MATEMATIČKI FAKULTET

UNIVERZITET U BEOGRADU

Davorka Jandrlić

Primena tehnika istraživanja podataka na uspostavljanje korelacije između neuređenih i antigenih regiona proteina

Magistarski rad

Mentor: dr Nenad Mitić

Beograd, 2010.

Sadržaj

1 Uvod

Važan zadatak u bioinformatici je predviđanje funkcionalnih osobina proteina na osnovu redosleda amino kiselina u proteinskoj sekvenci. Prostorna (3D) struktura proteina uslovljava funkciju proteina. Pridruživanje proteinske sekvence nekoj strukturnoj familiji ili identifikovanje značajnih motiva u sekvenci predstavlja osnovi u predviđanju funkcije proteina. Uspostavljanje korelacije između uređenih / neuređenih strukturnih delova proteina, zbog funkcije koju imaju, sa antigenim regionima proteina (epitopima) je od velikog potencijalnog značaja, zbog osnovnih i kliničkih proučavanja imunološkog odgovora, pravljenja vakcina i proučavanja i tretmana bolesti.

Postoje četiri pristupa za predviđanje epitopa:

- metode zasnovane na analizi proteinskih sekvenci, kojima je moguće predvideti samo kontinualne epitope,
- metode zasnovane na analizi 3D strukture proteina, kojima je moguće predvideti samo diskontinualne epitope,
- hibridne metode koje kombinuju sekvencionalnu i strukturalnu analizu proteina,
- konsenzus metode koje kombinuju predviđanje epitopa dobijeno različitim metodama.

Predviđanje T-ćelijskih epitopa, koje je obrađeno u ovom radu, zasnovano je na analizi proteinskih sekvenci i vrši se indirektno, tako što se pronalaze peptidi koji se vezuju za molekule MHC klasa: I i II. Pretpostavka je da bi kombinacija metoda koje, na osnovu sekvence, predviđaju strukturu proteina mogla da pruži odgovor na neke od značajnih imunoloških pitanja kao što su raspodela i učestalost epitopa u različitim strukturalnim (i funkcionalnim) delovima antigena, jačina vezivanja epitopa za molekule MHC klase I i II i fenomen širenja (eng. „spreading“) imunološkog odgovora koji je od posebanog značaja za autoimuna oboljenja i izazivanje imunološkog odgovora na tumor – pridružene antigene.

1.1 Struktura proteina

Proteini ili belančevine su makromolekuli koji čine osnovu živih organizama. Proteini ili polipeptidi su linearni lančani polimeri koji se sastoje od osnovnih monomernih jedinica aminokiselina. Razlika između peptida i polipeptida je što su peptidi kratki, do 100 amino kiselina, a polipeptidi dugački (preko 100 amino kiselina). Svaki proteinski polimer je sekvenca koja sadrži kombinacije 20 različitih L-α-aminokiselina povezanih peptidnom vezom (CO-NH). Aminokiseline su molekuli koji se sastoje od amino i karboksilne grupe koje su vezane za tzv. α-C atome i bočni radikal, koji može da varira od H atoma kod aminokiseline glicina, do složene heterociklične molekulske grupe aminokiseline triptofana.

Redosled amino kiselina u proteinu određuje prostornu strukturu proteina, a od prostorne strukture proteina direktno zavisi funkcija proteina

Postoje četiri nivoa strukture proteina:

- **Primarna struktura** predstavlja redosled amino kiselina u polipeptidnom lancu (sekvenca uzastopnih amino kiselina).
- **Sekundarna struktura** predstavlja lokalnu prostornu organizaciju (konformaciju) atoma polipeptidne kičme koja je definisana vodoničnim vezama između amido i karboksilne grupe u sekvenci aminokiselina u polipeptidu (pri čemu se priroda veza bočnih ostataka aminoksielina i njihove konformacije ne uzimaju u obzir). Torzioni uglovi Ramahandranovog dijagrama (eng. „Ramachandran φ and ψ dihedral torsion angles") između α-C atoma i C atoma u COOH grupi i N atoma u NH2 grupi određuju sekundarnu strukturu proteina.
- **Tercijarna struktura** je trodimenzionalna struktura čitavog polipeptidnog lanca.
- **Kvaternarna struktura** je prostorni raspored više polipeptida (podjedinica) koje čine protein.

Primarnu strukturu proteina čone njegova jedinstvena amino kiselinska sekvenca (niska) i raspored disulfidnih mostova. Broj i raspored amino kiselina varira od proteina do proteina. Direktna informacija o rasporedu je sadržana u genima. I najmanja promena u primarnoj strukturi može značajno da utiče na ukupnu strukturu i funkcionisanje proteina. Sekundarna struktura je konformacija polipeptidnog lanca zasnovana na vodoničnim vezama. Osnovni oblici koji se podrazumevaju pod sekundarnom strukturom su α-heliks, β-nabrana struktura (β-ravan) i β-zavoj. Sekundarna struktura proteina nije nepromenjiva, te su moguće konformacione promene vezane za funkcionisanje proteina, promene u okolini. Tercijarnu strukturu određuje raspored podjedinica i zasnovana je na nizu različitih interakcija. Reč je o interakcijama između delova polipeptidnog lanca udaljenih u primarnoj strukturi. Kvaternarna struktura je prostorni raspored polipeptida u proteinima koji imaju više podjedinica.

1.2 Uređena i neuređena struktura proteina

Mnogi proteinski regioni ili neki celi proteini nemaju definisanu 3D strukturu, kao što pokazuju eksperimentalni podaci dobijeni u *in vitro* uslovima. Oni pokazuju različite konformacione izomere u kojima se pozicije atoma i torzionih uglova polipeptidne kičme menjaju u toku vremena. Postojeći nazivi ovih proteina obuhvataju više izraza kao „urođena neuređena / neuvijena / denaturisana (struktura)", ali je ipak najšćešće u upotrebi „suštinski neuređeni /neuvijeni /nestrukturisani proteini" od (eng. "intrinsically disordered /unfolded /unstructured proteins". U ovom radu će se koristiti izraz uređeni, odnosno, neuređeni proteini. Oni mogu biti potpuno uređeni ili neuređeni ili se sastoje od uređenih i neuređenih regiona.

