09	0.15	<= SB
21	HLA-A*01:20	VDWSRLYR	gi_156547151_re	0.050	29195.07	50.00	
22	HLA-A*01:20	DWSRLYRD	gi_156547151_re	0.027	37133.28	50.00	

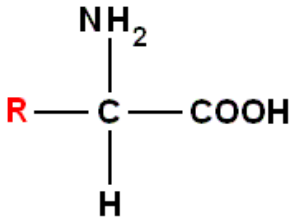
Slika 7. Primer izlaza programa NetMHCpan 2.4. za protein čija je identifikacija GI:156547151, za alelu HLA-A*01:20 i dužinu peptida 8. Program je generisao ukupno 71 peptid dužine 8 (na slici nije prikazan ceo izlaz), među kojima su po jedan jak i slab epitop

pos	HLA	peptide	Identity	Pos	Core	1-log50k(aff)	Affinity(nM)	%Rank	BindLevel
0	DRB1*0114	MNFYLLLAS	gi_63025190_ref	0	MNFYLLLAS	0.083	6749.56	32.00	
1	DRB1*0114	NFYLLLASS	gi_63025190_ref	0	NFYLLLASS	0.257	1270.82	4.00	
2	DRB1*0114	FYLLLASSI	gi_63025190_ref	0	FYLLLASSI	0.519	102.25	0.20	<= WB
3	DRB1*0114	YLLLASSIL	gi_63025190_ref	0	YLLLASSIL	0.413	283.81	0.80	<= WB
4	DRB1*0114	LLLASSILC	gi_63025190_ref	0	LLLASSILC	0.092	6206.87	32.00	
5	DRB1*0114	LLASSILCA	gi_63025190_ref	0	LLASSILCA	0.054	8920.44	32.00	
6	DRB1*0114	LASSILCAL	gi_63025190_ref	0	LASSILCAL	0.031	11146.05	50.00	
7	DRB1*0114	ASSILCALI	gi_63025190_ref	0	ASSILCALI	0.057	8635.83	32.00	

Slika 8. Primer izlaza programa NetMHCIIpan 2.0. za protein čija je identifikacija GI:63035190, za alelu DRB1*0114 i dužinu peptida 9. Program je generisao ukupno 105 peptida dužine 9 (na slici nije prikazan ceo izlaz), među kojima nema jakih, ali postoje dva slaba epitopa

5.2. Hidropatija

Hidropatija je hemijska osobina amino kiselina i odnosi se na njeno ponašanje u vodi. Za svaku amino kiselinu određen je nivo hidrofobnosti i nivo hidrofilnosti. Hidrofobnost je stepen odbojnosti prema vodi, odnosno stepen nerastvorivosti u vodi. Ovakvi molekuli se nazivaju hidrofobi. Suprotno njima, hidrofili se vezuju za molekule vode, pa hidrofilnost predstavlja stepen rastvorivosti u vodi. Hidropatija je važna osobina jer daje informacijo o funkcionisanju ćelijske membrane, koja je selektivno propustljiva.



Slika 9. Opšta formula amino kiselina

Poznato je da se proteini sastoje od više amino kiselina međusobno povezanim peptidnim vezama. Svaka amino kiselina sadrži amino grupu (-NH_2), karboksilnu grupu (-COOH) i različitu R-grupu, amino-kiselinski ostatak (R, eng. *residue*). R grupa određuje da li je protein hidrofoban ili hidrofilan.

Hidrofobnost amino kiseline određuje lokaciju te amino kiseline u konačnoj strukturi proteina, pa su amino kiseline u globularnim proteinima sa hidrofobnom R grupom smeštene unutar proteina i nisu u kontaktu sa vodom. Amino kiseline sa hidrofilnim R grupama su smeštene na spoljašnosti proteina i intereaguju sa vodom u citosolu.

Indeks hidrofobnosti amino kiselina govori o strukturi proteina, koja je u bliskoj vezi sa njegovom funcijom. Zato je ova osobina često uključena kao parametar u raznim alatima za predviđanje strukture proteina. Definisano je nekoliko skala za izračunavanje indeksa hidropatije, a najčešće korišćene su :

- Kajt Dulitl (eng. Kyte-Doolittle) – hifrofobne amino kiseline imaju pozitivnu vrednost
- Hop Vuds (eng. Hopp-Woods) – amino kiseline sa pozitivnom vrednošću su hidrofilne

Ove skale su prikazane na slici 10.

Hidrofobnost / hidrofilnost delova proteina se računa kao srednja vrednost hidrofobnosti svake amino kiseline tog proteina. Računanje se vrši tako što se protein skenira uz pomoć prozora određene veličine. Veličina prozora predstavlja broj amino kiselina i može da varira 5-25. Prilikom skeniranja u svakom koraku prozor se pomera i računa se srednja vrednost indeksa hidropatije. Prozor veličine 5-7 je dobar za pronalaženje hidrofilnih regiona, dok prozor veličine 19-21 pronalazi hidrofobne domene. Na kraju skeniranja rezultati se mogu grafički predstaviti, tzv. *grafikom hidropatije*.

Hydrophobicity Scales

	Kyte-Doolittle	Hopp-Woods
Alanine	1.8	-0.5
Arginine	-4.5	3.0
Asparagine	-3.5	0.2
Aspartic acid	-3.5	3.0
Cysteine	2.5	-1.0
Glutamine	-3.5	0.2
Glutamic acid	-3.5	3.0
Glycine	-0.4	0.0
Histidine	-3.2	-0.5
Isoleucine	4.5	-1.8
Leucine	3.8	-1.8
Lysine	-3.9	3.0
Methionine	1.9	-1.3
Phenylalanine	2.8	-2.5
Proline	-1.6	0.0
Serine	-0.8	0.3
Threonine	-0.7	-0.4
Tryptophan	-0.9	-3.4
Tyrosine	-1.3	-2.3
Valine	4.2	-1.5

Slika 10. Kajt Dulitl i Hop Vuds skale

6. Cilj rada

Cilj ovog rada je analiza odnosa između predviđenih HLA-I i HLA-II epitopa različite dužine i upoređivanje sa predviđenim uređenim i neuređenim regionima proteina, uz korišćenje različitih metoda predviđanja neuređenosti proteina. Analiza odnosa dužine epitopa i uređenosti proteina uključije karakteristike kao što su frekvencija epitopa, afinitet vezivanja za HLA molekule i srednja vrednost hidropatije za epitope. Analiza će biti urađena na grupi tumor-asociranih antigena, koji nose naziv kancer-testis antigeni (CTA) ili podeljeni antigeni (eng. *shared antigens*), tj. antigeni koji se delom pojavljuju u tumorskim ćelijama a delom u normalnim greminativnim ćelijama. Za predviđanje epitopa korišćeni su programi NetMHCpan za HLA I i NetMHCIIpan za HLA II klasu alela. S obzirom da su za predviđanje uređenosti proteina korišćena dva prediktora, VSL2 i IUpred, sva istraživanja su rađena posebno nad dobijenim podacima iz jednog i iz drugog programa i svi rezultati istrazivanja su međusobno upoređeni.

Cilj rada je:

- Ispitati koliko se epitopa nalazi u uređenim, odnosno neuređenim regionima proteina.
- Ispitati zastupljenost slabih i jakih epitopa u neuređenim/uređenim delovima proteina.
- Tehnikama istraživanja podataka, za obe klase MHC I i II, utvrditi da li postoje međuzavisnosti između atributa: uređenost proteina, vrsta epitopa, dužina, alela i peptid (koji predstavlja epitop).
- Za sve amino kiseline ispitati učestalost pojavljivanja na određenoj poziciji u epitopu, u zavisnosti od njegove dužine
- Ispitati na kojim delovima proteina se nalazi najviše epitopa

7. Materijal i metode

Prikupljeno je 143 proteina i svi pripadaju kancer-testis funkcionalnoj grupi proteina. Proteini su preuzeti iz *CTDatabase* baze podataka [15] u kojoj su sakupljeni kancer-testis antigeni. Baza je dostupna na sajtu *Cancer Immunity* žurnala[3]. Na prikupljene proteine primenjeni su programi NetMHCpan i NetMHCIIpan, koji ispituju da li se peptidi proteina vezuju za molekule MHC klase I ili II. Takođe, na proteine su primenjeni i prediktori VSL2 i IUpred za predviđanje uređenosti amino kiselina u proteinu. Većina ovih proteina ima dužinu manju od 1000 amino kiselina, ali ima i nekoliko dužih i za njih je obrada u programima NetMHCpan i NetMHCIIpan vremenski veoma zahtevna.

Družina proteina	Broj proteina
0-100	**10**
101-200	**34**
201-300	**16**
301-400	**31**
401-500	**9**
501-600	**7**
601-700	**7**
701-800	**8**
801-900	**5**
901-1000	**6**
preko 1000	**10**
Ukupno: **143**	

Tabela 2. Podela proteina korišćenih u radu prema broju amino kiselina

Većina postojećih programa za predviđanje neuređenih regiona koristi klizni prozor za preslikavanje individualne amino kiseline u određeni komponentni prostor (eng. *feature space*), gde binarni klasifikator može da klasifikuje simbole kao uređene ili neuređene koristeći različite algoritme mašinskog učenja [4]. Komponente se izdvajaju iz niza amino kiselina kroz prozor koji predstavlja kompozicionu osnovu. Program VSL2 je implementacija modela koji se zasniva na neuronskim mrežama i treniran je nad regionima čija je neuređenost potvrđena metodom kristalografije X-zracima ili metodom NMR.

[3] Cancer Immunity žurnal se objavljuje od strane Instituta za istraživanje raka (eng. *"Cancer Research Institute"*) i pruža informacije iz oblasti kancer imunologije i imunoterapije.

IUPred koristi metod koji se zasniva na proceni kapaciteta polipeptida da formiraju stabilne veze. Osnovna pretpostavka je da proteini globularnog oblika stvaraju veliki broj interakcija izmedju amino kiselina i tako stvaraju energiju koja stabilizuje njegovu (uređenu) strukturu koja je nastala prilikom uvijanja tog proteina. Suprotno tome, neuređeni proteini imaju sekvence koje nemaju kapacitet stvaranja interakcija između amino kiselina, samim tim i manju količinu energije. Količina energije koja se stvara tokom interakcija se može proceniti uzimajući u obzir da doprinos amino kiseline uređenosti/neuređenosti ne zavisi samo od njenog hemijskog tipa, već i od njenih potencijalnih (interaktivnih) partnera. Algoritam podrazumeva računanje koje uključuje matrice 20x20 čiji su parametri izvedeni korišćenjem globularnih proteina, čija je struktura poznata. Poredeći globularne i neuređene proteine uočena je jasna razlika u njihovom energetskom sadržaju [14].

Prikupljeni proteini i svi rezultati obrade proteina su smešteni u tabele relacione baze podataka IBM DB2. Tabele u bazi imaju sledeći izgled:

- Tabela **PROTEIN** sadrži sekvencu proteina i njenu dužinu.

Key	Name	Data type	Length	Nullable
🔑	ID	VARCHAR	20	No
	SEQUENCE	VARCHAR	10000	No
	LENGTH	SMALLINT	2	No

- Tabela **PROTEIN_DETAILS** sadrži zaglavlje u FASTA formatu proteina u kom su navedene detaljnije informacije o proteinu, GI identifikaciju proteina, familija kojoj pripadaju i CT identifikacija.

Key	Name	Data type	Length	Nullable
🔑	PROTEIN_ID	VARCHAR	20	No
	FASTA_HEADER	VARCHAR	500	No
	ORD_NUM_IN_DISPROT	VARCHAR	20	No
	FAMILIJA	VARCHAR	10	Yes
	CT_IDENTIFIKACIJA	VARCHAR	10	Yes

- Tabela **PREDIKTOR** sadrži rezultate prediktora VSL2 i IUpred primenjenih na proteine i predviđenu vrednost uređenja za svaku aminokiselinu u proteinu. Vrednost uređenja može da bude O za amino kiseline u uređenim regionima i D za amino kiseline u neuređenim regionima.

Key	Name	Data type	Length	Nullable
🔑	ID	INTEGER	4	No
	"PROGRAM"	VARCHAR	11	No
	AMINO_ACID	CHARACTER	1	Yes
	START_POS	SMALLINT	2	No
	END_POS	SMALLINT	2	No
	PREDICTION	DECIMAL	7	Yes
	DISORDER	CHARACTER	1	Yes
	PROTEIN_ID	VARCHAR	20	No
	PID	INTEGER	4	Yes

- Tabela **HLA1** (levo) sadrži rezultate dobijene obradom proteina programom NetMHCpan. Dodatno, za svaki peptid je određeno da li pripada uređenom, neuređenom

ili prelaznom regionu na osnovu rezultata dobijenih primenom VSL2 odnosno IUpred prediktora (atributi order_peptid_vsl2 i order_peptid_iupred). Takođe, tabela je proširena atributima allele_sidney i allele_multipred2 u kojima se nalaze podaci o supetipovima alela, koji su objašnjeni na strani 25. Atribut relative_start se odnosi na relativni početak epitopa koji je detaljno objašnjen na strani 37.

- Tabela **HLA2** (desno) sadrži rezultate dobijene obradom proteina programom NetMHCIIpan. Slično prethodnoj tabeli, HLA2 sadrži podatke o uređenosti peptida (order_peptid_vsl2 i order_peptid_iupred) i supertipovima alela (allele_greenbaum i allele_multipred2). U atribute order_jezgro_vsl2 i order_jezgro _iupred su upisani podaci o uređenosi jezgra (za sve dužine). Atribut relative_start se odnosi na relativni pocetak epitopa koji je detaljno objašnjen na strani 37.

Key	Name	Data type	Length	Nullable
🔑	UDALJENJE	SMALLINT	2	No
🔑	ALELA	VARCHAR	14	No
🔑	PEPTID	VARCHAR	11	No
🔑	PID	INTEGER	4	No
	DUZINA	SMALLINT	2	Yes
	LOGAFF	DECIMAL	4	Yes
	AFFINITY	DECIMAL	8	Yes
	RANK	DECIMAL	4	Yes
	BIND	CHARACTER	2	Yes
	ORDER_PEPTID_VSL2	CHARACTER	1	Yes
	ORDER_PEPTID_IUPRED	CHARACTER	1	Yes
	ALLELE_SIDNEY	VARCHAR	12	Yes
	ALLELE_MULTIPRED2	VARCHAR	12	Yes
	RELATIVE_START	SMALLINT	2	Yes

Key	Name	Data type	Length	Nullable
🔑	UDALJENJE	SMALLINT	2	No
🔑	ALELA	VARCHAR	14	No
🔑	PEPTID	VARCHAR	11	No
🔑	PID	INTEGER	4	No
	DUZINA	SMALLINT	2	Yes
	POZICIJA	SMALLINT	2	No
	JEZGRO	CHARACTER	9	No
	LOGAFF	DECIMAL	4	Yes
	AFFINITY	DECIMAL	8	Yes
	RANK	DECIMAL	4	Yes
	BIND	CHARACTER	2	Yes
	ORDER_JEZGRO_VSL2	CHARACTER	1	Yes
	ORDER_PEPTID_VSL2	CHARACTER	1	Yes
	ORDER_JEZGRO_IUPRED	CHARACTER	1	Yes
	ORDER_PEPTID_IUPRED	CHARACTER	1	Yes
	ALLELE_GREENBAUM	VARCHAR	12	Yes
	ALLELE_MULTIPRED2	VARCHAR	12	Yes
	RELATIVE_START	SMALLINT	2	Yes

- Tabela **AMINOACIDS** sadrži sve amino kiseline, njihove nazive, simbole i podatke o hidropatiji.

Key	Name	Data type	Length	Nullable
🔑	NAME	CHARACTER	27	No
	CODE1	CHARACTER	1	No
	CODE3	CHARACTER	3	No
	CODONS	SMALLINT	2	Yes
	CODE3U	CHARACTER	3	Yes
	HIDROTIPKD	CHARACTER	11	Yes
	HIDROFOBNOSTKD	DECIMAL	2	No
	HYDROPHOBICKD	CHARACTER	3	Yes
	HIDROTIPHW	CHARACTER	11	Yes
	HIDROFOBNOSTHW	DECIMAL	2	No
	HYDROPHOBICHW	CHARACTER	3	Yes

- Tabele **ALLELE_SUPERTYPE_HLA1_SYDNEY, ALLELE_SUPERTYPE_HLA1_MULTIPRED2, ALLELE_SUPERTYPE_HLA2_GREENBAUM, ALLELE_SUPERTYPE_HLA2_MULTIPRED2** sadrže podatke o supertipovima alela.

Key	Name	Data type	Length	Nullable
🔑	ALLELE	VARCHAR	20	No
	SUPERTYPE	VARCHAR	12	No

- Tabele **HLA1_AA** (levo) i **HLA2_AA** (desno) su izvedena iz tabela HLA1 i HLA2. HLA1_AA dodatno sadrži atribute AA01, AA02,…, AA11 koji predstavljaju poziciju amino kiseline u peptidu i sadrže amino kiselinu na datoj poziciji. Slično, HLA2_AA dodatno sadrži atribute AA01, AA02,…, AA09 koji predstavljaju poziciju amino kiseline u jezgru i sadrže amino kiselinu na datoj poziciji.

Key	Name	Data type	Length	Nullable
	UDALJENJE	SMALLINT	2	No
	ALELA	VARCHAR	14	No
	PEPTID	VARCHAR	11	No
	PID	INTEGER	4	No
	DUZINA	SMALLINT	2	Yes
	BIND	CHARACTER	2	Yes
	ORDER_PEPTID_VSL2	CHARACTER	1	Yes
	ORDER_PEPTID_IUPRED	CHARACTER	1	Yes
	SUPERTYPE_SIDNEY	VARCHAR	12	Yes
	SUPERTYPE_MULTIPRED2	VARCHAR	12	Yes
	AA01	CHARACTER	1	Yes
	AA02	CHARACTER	1	Yes
	AA03	CHARACTER	1	Yes
	AA04	CHARACTER	1	Yes
	AA05	CHARACTER	1	Yes
	AA06	CHARACTER	1	Yes
	AA07	CHARACTER	1	Yes
	AA08	CHARACTER	1	Yes
	AA09	CHARACTER	1	Yes
	AA10	CHARACTER	1	Yes
	AA11	CHARACTER	1	Yes

Key	Name	Data type	Length	Nullable
	UDALJENJE	SMALLINT	2	No
	ALELA	VARCHAR	14	No
	PEPTID	VARCHAR	15	No
	PID	INTEGER	4	No
	DUZINA	SMALLINT	2	Yes
	POZICIJA	SMALLINT	2	No
	JEZGRO	CHARACTER	9	No
	BIND	CHARACTER	2	Yes
	ORDER_JEZGRO_VSL2	CHARACTER	1	Yes
	ORDER_PEPTID_VSL2	CHARACTER	1	Yes
	ORDER_JEZGRO_IUPRED	CHARACTER	1	Yes
	ORDER_PEPTID_IUPRED	CHARACTER	1	Yes
	SUPERTYPE_GREENBAUM	CHARACTER	12	Yes
	SUPERTYPE_MULTIPRED2	VARCHAR	12	Yes
	AA01	CHARACTER	1	Yes
	AA02	CHARACTER	1	Yes
	AA03	CHARACTER	1	Yes
	AA04	CHARACTER	1	Yes
	AA05	CHARACTER	1	Yes
	AA06	CHARACTER	1	Yes
	AA07	CHARACTER	1	Yes
	AA08	CHARACTER	1	Yes
	AA09	CHARACTER	1	Yes

Alele su klasifikovane u grupe, koje predstavljaju supertipove, a svakom supertipu je zajednička specifičnost vezivanja za MHC molekule. Dakle, svaki supertip odlikuje motiv koji prepoznaju molekuli tog supertipa. Motivi koji se vezuju za klasu I su dobro karakterizovani i do sada je otriveno 88 motiva klase I. Motivi koji se vezuju za klasu II imaju više od jedne hidrofobne amino kiseline što dozvoljava višestruko moguće ravnanje [1], pa je ovakve motive teže opisati i pokrivenost za klasu II je manja. Za klasu I pokrivenost alela supertipovima je veća, ali ne i potpuna, pa neke alele nemaju svoj supertip. Za obe MHC klase su korišćene po dve klasifikacije. Za HLA I alele korićene su sledeće klasifikacije alela u supertipove:

1.1. Klasifikacija koju su u radu [16] predstavili Sidney i sar. Ova klasifikacija pokriva većinu HLA-A i HLA-B alela. Ukupno ima 12 supertipova i oni su označeni sa: A01, A01-103, A01-A24, A02, A03, A24, B07, B08, B27, B44, B58, B62 (i dodatno *Unclassified* za HLA-A i HLA-B alele koje nisu pokrivene postojećim supertipovima). U ovom radu za ovu klasifikaciju koristiće se naziv *Sidney* (prema autoru).

1.2. Klasifikacija dobijena programom MULTIPRED2 [17] za HLA I alele daje 13 supertipova: A1, A2, A3, A24, A26, B7, B8, B27, B44, B58, B62, C1 i C4. U ovom radu za ovu klasifikaciju koristiće se naziv *Multipred2.*

U radu su korišene dve klasifikacije za HLA II alele:

2.1. U radu [18] Greenbaum i sar. su predstavili klasifikaciju koja ima 7 supertipova (main DR, DR4, DRB3, main DQ, DQ7, main DP i DP2) koji pokrivaju HLA-DR, HLA-DQ, HLA-DP molekule. Obzirom da se u radu ne koriste alele HLA-DQ i HLA-DP, u

materijalu za ovu klasifikaciju se javljaju samo 3 supertipa koji se odnose na HLA-DR alele. U ovom radu za ovu klasifikaciju koristiće se naziv *Greenbaum.*

2.2. Klasifikacija programom MULTIPRED2 [17] koja daje 13 supertipova klase II: DR1, DR3, DR4, DR6, DR7, DR8, DR9, DR11, DR12, DR13, DR14, DR15 I DR16. U ovom radu za ovu klasifikaciju koristiće se naziv *Multipred2.*

Početni skup alela sadržao je 2932 HLA I alela i 654 HLA II alela. Eliminasane su alele koje se odnose na ne-ljudske primate miša, svinje, šimpanze i majmuna. Takođe, ustanovljeno je neke alele imaju identičnu pseudo-sekvencu. Takve alele su pronađene i obrisane tako da nema "duplih". U radu je (nakon eliminacije) korišćeno 1568 HLA I alela i 392 HLA II alela. Pokrivenost HLA I alela korišćenih u radu klasifikacijom 1.1. je 45.5%, a klasifikacijom 1.2. je 28.6%. Pokrivenost HLA II alela korišćenih u radu klasifikacijom 2.1. je 3.8%, a klasifikacijom 2.2. je 66.3%.

8. Rezultati dobijeni SQL upitima

Prikupljeno je 143 proteina i svi pripadaju istoj funkcionalnoj grupi (kancer-testis). Raspodela amino kiselina u uređenim i neuređenim regionim za oba prediktora je prikazana u tabeli 3. Poznato je da kancer-testis grupa proteina ima dosta neuređenih regiona, pa je rezultat predviđanja očekivan. Postoji neslaganje u predviđanju uređenosti nad istom grupom proteina programima VSL2 i IUpred, što je takođe očekivano obzirom da se zasnivaju na različitim pristupima.

Broj amino kiselina (u %)		
	Uređeni regioni	Neuređeni regioni
VSL2	**42.84**	**57.16**
IUpred	**66.60**	**33.40**

Tabela 3. Raspodela amino kiselina u uređenim/neuređenim regionima

Rezultati programa NetMHCpan i NetMHCIIpan daju veliki broj peptida čija ocena afiniteta nije zadovoljavajuća, odnosno takvi peptidi nisu epitopi. U bazu podataka su upisani samo peptidi čiji je *Bind level* WB (eng. *Weak Binders*) ili SB (eng. *Strong Binders*) i takvih je 6192797 za MHC klase I. Peptidi koji se vezuju za molekule MHC klase II su brojniji i ima ih 24460520, što je očekivano obzirom na broj prozora različitih dužina. Distribucija slabih i jakih epitopa različe dužine je data u tabelama 4. i 5.

Procenti u tabeli se računaju u odnosu na ukupan broj epitopa za određenu dužinu. Epitopi koji se vezuju za klasu I su dosta jednako raspoređeni u odnosu na dužinu pri čemu su epitopi dužine 9 neznatno brojniji. Broj epitopa koji se vezuju za klasu II raste sa dužinom prozora, što može da bude posledica činjenice da u peptidima čija je dužina veća od 9 predviđanjem može da se pronađe dva ili više epitopa (jezgra).

Za obe klase HLA molekula je veći je broj slabih epitopa, u odnosu na broj jakih, koji se za njih vezuju. Ova pravilnost je očekivana i važi za sve dužine peptida.

HLA I			
Dužina epitopa	**Epitopi**		**Ukupno po dužinama**
	Slabi (WB)	**Jaki (SB)**	
8	1303456 **(84.76%)**	234357 **(15.24%)**	1537813
9	1365264 **(83.06%)**	278500 **(16.94%)**	1643764
10	1297463 **(84.60%)**	236251 **(15.40%)**	1533714
11	1260561 **(85.32%)**	216945 **(14.68%)**	1477506
	Ukupno WB	**Ukupno SB**	**Ukupno epitopa**
	5226744 **(84.40%)**	966053 **(15.60%)**	**6192797**

Tabela 4.Distribucija epitopa različite dužine u uređenim/neuređenim regionima molekula HLA klase I

HLA II			
Dužina prozora	**Epitopi**		**Ukupno po dužinama**
	Slabi (WB)	**Jaki (SB)**	
9	149947 **(93.62%)**	10229 **(6.38%)**	160176
10	709524 **(88.93%)**	88355 **(11.07%)**	797879
11	1659456 **(85.27%)**	286613 **(14.73%)**	1946069
12	2884321 **(81.67%)**	647304 **(18.33%)**	3531625
13	3919667 **(78.53%)**	1071493 **(21.47%)**	4991160
14	4672405 **(76.26%)**	1454888 **(23.74%)**	6127293
15	5164597 **(74.78%)**	1741721 **(25.22%)**	6906318
	Ukupno WB	**Ukupno SB**	**Ukupno epitopa**
	19159917**(78.33%)**	5300603 **(21.67%)**	**24460520**

Tabela 5. Distribucija epitopa različite dužine u uređenim/neuređenim regionima molekula HLA klase II

Tabelama 6 i 7 dat je prikaz broja jakih i slabih epitopa u uređenim, neuređenim i prelaznim regionima. Iz tabela se vidi da se više epitopa nalazi u uređenim regionima. Ovo tvrđenje važi i za slabe i za jake epitope, za obe klase HLA molekula.

			Broj jakih epitopa	Broj slabih epitopa	Ukupno epitopa
HLA I	VSL2	Neuređeni regioni	304582 **(14.66%)**	1772600 **(85.34%)**	2077182
		Uređeni regioni	517162 **(16.44%)**	2627745 **(83.56%)**	3144907
		Prelazni regioni	144309 **(14.87%)**	826399 **(85.13%)**	970708
			Broj jakih epitopa	**Broj slabih epitopa**	**Ukupno epitopa**
	IUpred	Neuređeni regioni	107396 **(14.53%)**	631766 **(85.47%)**	739162
		Uređeni regioni	704796 **(16.12%)**	3668647 **(83.88%)**	4373443
		Prelazni regioni	153861 **(14.24%)**	926331 **(85.76%)**	1080192