Neuređeni regioni se eksperimentalno identifikuju na osnovu 3D strukture proteina. Tradicionalno, identifikovanje 3D strukture se izvodi eksperimentalnim metodama, od kojih su najznačajnije:

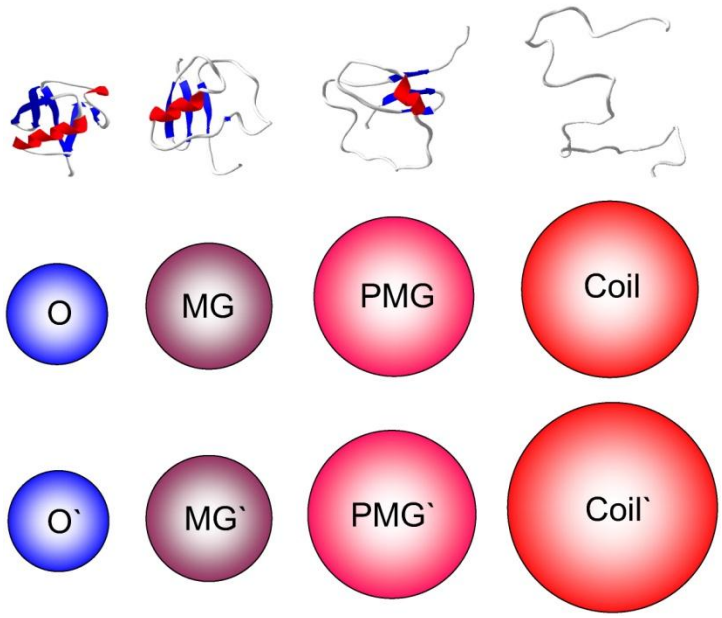
- Difrakciona kristalografija X - zracima
- Nuklearno magnetna rezonantna spektroskopija (NMR),

Ove metode su vremenski veoma zahevne i imaju još niz drugih ograničenja. Do sada je razvijeno oko 20 biofizičkih i biohemijshih metoda za određivanje neuređenih delova proteina. Razvijeno je više od 60 programa za predviđanje neuređene strukture. Programi za predviđanje neuređene strukture se dele na dve grupe na osnovu principa na kojima funkcionišu:

- programi zasnovani na fizikohemijskim osobinama aminokiselina u proteinu (PONDR, FoldUnFold, PreLINK, IUPred, GlobProt, FoldIndex), i
- programi zasnovani na metodama poravnanja (eng. „alignement") homologih proteinskih sekvenci (RONN, DISOPRED).

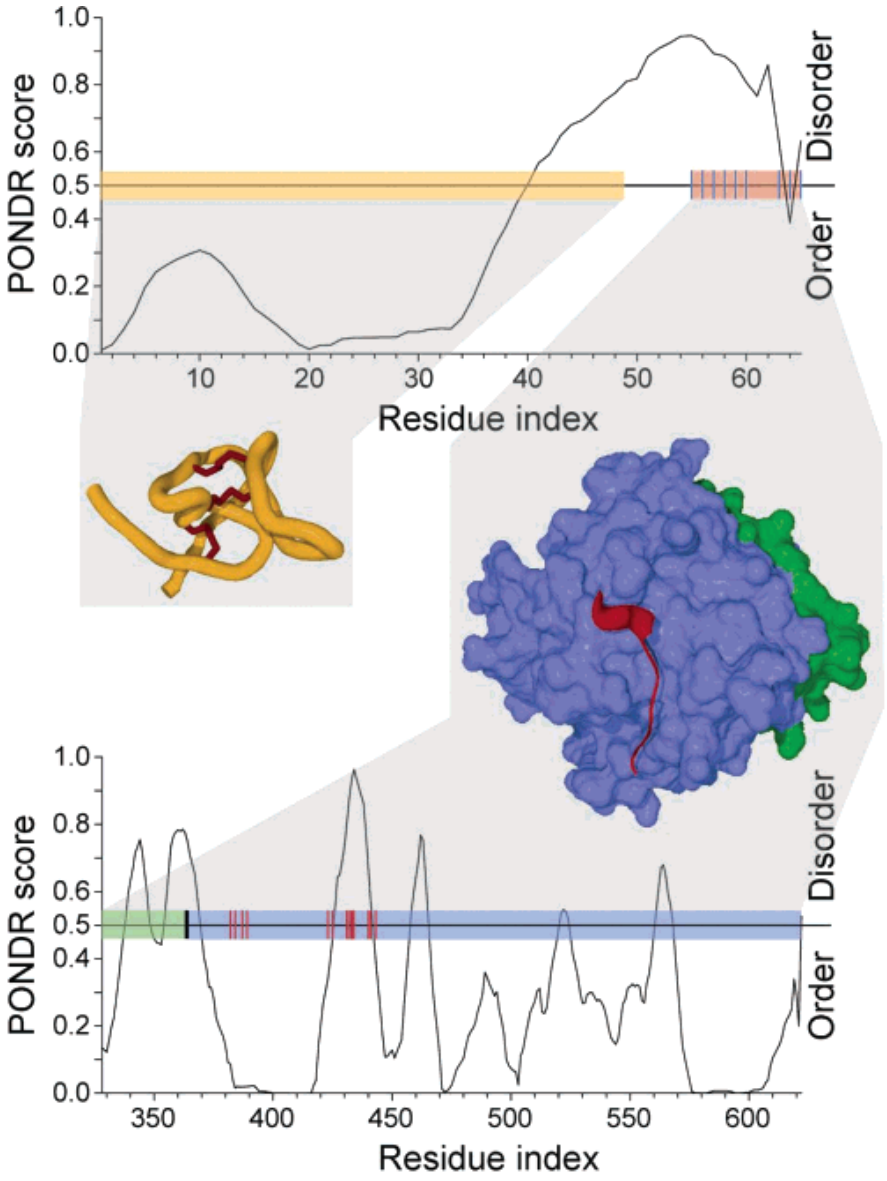
Na osnovu eksperimentalnih podataka i predviđanja neki autori su podelili neuređene regione na 3-5 grupa (a) kratke: 1–3, 4–15, 16–30, (b) duge: 30–100 i 100-200 i (c) veoma duge: >200 amino kiselina [17].

Razlike u obliku strukture neuređenih proteina su takođe velike. Nanejuređenija struktura je nasumično klupko (eng. "random coil"), koje odgovara najviše razvijenom stanju globularnih proteina, pre-topljiva globula (eng. „pre-molten globule") je izdužena, delimično strukturisana forma, topljiva globula (eng. „molten globule") je kompaktna neuređena struktura koja može sadržati značajne delove uređene strukture). Poslednje stanje je uređena (eng. „order" struktura). Navedene strukture su prikazane na slici 1.



Slika 1. Strukture proteina

Bilo koje od ovih stanja može biti prirodno stanje, tj. stanje koje je bitno za biološku funkciju. Neki neuređeni proteini mogu da prelaze iz neuređenog u uređeno stanje i obrnuto posle interakcije sa drugim makromolekulima ili posle promena u biohemijskim procesima, dok drugi ostaju u neuređenom obliku u toku obavljanja svoje funkcije. U Disprot bazi (DB) nalazi se preko 500 proteina koji sadrže neuređene regione različite dužine. U skladu sa odnosom strukture i funkcije proteini su svrstani u 17 kategorija [15]. Na slici 2. je ilustrovan proces predviđanja različitih 3D strukturnih regiona proteina na osnovu primarne strukture proteina (sekvence amino kiselina).



Slika 2. Uversky and Dunker 2006 [14]. Prikazano je predviđanje neuređenih regiona prema prediktoru PONDR VL_XT za dva proteina: Hirudin i Trombin. Njihova struktura je poznata i prikazana je na istom grafiku. Različite strukture su obojene različitim bojama radi poređenja sa rezultatima dobijenim predikcionom metodom. Žuta boja: N –

terminal u lancu Hirudina; Crvena boja predstavlja C – terminal. Plava i zelena boja kod Trombina predstavljaju gust i jednostavniji lanac, respektivno.