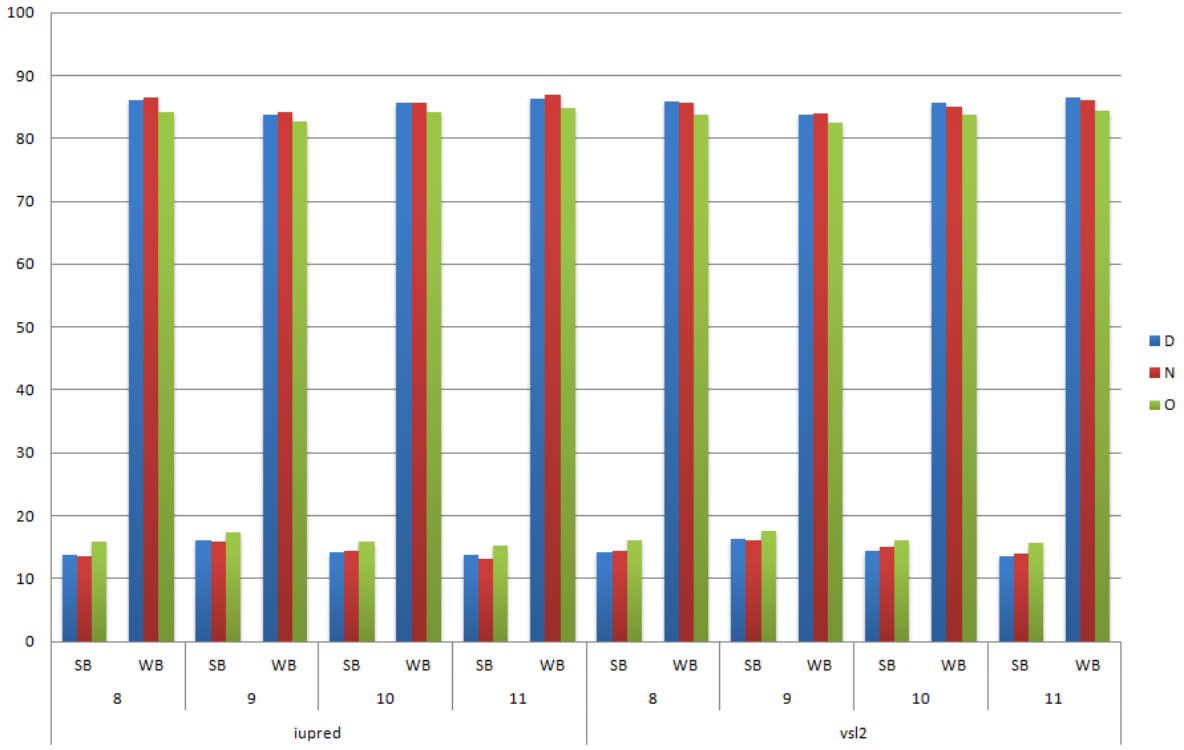
Tabela 6. Distribucija epitopa (jakih, slabih) u regionima različite strukture za HLA molekule klase I

			Broj jakih epitopa	Broj slabih epitopa	Ukupno epitopa
HLA II	VSL2	Neuređeni regioni	1259436 **(18.33%)**	5613271 **(81.67%)**	6872707
		Uređeni regioni	2856235 **(23.49%)**	9304125 **(76.51%)**	12160360
		Prelazni regioni	1184932 **(21.83%)**	4242521 **(78.17%)**	5427453
			Broj jakih epitopa	**Broj slabih epitopa**	**Ukupno epitopa**
	IUpred	Neuređeni regioni	238176 **(15.61%)**	1287552 **(84.39%)**	1525728
		Uređeni regioni	4024728 **(22.91%)**	13541370 **(77.09%)**	17566098
		Prelazni regioni	1037699 **(19.33%)**	4330995 **(80.67%)**	5368694

Tabela 7. Distribucija epitopa (jakih, slabih) u regionima različite strukture za HLA molekule klase II

SQL upitima nad tabelama HLA1 i HLA2 dobijeni su podaci o učestalosti pojavljivanja jakih i slabih epitopa u regionima različite uređenosti i različitih dužina, kao i frekventnost vezivanja za određeni supertip. U radu je prikazan samo deo rezultata, a svi rezultati su snmljeni na CD i priloženi kao dodatak radu. Rezultati su prikazani u procentima gde za određenu dužinu epitopa (prozora) i uređenost jaki i slabi epitopi u zbiru daju 100%. Radi preglednosti na nekim graficima nisu prikazani slabi epitopi. Grafički prikazi rezultata su prikazani u daljem tekstu.

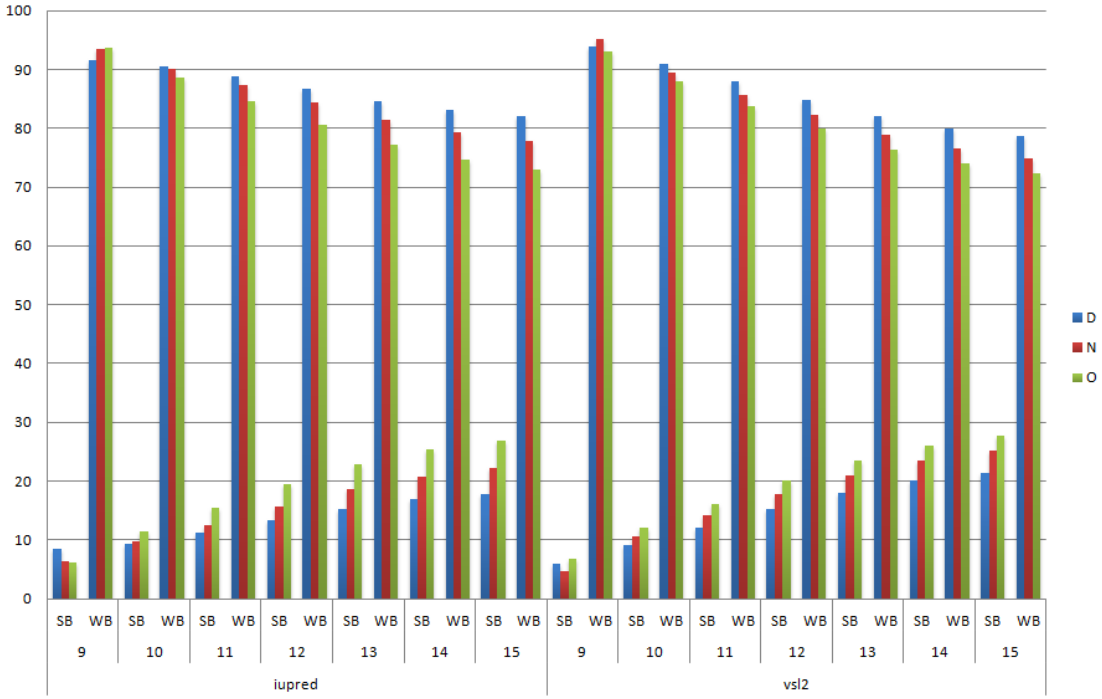
Na slici 11 dat je grafički prikaz odnosa jakih i slabih epitopa koji se vezuju za molekule HLA klase I u regionima različite uređenosti. Na grafiku se vidi da nema mnogo odstupanja u predviđanju prediktorima VSL2 i IUpred. Rezultati su dosta ujednačeni za sve dužine epitopa. Jakih epitopa dužine 9 ima nešto više u odnosu na epitope ostalih dužina. Za sve dužine važi da se najviše jakih epitopa nalazi u uređenim regionima.



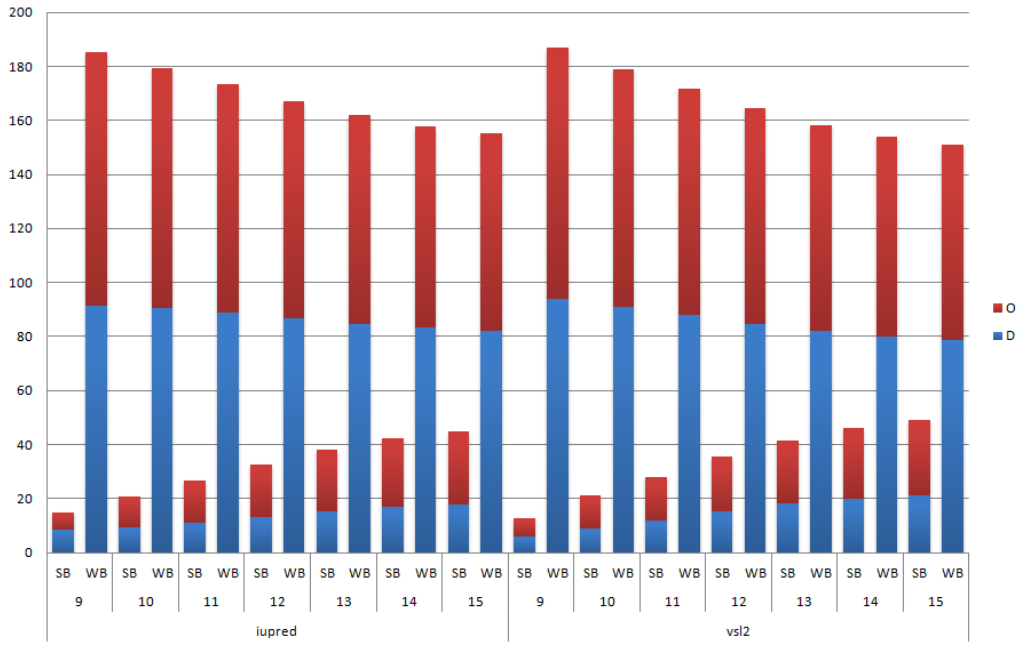
Slika 11. Grafički prikaz distribucije jakih i slabih epitopa različite dužine u uređenim, neuređenim i prelaznim regionima molekula HLA klase I

Na slici 12 dat je grafički prikaz odnosa jakih i slabih epitopa koji se vezuju za molekule HLA klase II u regionima različite uređenosti. Na grafiku se jasno vidi da broj jakih epitopa raste sa povećanjem njegove dužine. Postoje mala odstupanja u predviđanju prediktorima VSL2 i IUpred. Na osnovu predviđene uređenosti prediktorom IUpred najviše jakih epitopa dužine 9 se nalazi u neuređenim regionima, dok za ostale dužine najviše jakih epitopa ima u uređenim regionima. Prema predviđenoj uređenosti programom VSL2 za sve dužine važi da najviše jakih epitopa nalazi u uređenim regionima. Na slici 13 grafički je prikazan odnos jakih i slabih epitopa

koji se vezuju za molekule HLA klase II u uređenim i neuređenim regionima za sve dužine. Grafik jasno prikazuje pravilnost da se broj jakih epitopa povećava sa dužinom peptida.

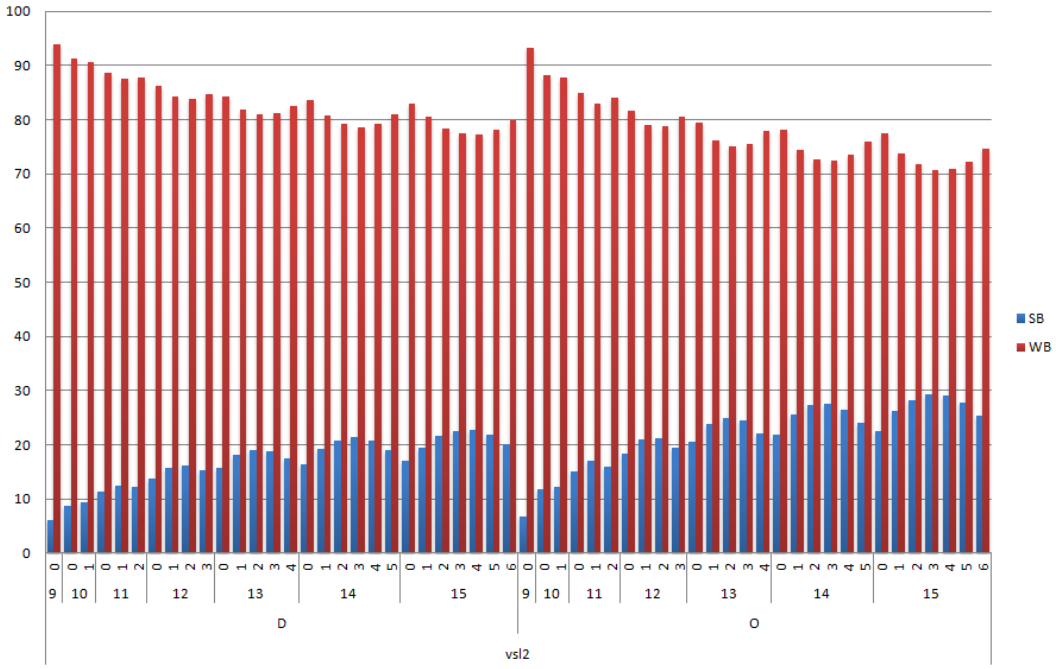


Slika 12. Grafički prikaz distribucije jakih i slabih epitopa različite dužine u uređenim, neuređenim i prelaznim regionima molekula HLA klase II

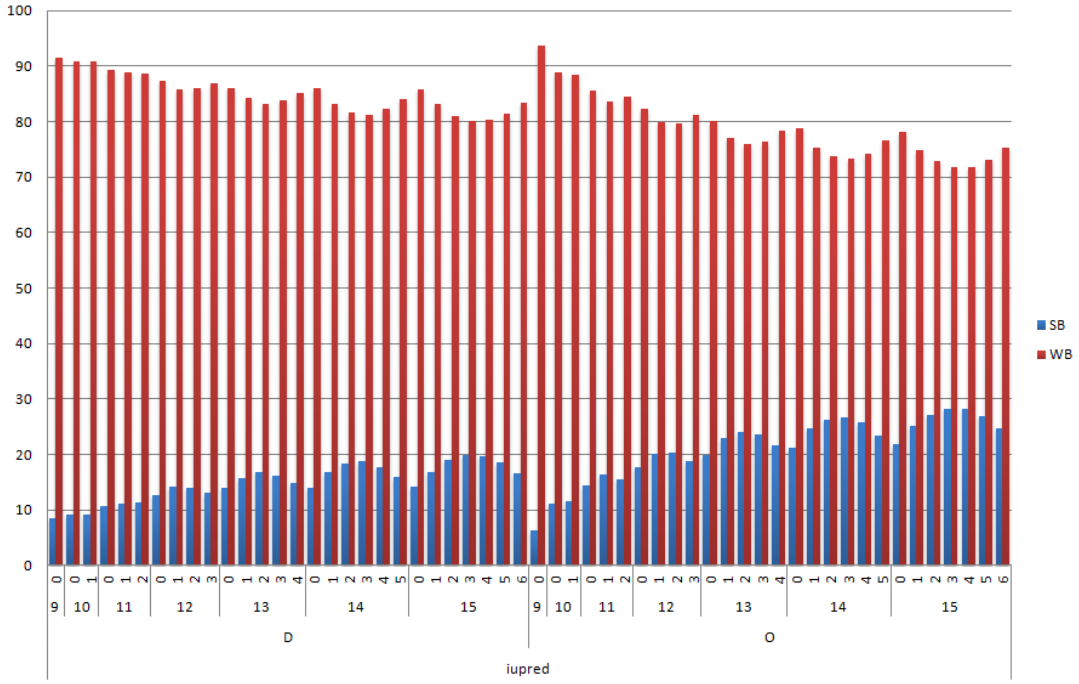


Slika 13. Distribucija jakih/slabih epitopa različite dužine u uređenim/neuređenim regionima molekula HLA klase II

Slika 14 daje prikaz distribucije jakih i slabih epitopa u zavisnosti od početne pozicije jezgra unutar peptida. Generalno gledano, bez obzira na uređenost, najviše je jakih epitopa čije je jezgro na sredini peptida bez obzira na njegovu dužinu.



Slika 14.a) Grafički prikaz broja (u %) epitopa u uređenim/neuređenim regionima u zavisnosti od pozicije jezgra unutar peptida, za sve dužine epitopa. Uređenost regiona je određena na osnovu rezultata programa VSL2.

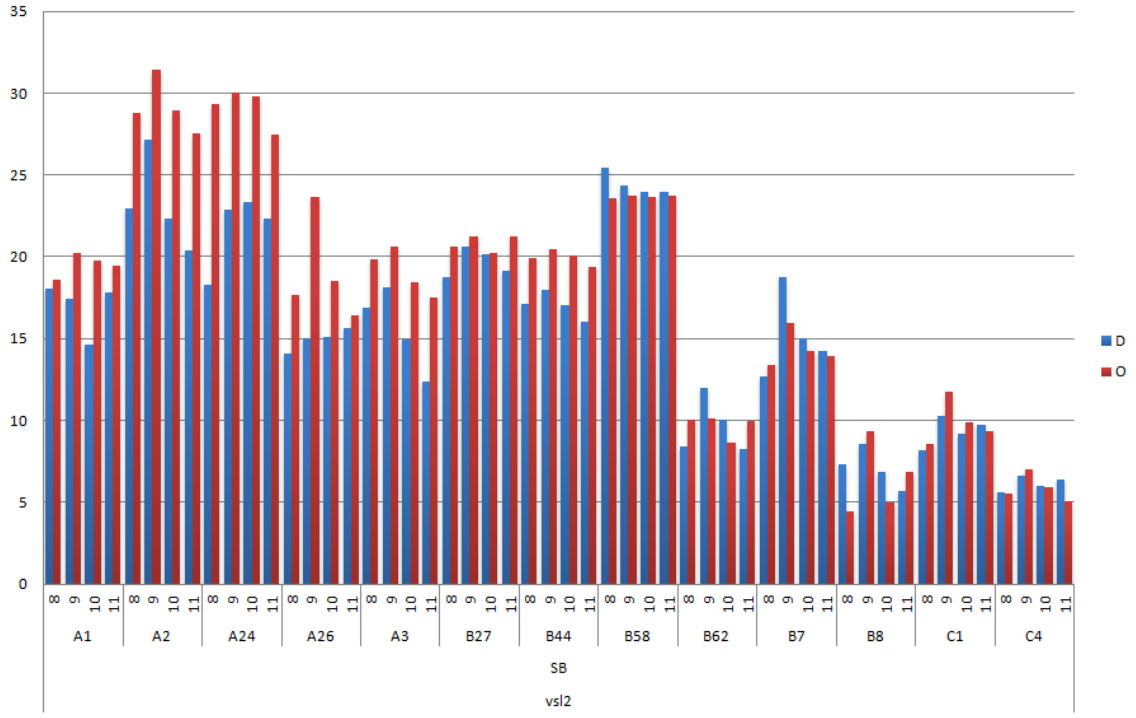


Slika 14.b) Grafički prikaz broja (u %) epitopa u uređenim/neuređenim regionima u zavisnosti od pozicije jezgra unutar peptida, za sve dužine epitopa. Uređenost regiona je određena na osnovu rezultata programa IUpred.

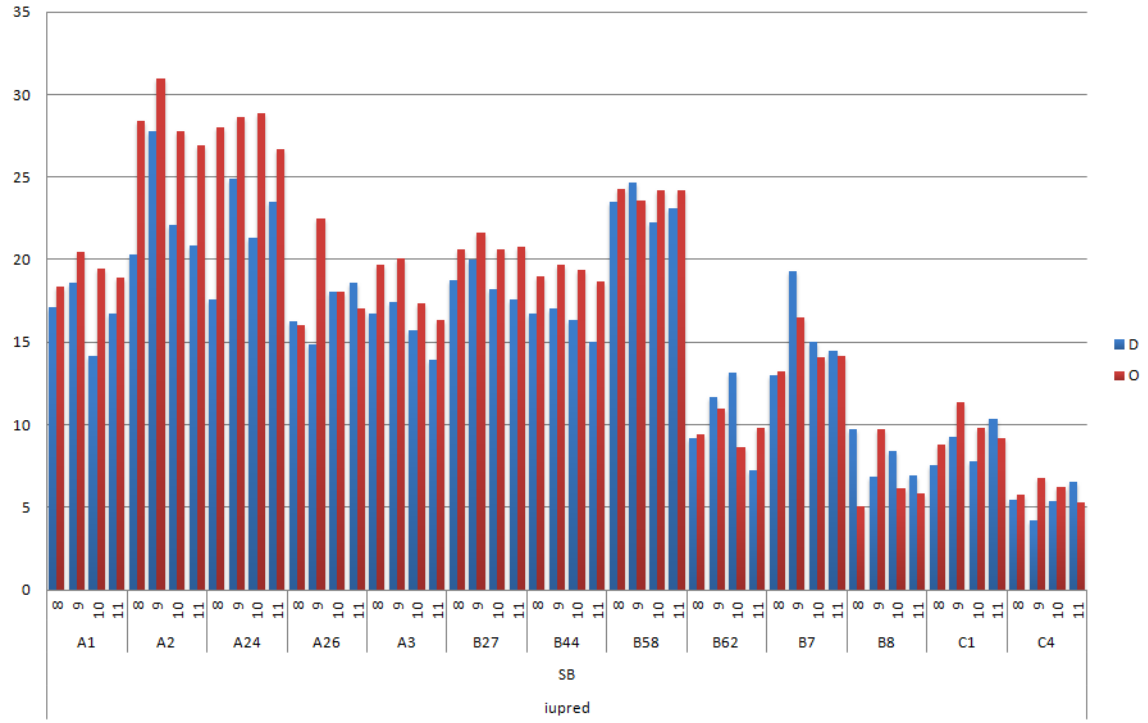
Na slici 15 za svaki supertip klasifikacije Multipred2 za molekule HLA klase I prikazano je (u %) koliko se jakih epitopa vezuje za alele koje pripadaju datom supertipu. Poznato je da pojedinim supertipovima, kao što su B7, B8 i B58, pripadaju alele koje se pretežno vezuju za neuređene epitope. Dobijeni rezultati to potvrđuju. Postoje neslaganja u predviđanju prediktorima VSL2 i IUpred, a neka od očiglednijih su:

- Supertip A26 za dužinu 8, 10 i 11
- Supertip B58 za dužinu 8, 10 i 11
- Supertip C4 za dužinu 8 i 9
- Supertip B8 za dužinu 9 i 11
- Supertip B27 za dužinu 10
- Supertip B62 za dužinu 10

Najviše ima epitopa koji se vezuju za alele koje pripadaju supertipovima A2, A24 i B58.



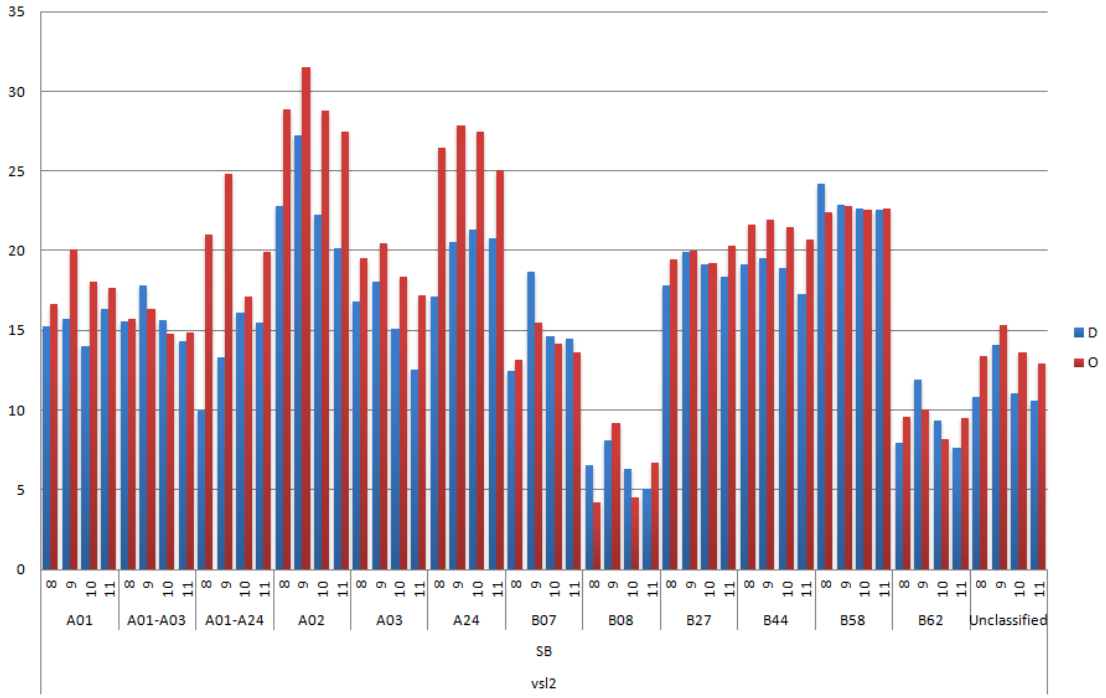
Slika 15.a) Grafički prikaz distribucije jakih epitopa koji se vezuju za HLA I alele po supertipovima klasifikacije Multipred2 na osnovu predviđanja uređenosti programom VSL2



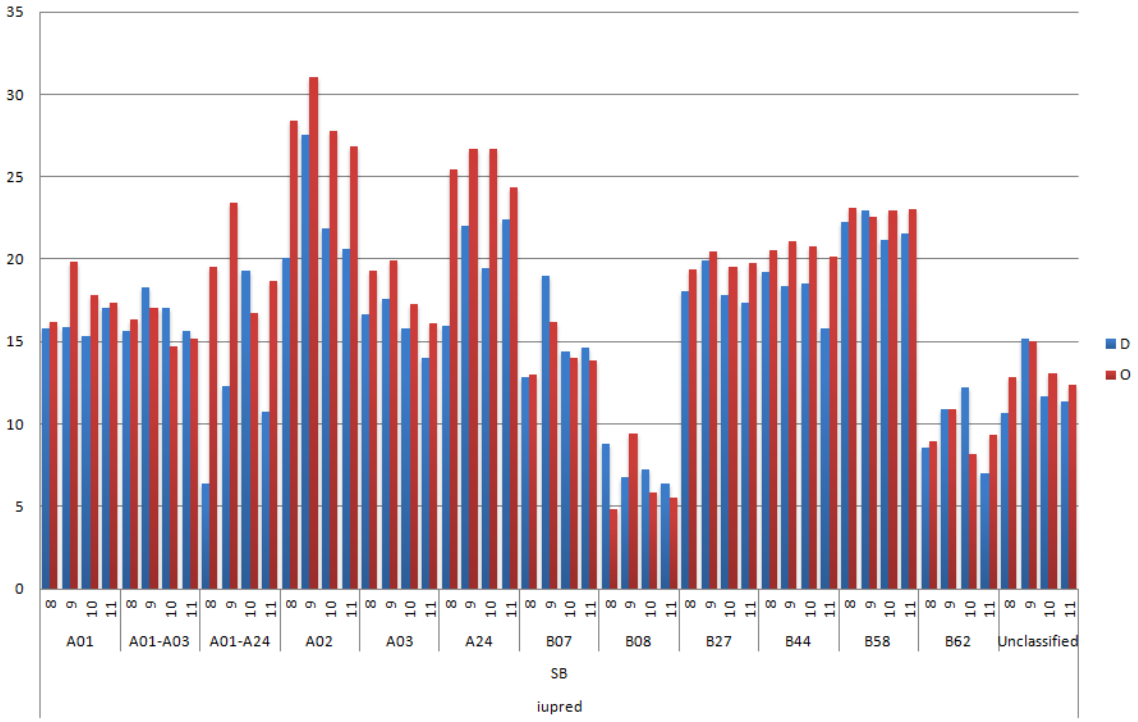
Slika 15.b) Grafički prikaz distribucije jakih epitopa koji se vezuju za HLA I alele po supertipovima klasifikacije Multipred2 na osnovu predviđanja uređenosti programom IUpred

Slično prethodnoj klasifikaciji, na slici 16 su prikazani grafici za klasifikaciju Sidney. Očekivano najviše ima epitopa koji se vezuju za alele koje pripadaju supertipovima A2, A24 i B58. Poređenjem predviđanja programima VSL2 i IUpred najveća neslaganja se javljaju za:

- Sve dužine supertipa B58
- Supertip A01-A03 za dužinu 11
- Supertip A01-A24 za dužinu 10
- Supertip B08 za dužinu 11
- Supertip B27 za dužinu 10
- Supertip B62 za dužinu 9 i 10



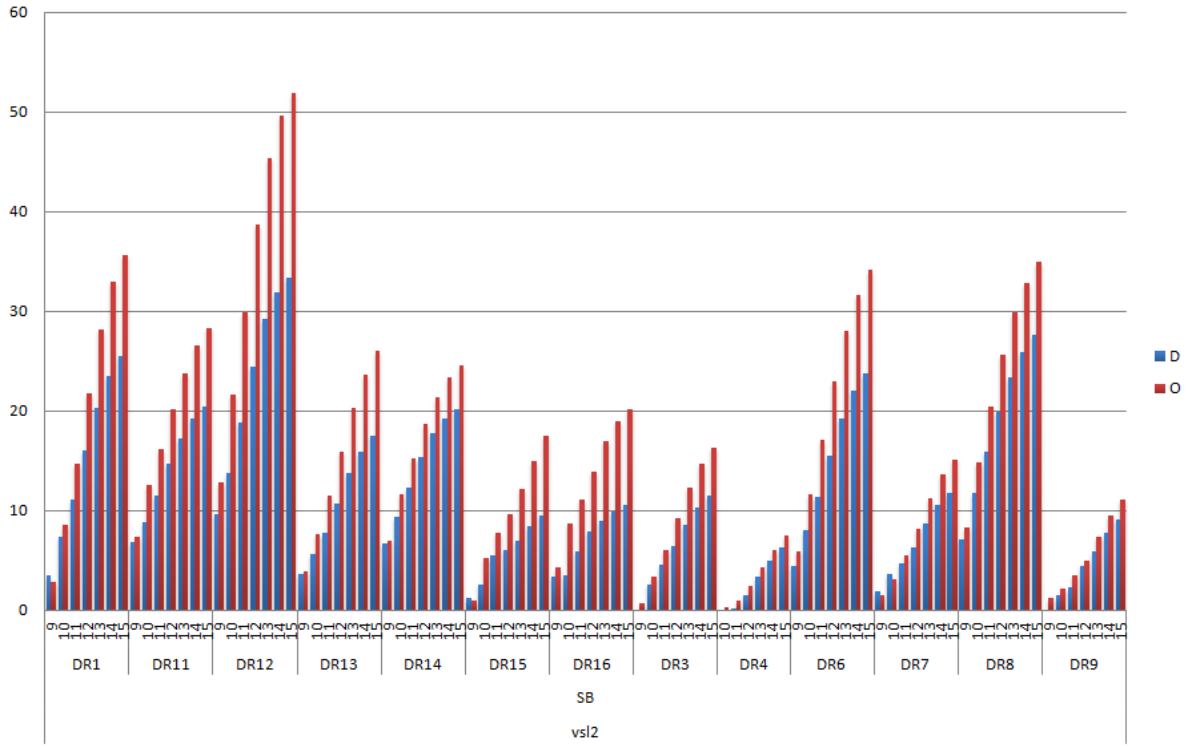
Slika 16.a) Grafički prikaz distribucije jakih epitopa koji se vezuju za HLA I alele po supertipovima klasifikacije Sidney na osnovu predviđanja uređenosti programom VSL2.