Do danas je poznat veliki broj funkcija neuređenih proteina: njihovo vezivanje sa drugim molekulima, kontrolni mehanizmi DNA regiona, aktiviranje enzima, životni vek proteina. Trenutno verovanje je da su neuređeni regioni takvi jer im to daje prednost da:

a) imaju veću površinu,
b) imaju konformacionu fleksibilnost da se vezuju za više "partnera" ,
c) imaju elemente molekularnog prepoznavanja koji prelaze u uvrnutu strukturu nakon vezivanja,
d) imaju pozicije koje se post-translatorno modifikuju,
e) obično sadrže kratke linearne motive koji su važni za interakciju proteina sa ligandima.

Iz ovoga sledi da su neuređeni proteini uključeni u najvažnije biološke procese kao što su ćelijska signalizacija, prepoznavanje, regulacija ćelijskog ciklusa koji su podeljeni na više od 30 podklasa. Kako ne postoji univerzalna definicija neuređenih regiona, ovde se podrazumeva da svaka amino kiselina u proteinu pripada ili uređenom ili neuređenom regionu.

Na nivou primarne strukture neuređene regione karakteriše slaba složenost (sastoje se od ponavljajućih kratkih fragmenata). Sadrže pretežno polarne i šaržirane amino kiseline[1], a retko voluminozne hidrofobne A.K.. Neuređeni regioni sadrže u povećanom broju: alanin (A), arginin (R), glicin (G), glutamin (Q), serin (S), glutaminsku kiselinu (E), lizin (K) i prolin (P), a u manjem broju A.K. koje pospešuju stvaranje uređene strukture kao što su triptofan (W), tirozin (Y), fenil-alanin (F), izoleucin (I), leucin (L), valin (V), cistein (C), asparagin (N) [16]. Koristeći TOP-IDI skalu bazi autori rada [16] su rangirali amino kiseline prema osobinama koje promovišu prelazak uređene u neuređenu strukturu kao što su hidrofobnost, polarnost, šarža i volumen. Redosled je W, F, Y, I, M, L, V, N, C, T, A, G, R, D, H, Q, K, S, E, P. U ovom radu se koristi jedna od metoda koja sa dovoljnom dobrom tačnošću identifikuje neuređene regione nezavisno od dužine: VSL2 prediktor.

1.3 Predviđanje neuređenih regiona u proteinu – VSL2 predictor

Predviđanje neuređenih proteina, odnosno proteina koji imaju bar jedan neuređen region, je od izuzetnog značaja u biologiji zbog funkcionalnih osobina takvih regiona. Za razvoj softvera koji bi predvideo neuređene regione analizirana je eksperimentalno dobijena struktura sa neuređenim regionima. Na taj način je razvijen model koji je treniran nad neuređenim regionima, dobijenim različitim eksperimentalnim metodama. Model je nazvan VSL2 prediktor. Program VSL2 prediktor koristi nekoliko ulaznih atributa: hidrofobnost, prisustvo određene kombinacije amino kiselina, šaržu, itd. Odnosno zasnovan je na fizikohemijskim osobinama amino kiselina koje ulaze u sastav proteina.

[1] U daljem tekstu A.K.

1.3.1 Kako je nastao VSL2 prediktor

Program VSL2 za predviđanje neuređenih regiona je nastao objedinjavanjem dva postojeća programa za predviđanje neuređenih regiona: VSL2-M1 i VSL2-M2. Ovi programi su optimizovani za pronalaženje neuređenih regiona u zavisnosti od dužine:

- VSL2-M1 pronalazi neuređene regione veličine manje ili jednake 30 amino kiselina,
- VSL2-M2 pronalazi neuređene regione dužine preko 30 amino kiselina.

Kasnije su ovi programi integrisani u jedan program za predviđanje nazvan VSL2 prediktor, koji jednako dobro predviđa i kraće i duže neuređene regione.

VSL2 program je postigao veliku tačnost na unakrsnim proverama od 81% u oba slučaja (i kratkih i dužih neuređenih regiona).

1.3.2 Šta se krije u pozadini?

Suštinski neuređeni/ neuvijeni/ nestrukturisani proteini odnosno na dalje samo neuređeni proteini ne podležu stabilnoj 3D strukturi pod osnovnim fiziološkim uslovima. Bez obzira na nedostatak specifične 3D strukture, ispostavilo se da ovi proteini i proteinski regioni nose veoma važne biološke funkcije (navedene u uvodu).

Programi koji predviđaju neuređene / uređene regione u proteinu (3D strukturu, a time i funkciju) uzimaju u obzir primarnu strukturu proteina (redosled i sastav amino kiselina). Većina postojećih programa za predviđanje neuređenih regiona koristi klizni prozor za pridruživanje individualnog simbola (amino kiseline) u određeni komponentni prostor (engl. “feature space”), gde uz pomoć binarnog klasifikatora klasifikuje simbole kao uređene ili neuređene uz pomoć različitih algoritama mašinskog učenja. Komponente (stavke) se izdvajaju iz niza amino kiselina kroz prozor koji predstavlja kompozicionu osnovu i jedinstvene osobine karakteristične za neuređen region.

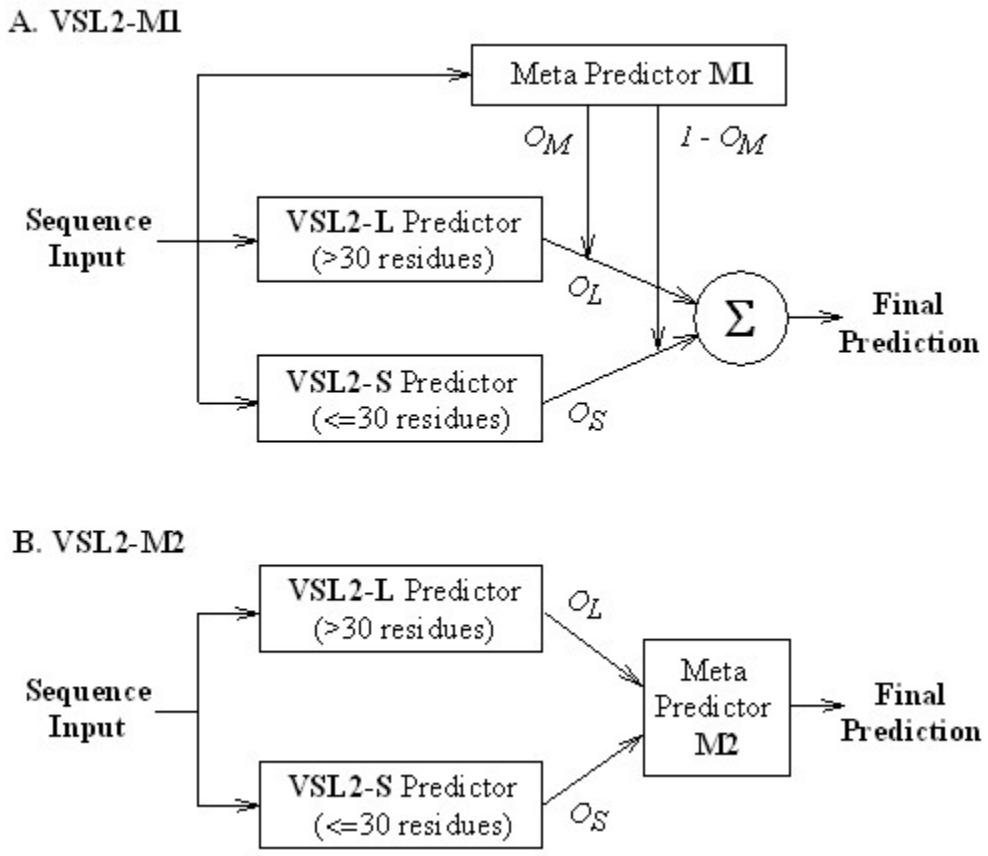
Neki prediktori (VL3 i DISOPRED2) izdvajaju komponente pomoću PSI-BLAST generisanog profila kako bi uključile evolucione informacije. Dokazane performanse ovih pristupa su u skladu sa zaključcima da neuređeni regioni imaju različite evolucione karakteristike.

1.3.3 VSL2 arhitektura

Oba VSL2-M1 i VSL2-M2 se sastoje od tro - komponentnog programa dvoslojne arhitekture. Na prvom nivou su specijalizovana dva prediktora: prediktor za kraće neuređene regione VSL2-S za regione kraće ili jednake 30 amino kiselina i prediktor za duže neuređene regione VSL2-L za regione sa preko 30 amino kiselina u neuređenim delovima.

U drugom nivou je meta-prediktor koji kombinuje izlaze oba prediktora u konačno predviđanje. Sve komponente prediktora su napravljene kao binarni klasifikator koji aproksimira posteriornu verovatnoću klase $p(c=1|x)$, gde je x komponenta (ulazni vektor), a c je labela klase. Za oba prediktora klasa 1 predstavlja neuređeni region a 0 uređeni region.

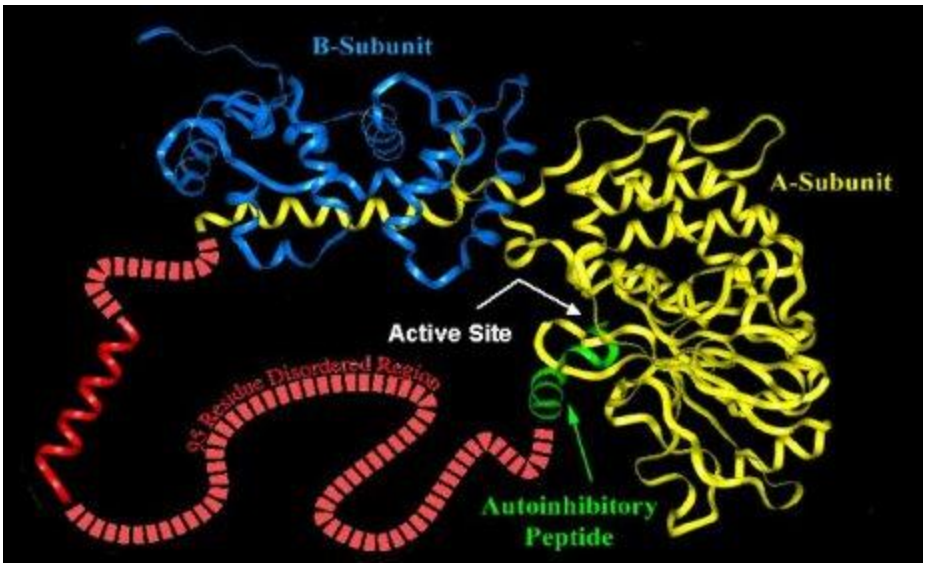
Arhitektura VSL2 programa za predviđanje je prikazana na slici 3.



Slika 3. Arhitektura VSL2 programa za predviđanje. Konačno predviđanje za VSL2_M1 se računa kao OL × OM + OS × (1 – OM), dok je za VSL2_M2 to izlaz prediktora M2. Ulaz za prediktor M2 su 2 × Win predikcije programa VSL2_S i VSL2_L za susedne simbole amino kiselina u proyoru dužine W.

1.4 **Primer izlaza i rezultata VSL2 programa**

VSL2 program uzima kao ulazni argument proteinsku sekvencu u FASTA formatu, a kao rezultat vraća, za svaku amino kiselinu date sekvence, predviđanje da li pripada uređenom ili neuređenom regionu. Mera predviđanja uzima vrednosti iz intervala [0, 1], i predstavlja verovatnoću sa kojom amino kiselina pripada neuređenom regionu. Amino kiseline kojima je, na ovaj način, pridružena mera veća od 0.5 pripadaju neuređenim regionima. Za protein čija je struktura prikazana na slici 4.:



Slika 4. Primer proteina sa neuređenom strukturom. Neuređeni regioni u proteinu su prikazani crvenom isprekidanom linijom.

dobija se sledeći izveštaj iz VSL2 programa (prikazan je samo deo rezultata):

Predicted Disordered Regions:
1-4
28-37
65-67
77-84
95-108
265-270
350-445

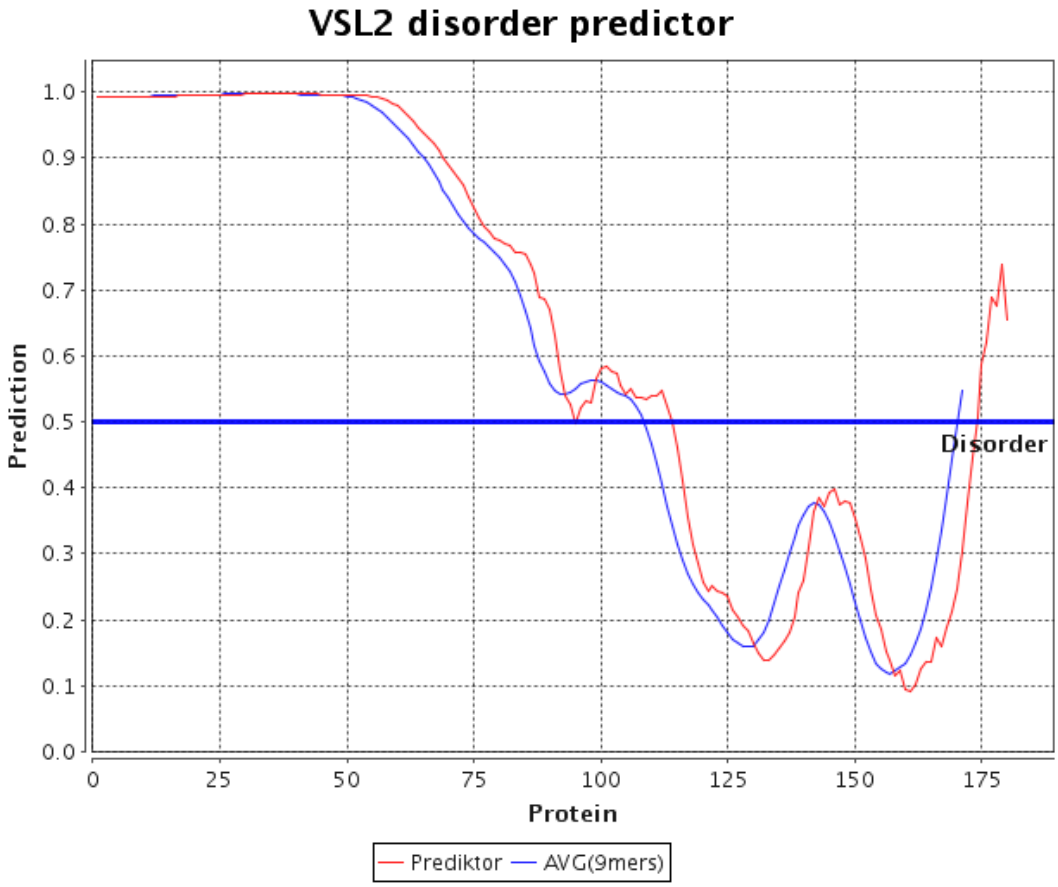
Prediction Scores:

NO.	RES.	PREDICTION	DISORDER
1	W	0.818899	D
2	G	0.743443	D
3	A	0.654804	D
4	L	0.518968	D
5	G	0.384689	.
6	H	0.295883	.
7	A	0.232575	.