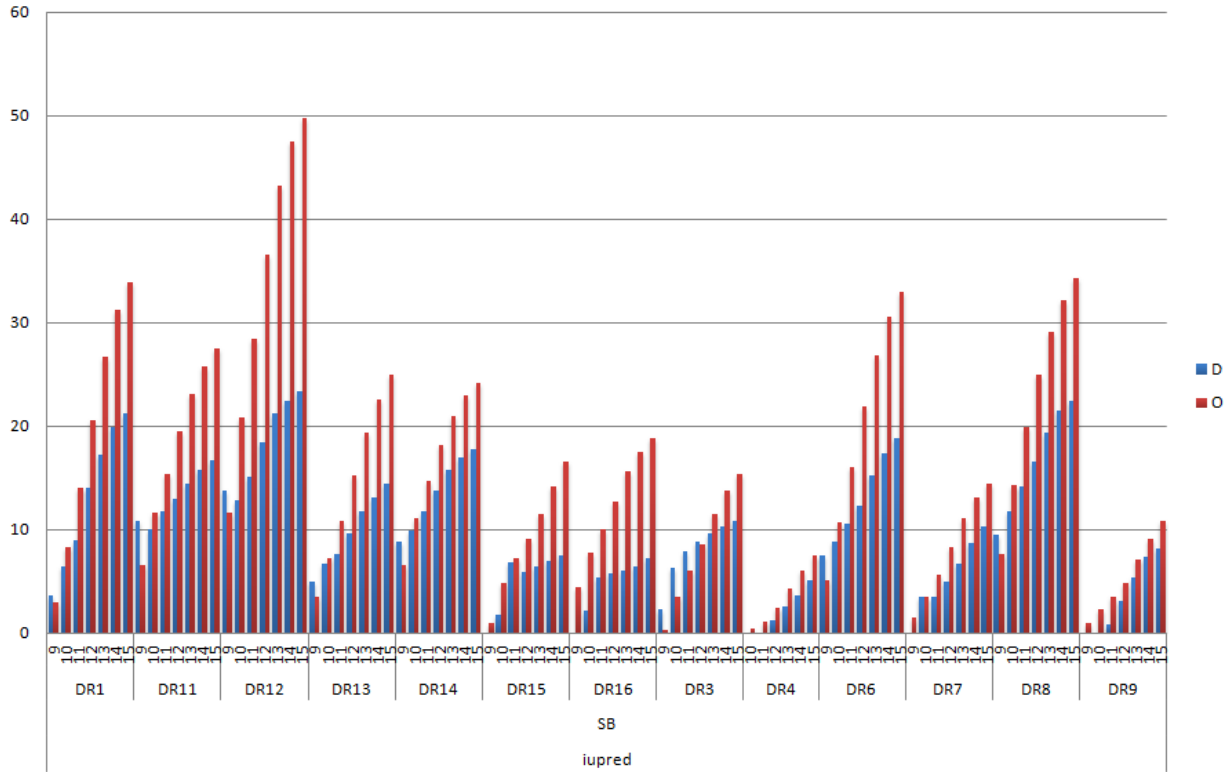
Slika 16.b) Grafički prikaz distribucije jakih epitopa koji se vezuju za HLA I alele po supertipovima klasifikacije Sidney na osnovu predviđanja uređenosti programom IUpred.

U nastavku su prikazani grafički rezultati za supertipove koji se odnose na HLA II alele i učestalost epitopa koji se za njih vezuju. Na slici 17 je dat grafik distribucije jakih epitopa po supertipovima klasifikacije Multipred2. Za oba prediktora i za sve supertipove važi da su jaki epitopi sa većim dužinama u većini u uređenim regionima. Izuzetak je supertip DR3, gde prema predviđanju IUpred programom jaki epitopi u većini se nalaze u neuređenim regionima, za dužine 9, 10, 11 i 12. Jaki epitopi koji su u većini u neuređenim regionima su dužine 9 i 10. Programi VSL2 i IUpred u ovom slučaju daju dosta slična predviđanja, osim za epitope dužine 9. Veća neslaganja se javljaju za:

- Supertip DR11 za dužinu 9
- Supertip DR12 za dužinu 9
- Supertip DR13 za dužinu 9
- Supertip DR14 za dužinu 9
- Supertip DR15 za dužinu 9
- Supertip DR16 za dužinu 9
- Supertip DR3 za dužinu 9, 10, 11 i 12
- Supertip DR6 za dužinu 9
- Supertip DR7 za dužinu 9 i 10
- Supertip DR9 za dužinu 9
- Supertip DR9 za dužinu 10 i 11
-

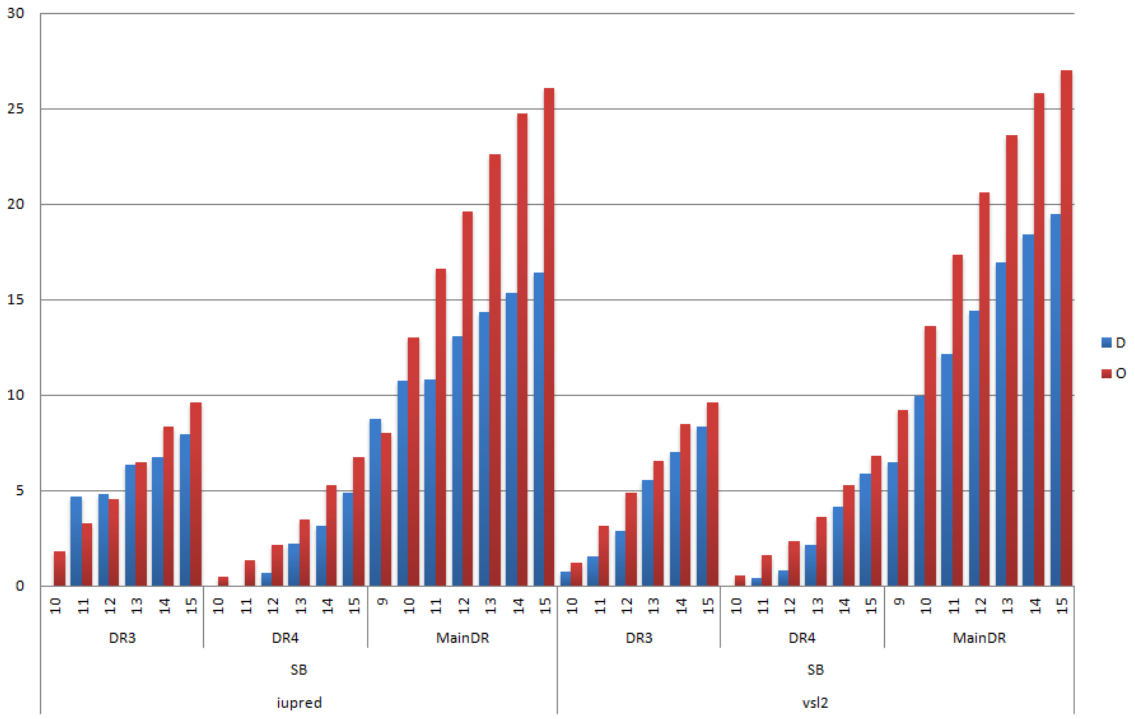


Slika 17.a) Grafički prikaz distribucije jakih epitopa koji se vezuju za HLA II alele po supertipovima klasifikacije Multipred2 na osnovu predviđanja uređenosti programom VSL2.



Slika 17.b) Grafički prikaz distribucije jakih epitopa koji se vezuju za HLA II alele po supertipovima klasifikacije Multipred2 na osnovu predviđanja uređenosti programom IUpred.

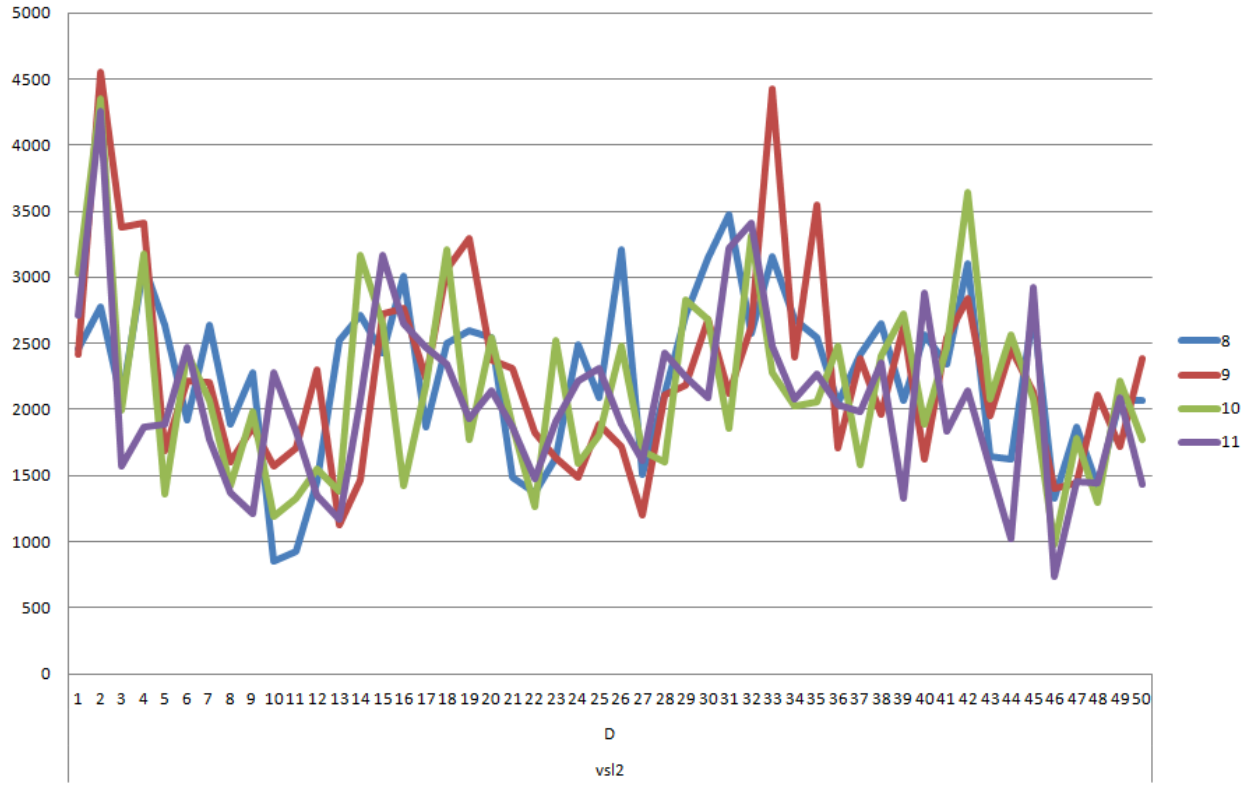
Kasifikacija Greenbaum pokriva mali broj alela (manje od 4%), pa se grafik na slici 18 odnosi na mali broj jakih epitopa. I ovde važi da su jaki epitopi većih dužina u većini u uređeni regionima. Za supertipove DR3 i DR4 nema jakih epitopa dužine 9. Predviđanje programom VSL2 jaki epitopi su većinom u uređenim regionima. Ovo tvrđenje važi za sve dužine prozora u sva tri supertipa. IUpred prediktorom se dobijaju različiti rezultati, gde se odstupanja javljaju za supertip DR3 (za dužinu 11 i 12) i supertip MainDR (za dužinu 9). U navedenim slučajevima jaki epitopi su većinom u neuređenim regionima.



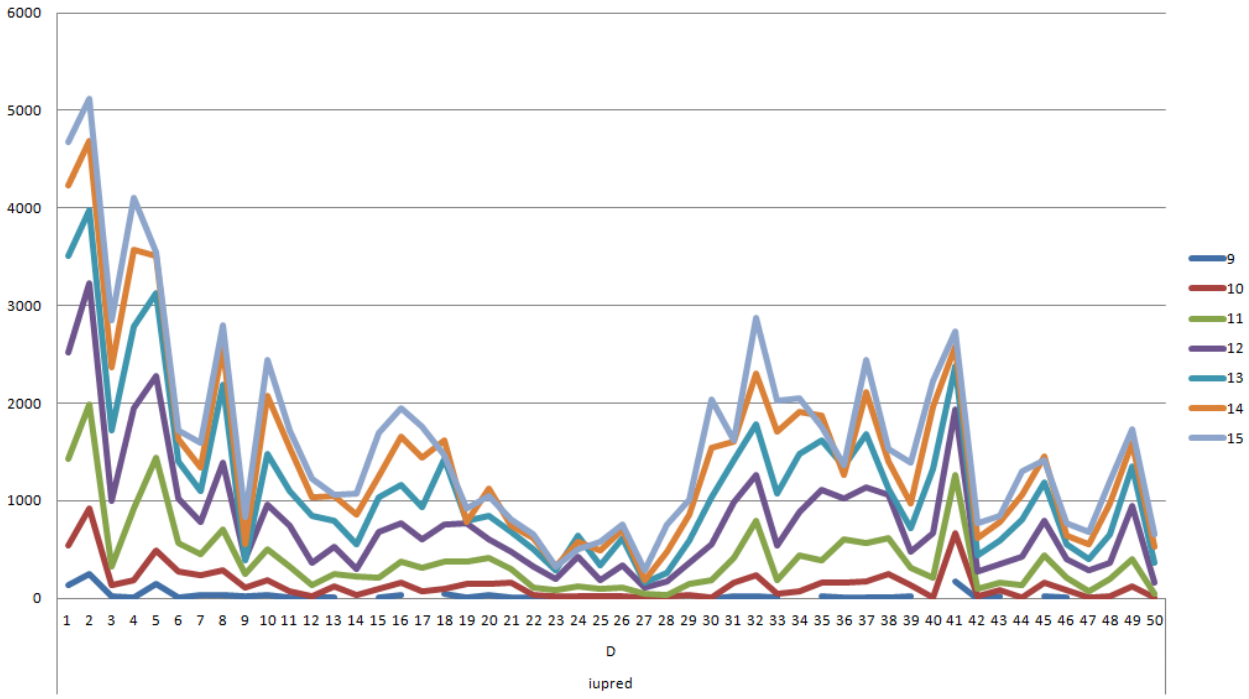
Slika 18. Grafički prikaz distribucije jakih epitopa koji se vezuju za Hla2 alele po supertipovima klasifikacije Greenbaum na osnovu predviđanja uređenosti programima VSL2 i IUpred.