8	T	0.187107	.
9	V	0.164851	.
10	A	0.175194	.
11	Y	0.193998	.
12	V	0.197843	.
13	A	0.215612	.

Rezultat pokretanja VSL2 programa se upisuje u tekstualnu datoteku u kojoj su na početku izdvojeni intervali koji su predviđeni kao neuređeni, a zatim je data ocena (mera predviđanja) za svaku amino kiselinu u proteinskoj sekvenci.

Grafički prikaz rezultata, dobijen pokretanjem EPDIS aplikacije (opisane u poglavlju 4.), koja obrađuje rezultate izlazne datoteke programa VSL2, dat je na slici 5.:



-

Slika 5. Grafički prikaz rezultata VSL2 programa, dobijen pokretanjem EPDIS aplikacije koja obrađuje izlaz VSL2 programa.

Crvena linija na grafiku predstavlja verovatnoću sa kojom svaka amino kiselina proteinske sekvence pripada neuređenom regionu. Granična vrednost koja razdvaja uređeni od neuređenog

regiona je 0.5. Na x osi je označena pozicija svake amino kiseline u sekvenci, a na y osi verovatnoća dobijena VSL2 programom. Plava linija na grafiku predstavlja srednju vrednost predikcione mere (verovatnoće) za 9 uzastopnih amino kiselina, čime je omogućeno analiziranje potencijalnih antigenih determinanti i njihova pripadnost uređenom / neuređenom regionu.

1.5 DISPROT baza podataka

DisProt baza podataka sa neuređenim proteinima povezuje strukturne i funkcionalne informacije o neuređenim proteinima. Zbog nedostatka organizovanih informacija o neuređenim regionima je upravo i napravljena ova baza podataka, koja bi omogućila dalja istraživanja nad neuređenim proteinima. Baza je javna i dostupna na http://www.disprot.org.

Standardna paradigma sekvenca – struktura - funkcija je zasnovana na činjenici da proteini podležu stabilnoj 3D strukturi i da upravo ta struktura uslovljava funkciju proteina, odnosno da postoji model koji predstavlja odnos strukture i funkcije. Enzimi zadovoljavaju ovakav model ponašanja, koji univerzalno objašnjava enzimatične funkcije. Međutim, proteinske funkcije vezane za signaliziranje, regulaciju i kontrolu kao što su protein-protein interakcija, protein-DNA interakcija, protein RNA-interakcija, post-translaciona modifikacija i aktivnosti povezivanja se i dalje proučavaju. Za mnoge od ovih aktivnosti se pretpostavlja (za neke je pokazano) da su uslovljenje proteinima koji ne podležu fiksnoj 3D strukturi. Takođe je i pokazano da ne samo da jedan neuređen protein može da se vezuje za nekoliko različitih proteina, već i da više neuređenih sekvenci mogu da se adaptiraju da odgovaraju jednom partneru.

Broj eksperimentalnih rezultata koji opisuju neuređene regione se brzo povećava baš iz razloga velikog interesovanja za funkcije signaliziranja, regulacije i kontrole. Odakle je i nastala potreba da se napravi odgovarajuća baza podataka sa informacijama o neuređenim regionima. Verzija baze koja se u ovom radu koristi je 4.9 i ima 517 neuređenih proteina, svrstanih u različite funkcionalne kategorije. Svi podaci su zasnovani na objavljenim eksperimentalnim rezultatima.

Disprot baza je implementirana kao relaciona baza podataka koristeći PostgreSQL. Disprot je podržana Apač veb serverom sa interfejsom implementiranim u PHP-u i JavaSkriptu. Neuređeni proteini mogu da se dobiju u fasta ili XML formatu. Za svaki protein je osim sekvence dato zaglavlje sa informacijama o proteinu (šifri i bazi na koju se odnosi, nazivu proteina kao i intervali sa neuređenim regionima dobijeni eksperimentalnim putem) [4].

Proteini iz DisProt baze predstavljaju samo jednu od grupa proteina koji se u ovom radu analiziraju. Detaljan opis proteina koji su obrađeni u radu je prikazan u poglavlju 3.

1.6 **Imunološki odgovor**

Antigen (skr. Ag, od prvobitnog eng. „antibody generator") je molekul koga prepoznaje imunološki sistem organizma, dok je epitop region ili fragment antigena koji se vezuje za odgovarajuće receptore na Ag-vezujućim ćelijama imunološkog sistema. Imunološkii sistem čine organizovana tkiva koja brane organizam od stranih molekula, infektivnih mikroorganizama i njihovih toksina. Postoje dva tipa imunološkog odgovora:

- Urođeni imunitet koji je nespecifičan i bez imunološke memorije i čini prvu liniju odbrane od stranih mikroorganizama.
- Adaptivni (stečeni) imunitet koji čine humoralni imunitet i ćelijski imunitet.

Adaptivni imunitet se javlja kod kičmenjaka, specifičan je za određeni antigen, ima imunološku memoriju i javlja se kasnije u toku imunloškog odgovora nego urođeni. Deli se na humoralni i ćelijski imunitet.

Humoralni imunitet se tako naziva jer se molekuli proteina (antitela), koji su glavni nosioci ovog tipa imuniteta, nalaze u telesnim tečnostima. Stvaraju ih ćelije koje se zovu B limfociti ili B ćelije. Antitela prepoznaju antigene, neutrališu infekcije izazvane mikroorganizmima, tako što ih uništavaju različitim mehanizmima odbrane. Humoralni imunitet je glavni mehanizam odbrane od mikroorganizama koji napadaju ćelije spolja, i usmeren je, uglavnom na prostorne (nelinearne ili diskontinualne) epitope antigena.

Ćelijski imunitet (ili ćelijama posredovani imunitet) se zasniva na T-limfocitima (ili T-ćelijama), i usmeren je na linearne epitope antigena. Jedna grana ćelijskog imuniteta (Th, Tr) ima ulogu da reguliše, kako adaptivni, tako i urođeni imunitet i odlučuje kakav tip imunološkog odgovora telo indukuje na određeni patogen. Usmeren je uglavnom na antigene iz spoljne sredine, kao što su Ag bakterija, (egzogeni put unošenja Ag) koje ćelije (nazvane „profesionalne Ag-prikazivačke ćelije"), unose endocitozom, degradiraju i „predstavljaju" na ćelijskoj površini. Druga grana ćelijskog imuniteta su citotoksični T limfociti (Tc). Ovaj put je usmeren, uglavnom na kontrolu sopstvenih, unutarćelijskih proteina i eliminaciju utrošenih proteina (endogeni put prezentacije Ag). Ako virus inficira ćeliju, viralni peptidi (epitopi) će biti predstavljeni preko ovog puta, omogućujući Tc limfocitima da prepoznaju i ubiju inficiranu ćeliju. I B i T limfociti nose na ćelijskoj membrani receptorne molekule (kod B limfocita su to antitela, a kod T limfocita T-ćelijski receptori, skr. TCR, od eng. „T-cell receptor").

Imunološki odgovor čini prepoznavanje antigena, aktivacija limfocita i efektorna faza eliminacije antigena. Adaptivni imunološki odgovori su inicirani prepoznavanjem specifičnih antigena. Adaptivni imunološki sistem sisara je evoluirao tako da izlaže fragmente (epitope) proteina, koji potiču od mikrobnih patogena (antigena), kao i sopstvene proteine (kao stalnu kontrolu sopstvenog imuniteta) ćelijama imunološkog sistema. Ove ćelije se dele na antigen-prikazivačke, efektorne i regulatorne. Fragmenti su peptidi, dužine do 25 aminokiselina koji se oslobađaju iz intaktnih proteina preko proteolitičkih mehanizama koji se odvijaju u

specijalizovanim organelama antigen-prikazivačkih ćelija. U narednom koraku se prenose na površinu ćelija u kompleksu sa proteinima glavnog histokompatibilnog kompleksa organizma, da bi ih (u kompleksu) prepoznale efektorne ćelije imunloškog sistema. Ćelije imunloškog sistema koje prepoznaju komplekse su pomažući / regulatorni (eng. „helper/regulatory", skr. Tr ili Th) T limfociti koji nose i oznake (T4 ili CD4) i citotoksićni T limfociti koji nose oznake T8 ili CD8 (što se naziva predstavljanje antigena). Molekuli glavnog histokompatibilnog kompleksa (eng. „major histocompatibility complex")[2] su genski regioni ili familije gena. Sastoje se od dve podklase „major histocompatibility complex I" i „major histocompatibility complex II" (skr. MHC I i MHC II) [18]. Kod čoveka nose naziv HLA I i HLA II, od eng. „human leukocyte antigens", jer su prvobitno imunološki definisani kao antigeni na leukocitima prilikom transfuzija krvi. Njihove kombinacije predstavljaju individualnu tkivnu i imunološku specifičnost organizma, koja je genetski definisana (genskim alelima klase MHC I i II). MHC molekuli imaju važnu ulogu u imunološkom sistemu i autoimunosti. Kao što je već opisano, antigeni (epitopi) se vezuju sa molekulima ovog kompleksa i prikazuju na površini ćelije. Postoji pet tipova gena HLA molekula klase I : HLA-A, HLA-B, HLA-C, HLA-E i HLA-G, a za HLA molekule klase II postoje tri lokusa: HLA-DP, HLA-DQ i HLA-DR.

HLA genski aleli su kodominanti i u jednom čoveku su najčešće izraženi kroz 6 različitih molekula klase MHC I i 12 ili više molekula MHC klase II. Inače postoji preko hiljadu HLA alela u celoj populaciji (tačnije 1469 klase I i preko 517 klase II). HLA lokus je najpolimorfniji poznati genski sistem. HLA aleli predstavljaju jedna od više formi DNA sekvence, a vezuju veliki spektar različith peptida, izvučenih iz 1000 do 10.000 proteinskih sekvenci - antigena. Osobina vezivanja sa peptidima potiče iz polimorfizma HLA molekula.

Pronalaženje T-ćelija CD8 i CD4 je važno za razumevanje patogeneze bolesti i predstavlja osnovu za razvijanje vakcine zasnovane na epitopima protiv infekcija, alergija, autoimunih bolesti, kancera, itd. Najselektivniji način u njihovom prepoznavanju je upravo pronalaženje epitopa (peptida) koji se vezuju za MHC molekule.

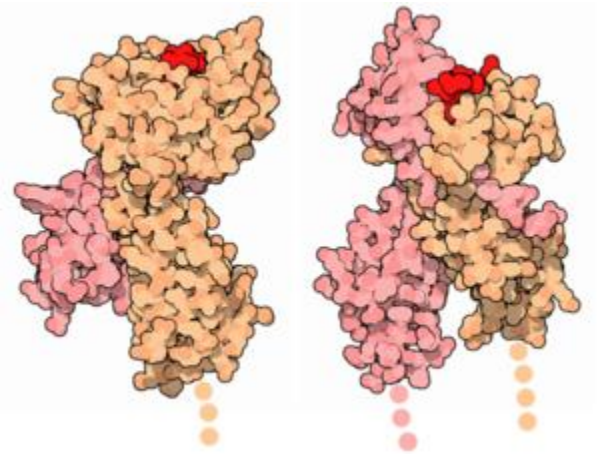
1.7 **Strukturna osnova MHC-peptid vezujućih regiona**

3D struktura klasičnih MHC I i MHC II molekula je neverovatno slična (slika 6), uprkos činjenici da je sličnost njihovih proteinskih sekvenci ispod 20%. Različiti MHC molekuli predstavljaju različite podskupove peptida (epitopa). MHC molekuli se vezuju sa kratkim peptidima iz antigena i prikazuju ih na površini T-ćelijama. Mehanizam vezivanja je za sada najselektivniji korak u identifikaciji T-ćelija. MHC aleli su grupisani po svojoj strukturi. Predviđanje vezivanja sa MHC klasom I je do sada dobro proučeno, i metode koje predviđaju vezivanje peptida sa molekulima klase MHC I su velike tačnosti (čak do 95% jer ovi peptidi imaju ograničenu dužinu i dobro je poznato kako dolazi do njihovog „isecanja"). Predviđanje

[2] „Major histocompatibility complex" u daljem tekstu MHC

vezivanja antigenih epitopa sa molekulima MHC klase II je nešto slabije tačnosti, oko 81% , i još uvek je nedovoljno istraženo.

Glavni problem u predviđanju peptida koji se vezuju sa molekulima MHC klasa je što se u 100 do 200 peptida pronalazi samo jedan koji se vezuje sa navedenim molekulima. Taj peptid se naziva epitop. Predviđanje peptida koji se vezuju za molekule MHC klase II je znatno teže nego za MHC klasu I zbog različite strukture proteina ove dve klase. Struktura MHC klasa je prikazana na slici 7.:

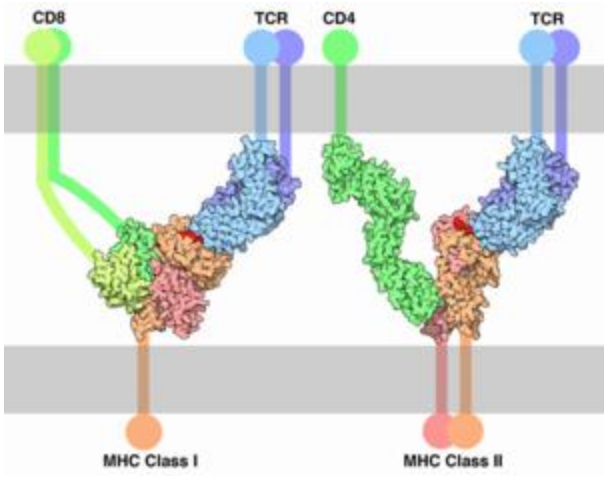


Slika 6. MHC I (levo) i MHC II lanac (desno)

MHC I proteini se nalaze u svim ćelijama sa jedrima. Sastoje se od alfa lanca i beta2 mikroglobulina. Vezuju antigeni fragment (epitop) i predstavljaju ga citotoksičnim T-ćelijama preko T-ćelijskog receptora (TCR) uz vezivanje za CD 8 koreceptor, (slika 7). CD8 koreceptor je receptor koji karakteriše citotoksične T limfocite (pa se stoga nazivaju T8 ili CD8 limfociti).