8.1. Relativno udaljenje epitopa u proteinu

Istraživanje u ovom radu je sprovedeno na 143 proteina. Od potencijalnog značaja je da se odredi na kojim delovima proteina se nalaze epitopi dobijeni predviđanjem programima NetMHCpan i NetMHCIIpan. S obzirom da su proteini različitih dužina, početne pozicije epitopa su skalirane i ta nova vrednost predstavlja *relativno udaljenje* epitopa (u odnosu na početak proteina). Skaliranje je vršeno u odnosu na najduži protein, čija je dužina 2789 amino kiselina. Relativna dužina proteina (2789 amino kiselina) je podeljena u intervale dužine 10 amino kiselina i za svaku dužinu epitopa (različite uređenosti) prebrojani su epitopi čiji je relativni početak u datom intervalu. Radi preglednosti prikazan je samo deo rezultata, a svi grafici su priloženi kao dodatak radu. Rezultati za obe klase HLA prikazani su grafički na slikama 19 i 20. X osa na grafiku predstavlja redni broj intervala (dužine 10) i u ukupnom materijalu ih ima 278.



Slika 19. Raspodela HLA I epitopa svih dužina, koji pripadaju neuređenim delovima, u prvih 500 amino kiselina relativne dužine proteina.



Slika 19. Raspodela HLA II peptida svih dužina (u okviru kojih se nalaze epitopi), koji pripadaju neuređenim regionima

Epitopi koji se vezuju za klasu I zahtevaju detaljniju analizu položaja u proteinu. Nema pravilnosti u njihovoj raspodeli uzimajući u obzir njihovu dužinu, osim što su za nekoliko intervala epitopi dužine 9 znatno brojniji. Epitopi koji se vezuju za klasu II imaju sličan trend rasta i opadanja za sve dužine peptida (prozora). Za većinu intervala važi da broj epitopa raste sa povećanjem dužine peptida (u kojima se epitop nalazi). Na primeru prikazanom na slici 19 vidi se da na pojedinim intervalima nema epitopa dužine 9.

8.2. Učestalost pojavljivanja amino kiselina u epitopima

U odeljku 3.1. spomenuta je veza učestalosti pojavljivanja amino kiselina u uređenim/neuređenim proteinima. U okviru ovog istraživanja izračunata je (u %) učestalost pojavljivanja amino kiselina u proteinima nad kojima se vrši istraživanje. Učestalost pojavljivanja amino kiselina (AK) u celokupnom materijalu prikazana je tabelom 8.

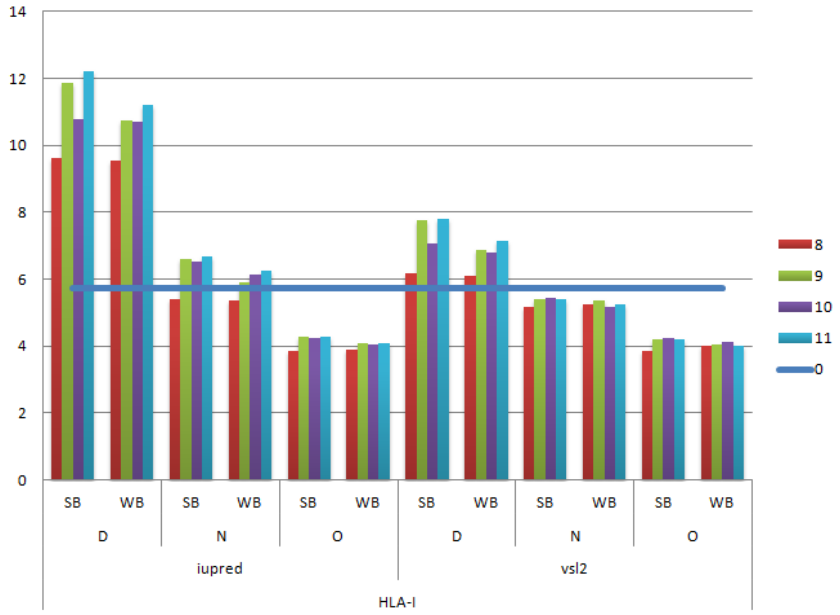
AK	Zastupljenost u %	AK	Zastupljenost u %
A	**5.63**	**L**	**9.48**
R	**4.94**	**K**	**7.34**
N	**4.48**	**M**	**2.39**
D	**5.09**	**F**	**3.32**
C	**1.95**	**P**	**5.73**
Q	**5.34**	**S**	**8.76**
E	**9.19**	**T**	**5.16**
G	**5.06**	**W**	**0.93**
H	**2.25**	**Y**	**2.51**
I	**4.80**	**V**	**5.64**

Tabela 8. Učestalost pojavljivanja amino kiselina u proteinima

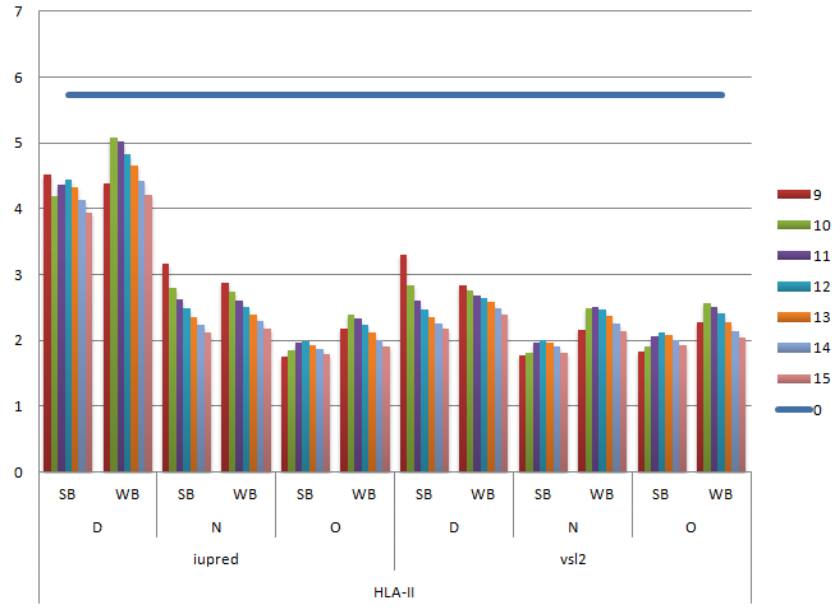
Takođe, izračunata je i učestalost pojavljivanja amino kiselina u epitopima u uređenim/neuređenim regionima i taj rezultat je upoređen sa učestalošću u celokupnom materijalu. Rezultati su predstavljeni grafički za svaku amino kiselinu pojedinačno. Svi grafici su priloženi kao dodatak radu.

Na slici 20 prikazani su rezultati dobijeni za amino kiselinu prolin u epitopima koji se vezuju za HLA klasu I i II redom. Na grafiku se jasno vidi razlika u predviđanju prediktorima VSL2 i IUpred i najizraženija je za neuređene regione. Rezultati su prikazani u procentima tako da za određenu dužinu, uređenost i vrstu epitopa (SB, WB) zbir procenata svih amino kiselina daje

100%. Na slici 20 a) se vidi da je prolin u epitopima u neuređenim regionima procentualno više zastupljen nego u celokupnom materijalu. Dok za epitope u uređenim regionima važi suprotno.



Slika 20. a. Učestalost pojavljivanja amino kiseline prolin u epitopima koji se vezuju za molekule HLA klase I. Horizontalnom plavom linijom prikazana je učestalost pojavljivanja u celokupnom materijalu.



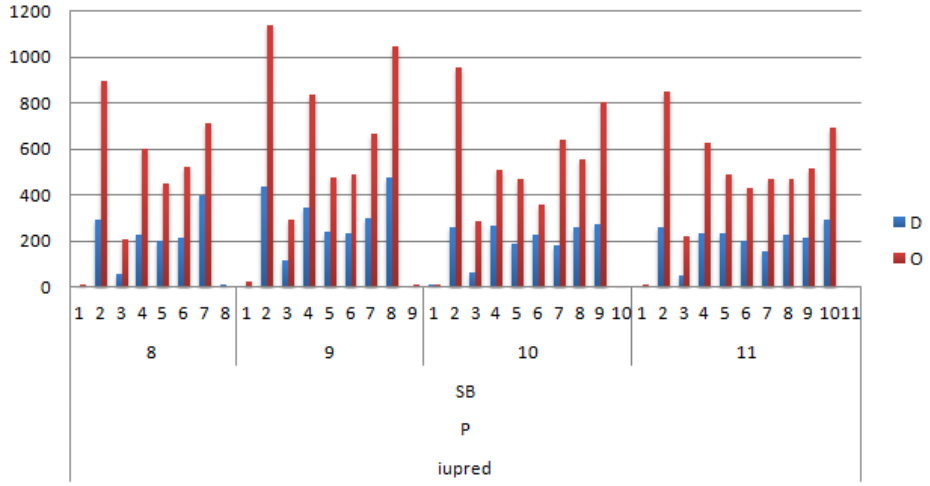
Slika 20. b) Učestalost pojavljivanja amino kiseline prolin u epitopima koji se vezuju za molekule HLA klase II. Horizontalnom plavom linijom prikazana je učestalost pojavljivanja u celokupnom materijalu.

Epitopi koji se vezuju za klasu II procentulano sadrže manje prolina u odnosu na njegovu zastupljenost u celokupnom materijalu, bez obzira na uređenost. Na slici 20 b) vidi se odstupanje u predviđanju prediktorima VSL2 i IUpred. Prema rezultatima programa IUpred, prolin u neuređenim epitopima se javlja skoro dva puta više nego u uređenim.

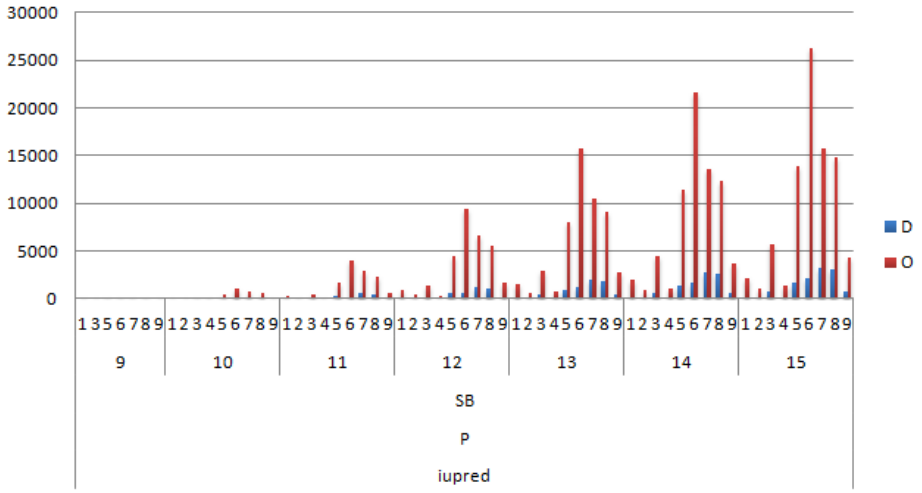
8.3. Učestalost pojavljivanja amino kiselina na određenoj poziciji unutar epitopa

Osim učestalosti pojavljivanja u celokupnom materijalu i u epitopima, za svaku amino kiselinu izračunata je frekventnost na konkretnoj poziciji u epitopu. Svi epitopi su razbijeni na pojedinačne amino kiseline, vodeći računa o njihovoj poziciji u epitopu. Zatim je računata frekventnost pojavljivanja svake amino kiseline za svaku poziciju u epitopu. Ovakvo istraživanje je sprovedeno posebno na uređene i neuređene epitope. Rezultati su grafički prikazani i priloženi su kao dodatak radu. Vrednosti u rezultatima su normalizovane brojem različitih alela u materijalu.

Na slici 21 dat je grafički prikaz zastupljenosti prolina na određenoj poziciji unutar epitopa koji se vezuju za molekule HLA klase I i II redom.



Slika 21. a) Zastupljenost prolina na različitim pozicijama unutar epitopa različitih dužina koji se vezuju za HLA I



Slika 21.b) Zastupljenost prolina na različitim pozicijama unutar epitopa koji se vezuju za HLA II. Epitopi (jezgro) se nalaze unutar peptida duzine 9-15

Epitopi koji se vezuju za HLA I skoro nikad nemaju prolin na prvoj i poslednjoj poziciji, bez obzira na dužinu i uređenost epitopa. HLA2 epitopi imaju najviše prolina na 6. poziciji.