MHC II proteini se sastoje od dva lanca: alfa i beta lanca. Vezuju antigeni fragment (epitop) i predstavljaju ga pomažućim T-ćelijama preko T-ćelijskog receptora (TCR) uz vezivanje za CD 4 koreceptor, (slika 7). CD4 koreceptor je receptor koji karakteriše pomažuće T limfocite (pa se stoga nazivaju T4 ili CD4 limfociti).

Molekuli MHC klase II su otvoreni na oba kraja, dok su molekili klase I zatvoreni na svakom kraju. Antigeni prikazani klasom MHC II su duži (obično između 15 i 24 amino kiseline), dok je za klasu MHC I dužina peptida između 8 i 11 amino kiselina.



Slika 7. Strukture molekula MHC klase I (levo) i MHC klase II (desno)

MHC klase I i II se razlikuju i u načinu "predstavljanja" peptida. Epitopi koji se vezuju za molekule MHC klase I su dobro karakterizovani, i uspostavljena su neka pravila za pojavljivanje amino kiselina na drugom i devetom mestu. Sa druge strane, epitopi koji se vezuju za molekule MHC klase II imaju više od jedne hidrofobne amino kiseline što dozvoljava višestruko moguće ravnanje.

1.8 **Antigeni regioni (epitopi) i struktura proteina – „širenje epitopa"**

Epitopi mogu biti klasifikovani kao linearni (kontinualni) i prostorni (diskontinualni). Linearne epitope prepoznaju prvo (i pretežno) T limfociti, a prostorne B limfociti. Softveri koji su zasnovani na proteinskim sekvencama (primarnoj strukturi proteina) daju predviđanje T-ćelijskih epitopa (za pomažuće ili citotoksične limfocite) i sa uspehom se koriste u pravljenju vakcina duže od decenije, tako što sužavaju izbor proteinskih antigena i smanjuju broj eksperimenata [19]. Većina ovih programa predviđa vezivanja linearnih sekvenci amino kiselina (peptida veličine 9-11) određenog antigena za MHC I ili MHC II molekule i ne uzima u obzir ćelijsku lokalizaciju antigena, njegovu specifičnu proteolizu i uticaj 3D strukture. Analize poznatih citotoksičnih i pomažućih T-ćelijskih epitopa su pokazale određene obrasce u aminokiselnskom sastavu, koji mogu biti specifični za određenu grupu MHC alela [20]. T-ćelijski epitopi predstavljaju peptidne fragmente od 9-12 amino kiselina, koji najčešće čine amfipatske helikse (helikse koje čini pravilna smena hidrofilnih i hidrofobnih aminokiselina, tako da je jedna strana heliksa hidrofilna, a druga hidrofobna). Priroda linearnih T-ćelijskih epitopa definisana je, dakle, pretežno kao, uređena struktura (heliks). Ipak do danas nema sistematskih podataka o zastupljenosti i karakteru (hidrofobnost, polarnost, šarža) epitopa u uređenim i neuređenim delovima proteina. Takođe se veoma malo zna o uticaju konformacije (3D strukture) na indukovanje proteinskih antigena u antigen-prikazivačkim ćelijama. Pokazano je da kod određenih proteina ova zavisnost postoji, [21, 22], što se, sledstveno, odražava na broj i vrstu epitopa. Kako se vrste antigen-prikazivačkih ćelija i mehanizmi obrade antigena (MHC I i

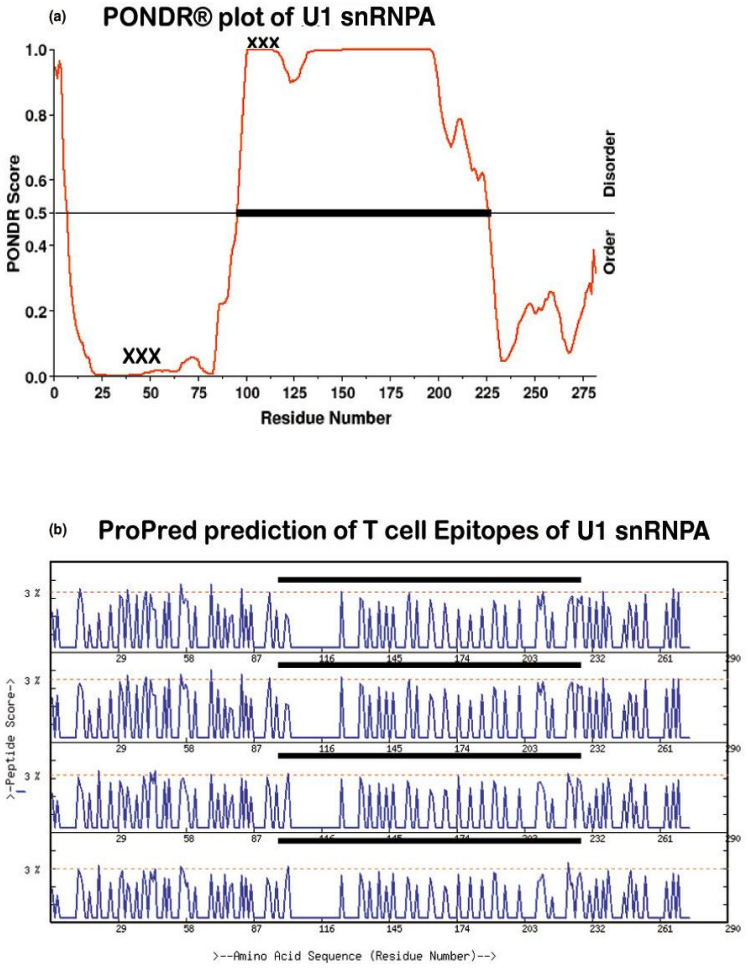
MHC II putevi) razlikuju [18], to se i epitopi jednog istog antigena (antigene determinante) za ova dva puta (klase MHC I i II) razlikuju.

Do sada nije sistematski analiziran odnos uređene / neuređene strukture u proteinu sa brojem i karakterom epitopa. Poznato je da u organelama u kojima se proteolitički razgrađuju antigeni dolazi do „rasplitanja" proteina i da je u proteozomima, organelama koje vrše preradu antigenih epitopa koji se vezuju za MHC-I molekule, ono ATP-zavisno (energetski zavisno). Pokazano je i da 3D struktura kontroliše i prikazuje epitope T4 limfocitima (MHC II put), gde pod dejstvom edoproteaza dolazi do delimičnog „otključavanja" 3D strukture [22]. Energetski favorizovano rasplitanje moglo bi da utiče da se u MHC-II putu prikazuje ukupno više epitopa koji se nalaze u uređenim strukturama proteina.