9. Rezultati dobijeni istraživanjem podataka

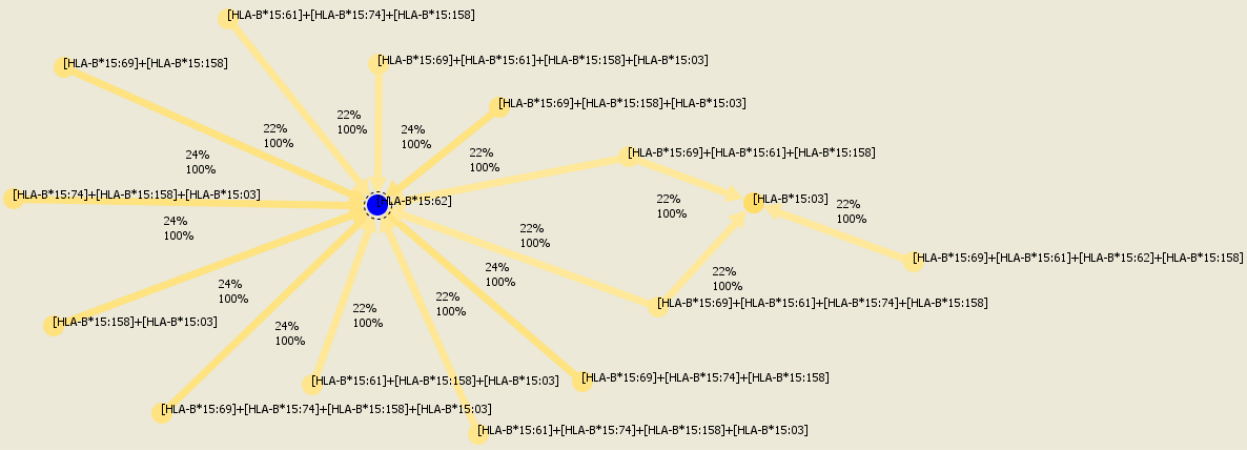
Istraživanje podataka je disciplina koja se bavi pronalaženjem informacija skrivenih u velikom broju podataka i koristi se u različitim oblastima. Cilj istraživanja podataka je pronalaženje modela koji najbolje opisuje podatke sa kojima se radi. U ovom radu od tehnika istraživanja podataka korišćene su:

- **Pravila pridruživanja** (eng. *association rules*) – najpoznatiji primer pravila pridruživanja je analiza potrošačke korpe (eng. *Market basket analysis*). Analizom potrošačke korpe se otkriva koje se stvari prodaju zajedno u isto vreme. Izdvajaju se samo značajne kombinacije, odnosno česti nizovi proizvoda i pravila o povezanosti elemenata kupovine tj. asocijativna pravila. Ova pravila su u formi A, B => C sa odgovarajućim verovatnoćama.
- **Klasterovanje** (eng. *clustering*) - pronalazi se prirodno grupisanje slučajeva na osnovu niza atributa, tako da atributi unutar jedne grupe imaju prilično slične vrednosti, a među grupama postoji značajna razlika. Dobijene grupe se nazivaju klasteri. Postoji više algoritama klasterovanja, koji su razvijeni za različite vrste podataka. U ovom radu se koristi klasterovanje zasnovano na neuronskim mrežama jer daje najbolje rezultate, odnosno dobru ocenu klasterovanja.

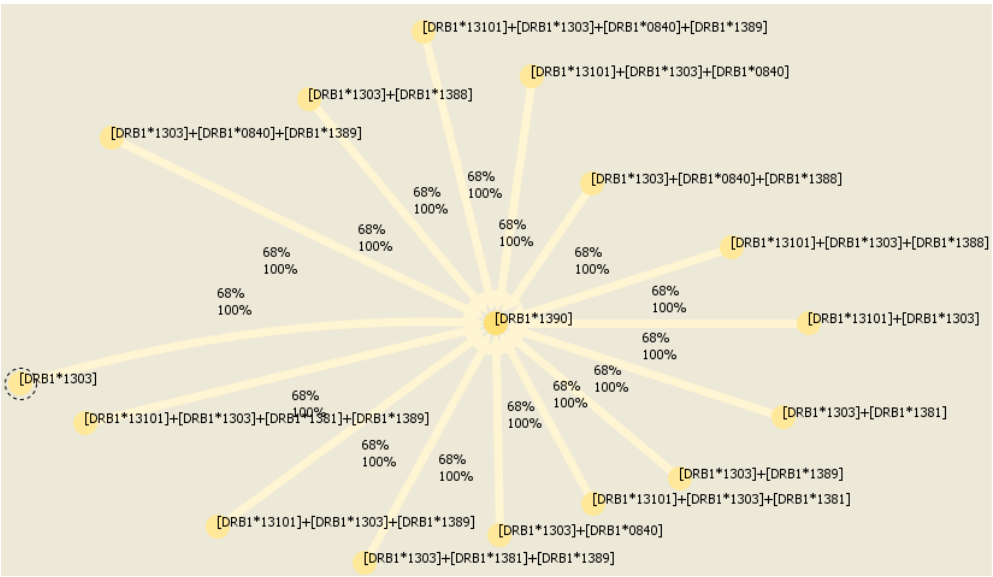
Alat za istraživanje podataka korišćen u ovom radu je InfoSphere Warehouse, proizvod kompanije IBM. Obezbeđuje pristup strukturalnim i nestrukturalnim podacima, kao i operacionim i transakcionim podacima. Svi modeli dobijeni ovim programom su priloženi kao dodatak radu. U daljem tekstu dat je kratak pregled dobijenih modela

9.1. Pravila pridruživanja

Primenom tehnika pravila pridruživanja izdvojena su pravila sa najvećom podrškom i nivoom poverenja 100%. Na ovaj način izdvojena su pravila koja se odnose na promiskuitetne alele, odnosno alele koje se vezuju za isti epitop. Na slici 22 i 23 prikazane su promisluitetne alele koje se vezuju za HLA I i HLA II epitope redom. Grafovi su generisani od izabranih pravila sa najvećom podrškom i maksimalnim poverenjem.



Slika 22. Promiskuitetne HLA I alele



Slika 23. Promiskuitetne HLA II alele

Tehnikom pravila pridruživanja izdvojena su pravila koja se odnose na jake epitope. Sva interesantna pravila imaju malu podršku. Sa druge strane, pravila koja imaju visoku podršku imaju *lift* ocenu 1 i takva pravila su na neki način očigledna.

Ova tehnika takođe je primenjena na familiju MAGE. U materijalu koji se koristi u ovom radu ima 12 proteina koji pripadaju ovoj familiji. Svi modeli i rezultati su priloženi kao dodatak radu. Na slici 24 prikazana su najznačajnija pravila koja uključuju jake epitope koji se vezuju za HLA I alele.

Rule	Support	▼ Confid...	Lift
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB] ==> [ALELA=HLA-B*15:03]	19.4477%	100.0000%	3.54
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB] ==> [ALELA=HLA-B*15:62]	19.4477%	100.0000%	3.18
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB] ==> [ALLELE_MULTIPRED2=B27]	19.4477%	100.0000%	2.24
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB] ==> [ALLELE_SIDNEY=B27]	19.4477%	100.0000%	2.20
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:03] ==> [ALELA=HLA-B*15:62]	19.4477%	100.0000%	3.18
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:03] ==> [ALLELE_MULTIPRED2=B27]	19.4477%	100.0000%	2.24
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:03] ==> [ALLELE_SIDNEY=B27]	19.4477%	100.0000%	2.20
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:62] ==> [ALELA=HLA-B*15:03]	19.4477%	100.0000%	3.54
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:62] ==> [ALLELE_MULTIPRED2=B27]	19.4477%	100.0000%	2.24
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:62] ==> [ALLELE_SIDNEY=B27]	19.4477%	100.0000%	2.20
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:69] ==> [ALELA=HLA-B*15:03]	19.3125%	100.0000%	3.54
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:69] ==> [ALELA=HLA-B*15:62]	19.3125%	100.0000%	3.18
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:69] ==> [ALLELE_MULTIPRED2=B27]	19.3125%	100.0000%	2.24
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALELA=HLA-B*15:69] ==> [ALLELE_SIDNEY=B27]	19.3125%	100.0000%	2.20
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALLELE_MULTIPRED2=B27] ==> [ALELA=HLA-B*15:03]	19.4477%	100.0000%	3.54
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALLELE_MULTIPRED2=B27] ==> [ALELA=HLA-B*15:62]	19.4477%	100.0000%	3.18
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALLELE_MULTIPRED2=B27] ==> [ALLELE_SIDNEY=B27]	19.4477%	100.0000%	2.20
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALLELE_SIDNEY=B27] ==> [ALELA=HLA-B*15:03]	19.4477%	100.0000%	3.54
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALLELE_SIDNEY=B27] ==> [ALELA=HLA-B*15:62]	19.4477%	100.0000%	3.18
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[ALLELE_SIDNEY=B27] ==> [ALLELE_MULTIPRED2=B27]	19.4477%	100.0000%	2.24
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[BIND=WB] ==> [ALELA=HLA-B*15:03]	19.4477%	100.0000%	3.54
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[BIND=WB] ==> [ALELA=HLA-B*15:62]	19.4477%	100.0000%	3.18
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[BIND=WB] ==> [ALLELE_MULTIPRED2=B27]	19.4477%	100.0000%	2.24
[ALELA=HLA-B*15:158]+[ALELA=HLA-B*15:74]+[BIND=SB]+[BIND=WB] ==> [ALLELE_SIDNEY=B27]	19.4477%	100.0000%	2.20
[ALELA=HLA-B*15:158]+[BIND=SB] ==> [ALELA=HLA-B*15:03]	19.4670%	100.0000%	3.54
[ALELA=HLA-B*15:158]+[BIND=SB] ==> [ALELA=HLA-B*15:62]	19.4670%	100.0000%	3.18
[ALELA=HLA-B*15:158]+[BIND=SB] ==> [ALLELE_MULTIPRED2=B27]	19.4670%	100.0000%	2.24
[ALELA=HLA-B*15:158]+[BIND=SB] ==> [ALLELE_SIDNEY=B27]	19.4670%	100.0000%	2.20

Slika 24. Najznačajnija pravila za MAGE familiju proteina. Pravila se odnose na jake HLA I epitope.

9.2. Klasterovanje

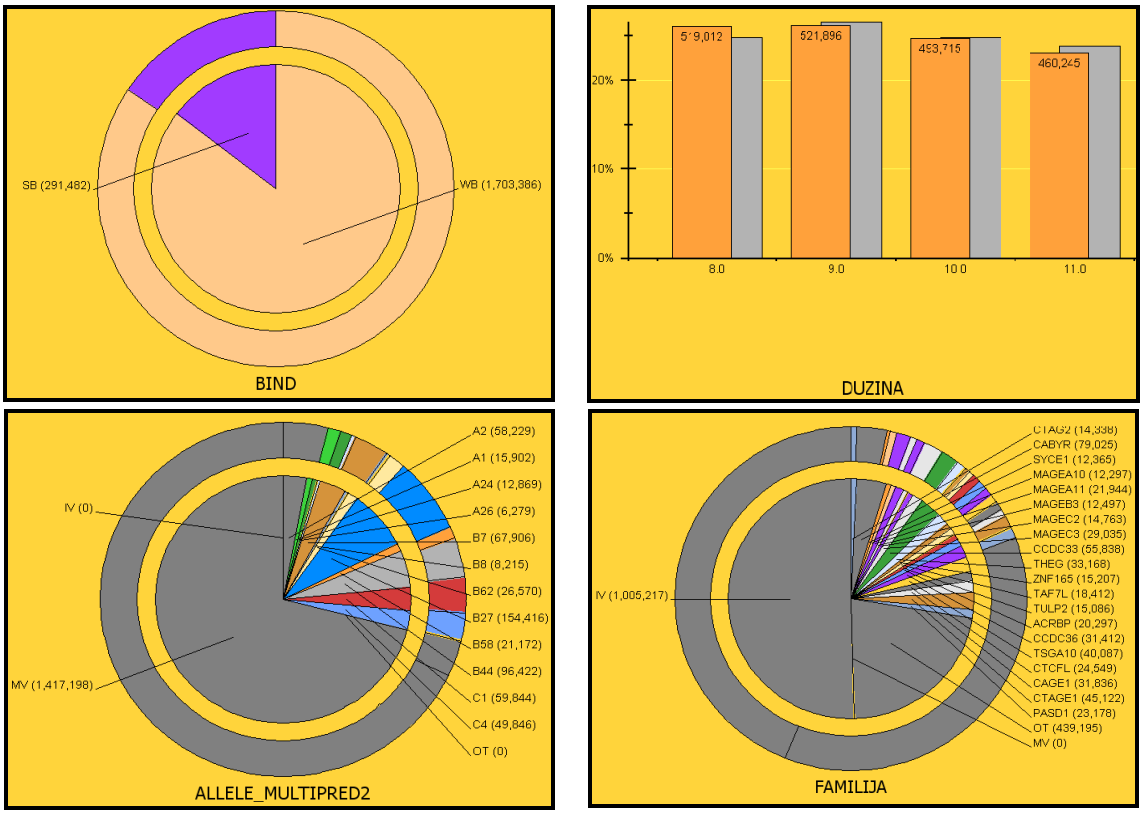
Za obe klase molekula HLA urađeno je po dva klasterovanja, zbog uređenosti epitopa koja je određivana na osnovu rezultata dva programa (VSL2 i IUpred). U dobijenim modelima vidi se da atributi koji se odnose na uređenost (*D, N, O*) i vrstu epitopa (*SB, WB*) najviše utiču na klasterovanje. Zbog polimorfnosti HLA molekula, odnosno velikog broja alela, uključivanje atributa koji se odnose na alele smanjuje kvalitet klasterovanja. Zato u ovom radu u klasterovanju ovi atributi ne učestvuju aktivno.

9.2.1. Klasterovanje HLA I epitopa

Klasterovanjem u kom je uređenost epitopa rezultat VSL2 prediktora, dobija se model čija je mera tačnosti 0.932. Podaci su podeljeni u 6 klastera. Informacije o klasterima prikazane su tabelom 9. U svim klasterima odnos slabih i jakih epitopa oponaša njihov odnos u celokupnom materijalu, odnosno slabih epitopa je mnogo više. 1.26% epitopa čine neuređeni epitopi koji se vezuju za alele koje pripadaju supertipu A3. Klasteri koji sadrže uređene epitope razlikuju se samo u dužini epitopa. Da je prilikom klasterovanja zadat manji broj klastera, verovatno bi ova dva klastera činili jedan. Najveći klaster koji sadrži jake i slabe epitope prikazan je na slici 25.

Veličina klastera	Vrsta epitopa	Uređenost	Dužina	Supertip
32.21%	Jaki i slabi epitopi	Neuređenji	Sve dužine	Svi supertipovi
28.49%	Jaki i slabi epitopi	Uređeni	8, 9 i 10	Svi supertipovi
22.24%	Jaki i slabi epitopi	Uređeni	10 i 11	Svi supertipovi
15.65%	Jaki i slabi epitopi	Prelazni	Sve dužine	Svi supertipovi
1.26%	Jaki i slabi epitopi	Neuređeni	Sve dužine	A3
0.14%	Jaki i slabi epitopi	Svi	Sve dužine	B27

Tabela 9. Klasterovanje epitopa čija je uređenost (indirektno) određena VSL2 prediktorom



Slika 25. Jaki i slabi epitopi u neuređenim regionima

Klasterovanjem u kom je uređenost epitopa rezultat IUpred prediktora, dobija se model čija je mera tačnosti 0.931. Podaci su podeljeni u 6 klastera. Informacije o klasterima prikazane su tabelom 10. Prediktorom IUpred predviđeno je više uređenih regiona, pa samim tim i epitopa u uređenim regionima. Zato se klasterovanjem dobija 3 klastera sa uređenim epitopima koji u zbiru daju 70.62% materijala.

Veličina klastera	Vrsta epitopa	Uređenost	Dužina	Supertip
30.38%	Slabi epitopi	Uređeni	8 i 9	Svi supertipovi
28.86%	Slabi epitopi	Uređeni	10 i 11	Svi supertipovi
14.96%	Slabi epitopi	Prelazni	Sve dužine	Svi supertipovi
11.94%	Jaki i slabi epitopi	Neuređeni	Sve dužine	Svi supertipovi
11.38%	Jaki epitopi	Uređeni	Sve dužine	Svi supertipovi
2.48%	Jaki epitopi	Prelazni	Sve dužine	Svi supertipovi

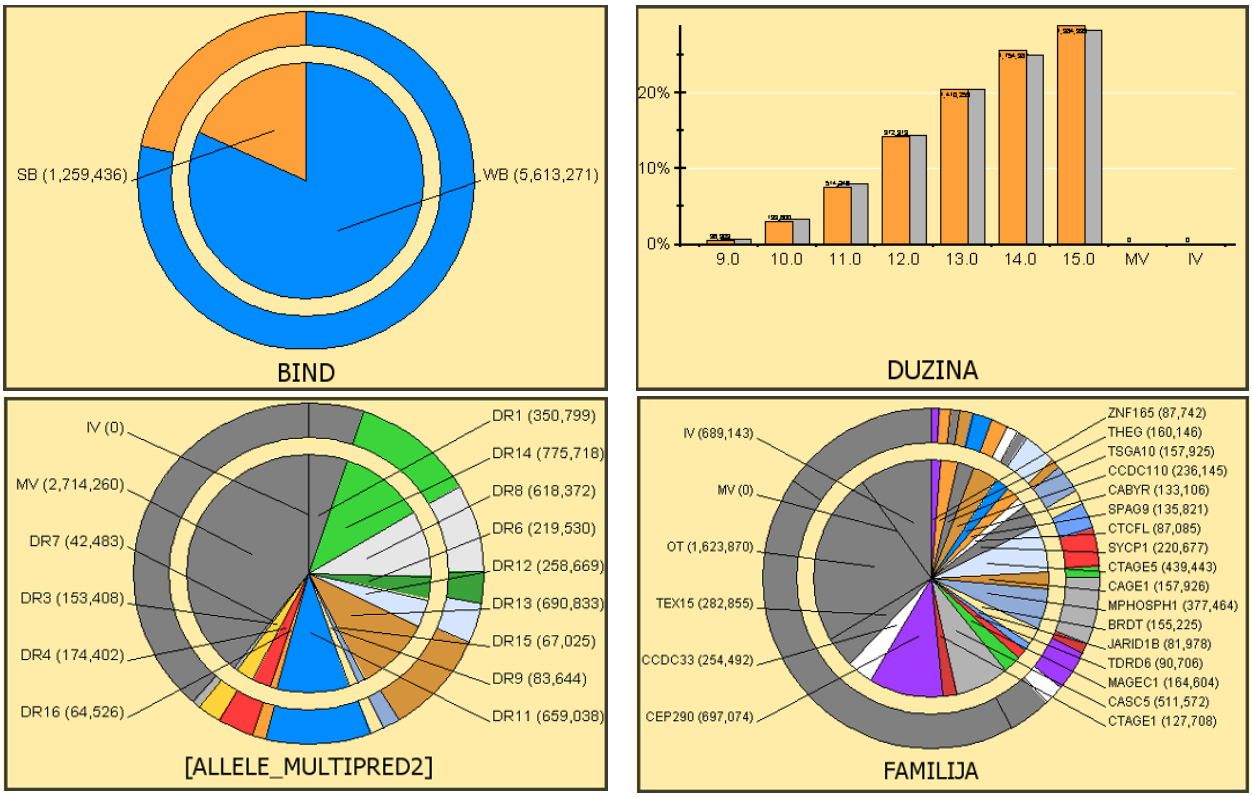
Tabela 10. Klasterovanje epitopa čija je uređenost (indirektno) određena IUpred prediktorom

9.2.1. Klasterovanje HLA II epitopa

Klasterovanjem u kom je uređenost epitopa rezultat VSL2 prediktora, dobija se model čija je mera tačnosti 0.91. Podaci su podeljeni u 5 klastera. Informacije o klasterima prikazane su tabelom 11. U svim klasterima odonos dužina peptida (u kojima se nalazi epitop) oponaša njihov odnos u celokupnom materijalu i najeviše je peptida dužine 15. Drugi po veličini klaster, koji sadrži neuređene epitope, prikazan je na slici 25.

Veličina klastera	Vrsta epitopa	Uređenost	Dužina	Supertip
38.04%	Slabi epitopi	Uređeni	Sve dužine	Svi supertipovi
28.1%	Jaki i slabi epitopi	Neuređeni	Sve dužine	Svi supertipovi
17.34%	Slabi epitopi	Prelazni	Sve dužine	Svi supertipovi
11.68%	Jaki epitopi	Uređeni	Sve dužine	Svi supertipovi
4.84%	Jaki epitopi	Prelazni	Sve dužine	Svi supertipovi

Tabela 11. Klasterovanje epitopa čija je uređenost (indirektno) određena IUpred prediktorom



Slika 26. Jaki i slabi epitopi u neuređenim regionima

Klasterovanjem u kom je uređenost epitopa rezultat IUpred prediktora, dobija se model čija je mera tačnosti 0.912. Podaci su podeljeni u 7 klastera. Informacije o klasterima prikazane su tabelom 12. Samo 6.24% epitopa pripada neuređenim regionima.

Veličina klastera	Vrsta epitopa	Uređenost	Dužina	Supertip
53.73%	Slabi epitopi	Uređeni	Sve dužine	Svi supertipovi
17.71%	Slabi epitopi	Prelazni	Sve dužine	Svi supertipovi
16.45%	Jaki epitopi	Uređeni	Sve dužine	Svi supertipovi
6.24%	Jaki i slabi epitopi	Neuređeni	Sve dužine	Svi supertipovi
3.91%	Jaki epitopi	Prelazni	Sve dužine	Svi supertipovi
1.63%	Slabi epitopi	Uređeni	Sve dužine	Svi supertipovi
0.33%	Jaki epitopi	Prelazni	Sve dužine	Svi supertipovi

Tabela 12. Klasterovanje epitopa čija je uređenost (indirektno) određena VSL2 prediktorom

10. Zaključak

U ovom radu predviđa se uređenost strukture proteina programima VSL2 i IUpred. Pronalaze se antigeni regioni u proteinu i analizira se veza i učestalost pojavljivanja antigenih regiona u različitim strukturama proteina za sve ljudske alele. Antigeni regioni (epitopi) se pronalaze korišćenjem programa NetMhcPan i NetMhciiPan koji predviđaju afinitet vezivanja peptida za molekule HLA klase I i II redom. Dobijeni rezultati su smešteni u tebele relacione baze podataka. Pored standardnih SQL upita korišćene su i napredne tehnike obrade i istraživanja podataka (klasterovanje i pravila pridruživanja).

U ovom radu analizirano je 143 kancer-testis proteina. Korišćeno je 1568 HLA I i 392 HLA II alela. Dobijeni su sledeći rezultati:

1. Većina epitopa koji se vezuju za HLA I se nalazi u uređenim regionima, bez obzira na njegovu dužinu. Ovim se potvrđuje pretpostavka da neuređeni delovi proteina predstavljaju slabe antigene. Odnos O/D se razlikuje u zavisnosti od programa kojim se predviđa uređenost:

    1.1. Prema rezultatima VSL2 programa, broj HLA I epitopa u uređenim regionima je 1.52 puta veći nego u neuređenim. Jaki epitopi u uređenim regionima su 1.7 puta brojniji u odnosu na jake epitope u neuređenim regionima.

    1.2. Prema rezultatima IUpred programa, broj HLA I epitopa u uređenim regionima je 5.2 puta veći nego u neuređenim. Jaki epitopi u uređenim regionima su 6.56 puta brojniji u odnosu na jake epitope u neuređenim regionima.

2. Slično prethodnim rezultatima, većina epitopa koji se vezuju za HLA II se nalazi u uređenim regionima. Odnos O/D u zavisnosti od programa kojim se predviđa uređenost je sledeći:

    2.1. Prema rezultatima VSL2 programa, broj HLA II epitopa u uređenim regionima je 1.78 puta veći nego u neuređenim. Jaki epitopi u uređenim regionima su 2.27 puta brojniji u odnosu na jake epitope u neuređenim regionima

    2.2. Prema rezultatima IUpred programa, broj HLA II epitopa u uređenim regionima je 11.52 puta veći nego u neuređenim. Jaki epitopi u uređenim regionima su 16.9 puta brojniji u odnosu na jake epitope u neuređenim regionima.

3. HLA I epitopi različitih dužina su ujednačeno zastupljeni, dok broj HLA II epitopa raste sa povećanjem dužine prozora (peptida) u kojem se epitop nalazi.

Za epitope koji se vezuju za HLA I, odnosno HLA II ispitano je i:

- Odnos jakih i slabih epitopa različitih dužina u regionima različite uređenosti
- Odnos jakih i slabih epitopa u uređenim/neuređenim regionima u zavisnosti od pozicije jezgra unutar peptida, za sve dužine epitopa
- Odnos jakih i slabih epitopa različitih dužina, koji su grupisani u supertipove alela (za koje se vezuju), u regionima različite uređenosti
- Analiza položaja uređenih/neuređenih epitopa u proteinima (pri krajevima ili sredini)
- Analiza učestalosti pojavljivanja pojedinačnih amino kiselina u epitopima (u odnosu na celokupan materijal)
- Analiza učestalosti pojavljivanja pojedinačnih amino kiselina na svakoj poziciji unutar epitopa

10.1. Dalji rad

Postojeća grupa kancer-testis proteina se može podeliti u manje grupe prema familiji kojoj pripadaju ili CT identifikaciji koja je pridružena svakom proteinu u tabeli PROTEIN_DETAILS. Detaljnije istraživanje bi podrazumevalo proveru utvrđenih ponašanja po familijama (grupama).

11. Korišćena literatura

[1] Golubović D: *"Primena tehnika istraživanja podataka u cilju uspostavljanja korelacije između antigenih regiona i neurđenih delova proteina"* (2010), Magistarski rad, Beograd: Matematički fakultet

[2] Uversky VN: *"The Mysterious Unfoldome: Structureless, Underappreciated, Yet Vital Part of Any Given Proteome",*Journal of Biomedicine and Biotechnology Volume 2010, 2010a, Article ID 568068.

[3] Uversky VN: *"Intrinsically disordered proteins from A to Z",The international journal of biochemistry & cell biology 2011, 43:1090-103.*

[4] Peng K, Radivojac P, Vučetić S, Dunker AK, Obradović Z: „*Length-dependent prediction of protein intrinsic disorder". BMC Bioinformatics* 2006, **7**:208.

[5] Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK: „*Intrinsic disorder in cell-signaling and cancer-associated proteins". J Mol Biol* 2002, 323(3):573-584.

[6] Uversky VN: *"Intrinsically disordered proteins may escape unwanted interactions via functional misfolding",*Biochim Biophys Acta. 2011, 1814:693–712

[7] Uversky VN, Oldfield CJ, Dunker AK: *"Intrinsically disordered proteins in human diseases: introducing the D2 concept"*. Annu Rev Biophys 2008, 37:215-246.

[8] Rajagopalan K, Mooney SM, Parekh N, Getzenberg RH, Kulkarni P: *"A Majority of the Cancer/Testis Antigens are Intrinsically Disordered Proteins",* Journal of Cellular Biochemistry, vol. 112 issue 11, November 2011. pages 3256-3267

[9] AM Powell, MM Black: *"Epitope spreading: protection from pathogens, but propagation of autoimmunity?"*, Clinical and experimental dermatology, 2001, 26, 427-433

[10] Purcell, AW, & Gorman, JJ: *"Immunoproteomics, Mollecular and cellular proteomics"*, 2004, 3, 193-208

[11] Carl, PL, Temple, BRS, Cohen, PL, *Most nuclear systemic autoantigens are extremely disordered proteins*: implications for the etiology of systemic autoimmunity, Arthritis Research & Therapy, 2005, 7, R1360-R1373

[12] Mitić N, Pavlović N, Jandrlić D: *"T-cell epitope prediction – correlation to disorder/order protein structure prediction"* (nepublikovani rezultati)

[13] NetMHCpan: http://www.cbs.dtu.dk/services/NetMHCpan (stanje 19.9.2012.)

[14] NetMHCiiPan: http://www.cbs.dtu.dk/services/NetMHCIIpan (stanje 19.9.2012.)

[15] Cancer Immunity: http://cancerimmunity.org/resources/ct-gene-database (stanje 19.9.2012.)

[16] Sidney J, Peters B, Frahm N, Brander C, Sette A.: "*HLA class I supertypes: a revised and updated classification"*, BMC Immunol 2008; 9:1.

[17] MULTIPRED2: http://cvc.dfci.harvard.edu/multipred2/index.php (stanje 29.9.2012.)

[18] Greenbaum J, Sidney J, Chung J, Brander C, Peters B, Sette A:*" Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes"*, Immunogenetics, 2011, 63:325–335

11.1. Korisna literatura

[1] Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB*: "Protein disorder prediction: Implications for structural proteomics",* Structure Volume 11, Issue 11, 2003

[2] Uversky VN, Dunker AK*: "Understanding protein non-folding"*, Biochim. Biophys. Acta - Proteins and Proteomics. 1804 (6) 1231–1264, 2010b

[3] Dosztanyi Z, Csizmok V, Tompa P, Simon I. *"IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.",* Bioinformatics. 2005;21(16):3433-4, PMID: 15955779

[4] Tong JC, Tan TW, Ranganathan S: *"Methods and protocols for prediction of immunogenic epitopes"*, Bioinformatics, 96 – 108, 2006.

[5] Brusić V, Flower DR: *"Bioinformatics tools for identifying T-cell epitopes"*; Drug Discovery Today: BIOSILICO Volume 2, Issue 1, 1 January 2004, Pages 18-23.

[6] Dis Prot, verzija 6.0, http://www.disprot.org (stanje 19.9.2012.)

[7] Zhang GL, Deluca DS, Keskin DB, Chitkushev L, Zlateva T, Lund O, Reinherz EL, Brusic V.:*"MULTIPRED2: A computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles",* J Immunol Methods, 2011 Nov 30;374(1-2):53-61