Pretpostavlja se da neuređeni regioni zbog svoje konformacione fleksibilnosti, maskirane nukleinskim kiselinama i drugim proteinima čine siromašne antigene, „nevidljive" za imunološki sistem, naročito za B, ali i T limfocite. Poznato je da 90% B limfocita prepoznaje konformacione epitope (uređena struktura) na antigenima [19]. Izrazito neuređeni proteini su, takođe, veoma osetljivi na dejstvo proteaza i pretpostavlja se da bi epitopi iz ovih regiona imali slab afinitet za MHC II molekule, što čini da ne bi bili efikasno prikazani kao T-ćelijski antigeni [1]. Kako T-pomažući limfociti ćelije, koji prepoznaju epitope MHC II antigena, imaju ulogu da predstavljaju epitope kako citotoksičnim tako i B limfocitima to bi imunološki odgovor ovih limfocita na epitope u neuređenim regionima proteina bio slabo zastupljen u ukupnom repertoaru imunološkog sistema. Ali bi, upravo zbog slabog afiniteta vezivanja za ove epitope, imali mogućnost da izbegnu imunološko brisanje, tj. odstranjvanje limfocita koji reaguju sa sopstvenim antigenima u toku fetalnog razvoja [26]. Ova hipoteza, koju su izneli Karl i saradnici u radu [1] je bila osnova za model nastanka sistemskog autoimuniteta (patološkog imunološkog odgovora na sopstvene-Ag, koji se javlja na nivou celog organizma). U radu [1] je korišćen program PONDR za predviđanje neuređenih regiona, dostupnom na adresi http://www.pondr.com/, i ProPred program za predviđanje epitopa. Pokazano je da većina sistemskih nuklearnih (lociranih u jedru ćelije) autoantigena predstavlja ekstremno neuređene proteine. Zašto neki proteini postaju autoantigeni nije dovoljno poznato, tako da je ovaj rezultat značajan doprinos određivanju svojstava autoantigenih proteina. Među strukturnim osobinama proteina, za koje se pretpostavlja da dovode do pojave autoimuniteta su visoko šaržirani i ponavljajući površinski elementi, vezane nukleinske kiseline i struktura uvrnutog klupka (eng. „coiled-coil"), što predstavlja elemente i neuređene i uređene strukture. U istom radu je pretpostavljeno (a za nekoliko primera i dokazano), da u ekstremno neuređenim autoantigenima, u dugačkim neuređenim regionima skoro da nema (ili uopšte nema) epitopa koji se vezuju za molekule MHC-II klase (pri čemu je praćena učestalost, ali ne i visina „vrhova" koja ukazuje na afinitet vezivanja, odnosno bolju mogućnost prezentovanja epitopa).

Da li se neuređeni regioni ikada preklapaju sa antigenim regionima (epitopima). Na ovo pitanje odgovor je – „da"

Na slici 8. je prikazan rezultat dobijen u radu [1] za protein U1 snRNPA, gde se vidi da postoje epitopi i u neuređenim regionima. U istom radu je razmatran i protein EBNA1 i utvrđena je korelacija između uređenih regiona i epitopa. Bez obzira što protein EBNA1 ima dugačak neuređen region, epitopi su prepoznati samo u uređenim regionima. Iako je već rečeno, treba naglasiti da se u [1] pod epitopima ne podrazumevaju peptidi koji zadovoljavaju određeni nivo afiniteta vezivanja, već je razmatrana učestalost "vrhova" na dijagramu. U daljem radu se pod epitopima podrazumevaju peptidi koji zadovoljavaju određeni nivo afiniteta vezivanja za molekule MHC kompleksa.



Slika 8. Predviđeni neuređeni regioni i T – ćelijski epitopi za protein U1 snRNPA. Na prvom grafiku je jaki imunogenski peptid označen sa XXX, koji izaziva širenje autoimune bolesti. Slabi imunogenski peptid je označen sa xxx, na koji se autoimuna bolest proširila. Rezultat programa ProPred je prikazan slikom b) i predstavlja epitope MHC klase II za alele: HLA*DRB1_01, HLA*DRB1_0102, HLA*DRB1_0301 i HLA*DRB_0305. Crna linija označava eksperimentalno dobijeni neuređen region [1].

Na osnovu raspodele epitopa u uređenim i u neuređenim regionima nekih nuklearnih autoantigena, izneta je pretpostavka da širenje imunološkog odgovora započinje na epitopima u uređenim regionima i da se može proširiti duž neuređenih regiona i tako dovesti do pojave autoimune bolesti. Širenje epitopa predstavlja ekstenziju imunološke reaktivnosti sa inicijalnog regiona jake antigenetičnosti kroz polipeptid na drugi epitop, ili sa epitopa jednog polipetida na drugi (najčešće susedni) polipeptid, što vodi mnogo bržem i intezivnijem sekundarnom odgovoru, kao i mnogo dužem imunološkom pamćenju. U autoimunitetu ono, najčešće, započinje „molekularnom mimikrijom“ ili unakrsnom-reaktivnošću, sličnošću određenog mikrobnog epitopa i epitopa domaćina. Fenomen „širenja“ imunološkog odgovora je, najverovatnije, normalna pojava u imunološkom odgovoru na mikroorganizme – jer je imunološki sistem evoluirao tako da napada što više ciljeva [27]. Autoimuno ili patogeno širenje epitopa, bi bilo posledica zakazivanja kontrolnih mehanizama imunološkog sistema. Kakva je tačno uloga epitopa u uređenim regionima u otpočinjanju širenja autoimunog procesa i kako tačno dolazi do aktivacije rezervoara B ćelija koje potencijalno napadaju neuređene regione proteina? Pretpostavlja se da razlog postoji u velikim razmerama i nivou ekspresije (ispoljavanja) proteina koji se ponašaju kao autoantigeni, polivalentne prirode većine nuklearnih sistemskih autoantigena i tome što se javljaju u makromolekularnim (kompleksima sa npr. nukleinskim kiselinama). Ove dve poslednje osobine su takođe i odlike neuređenih regiona proteina.

Suprotno fenomenu autoimuniteta postoji situacija kada je autoimuni odgovor poželjan, u antitumorskim vakcinama, kada imunizacijom fragmentima tumorskih-pridruženih Ag (TAA) treba izazvati imunitet na tumor. Tumor pridruženi antigeni su, kao i autoantigeni „sopstveni“ (eng. „self“) proteini, retko genetski izmenjeni, već uglavnom dolazi do promene njihove genske ekspresije (ispoljavanja). Da bi neki tumorski antigen postao potencijalni cilj za imunoterapiju, mora da ima ograničenu ekspresiju u normalnom tkivu, upravo da bi se sprečila pojava sistemskog autoimuniteta. Kancer-testis (CT) antigene čini 14 familija gena koji se učestalo ispoljavaju u različitim tumorima, ali je njihova normalna ekspresija ograničena na testise, fetalni ovarijum ili placentu, koji predstavljaju imunološki privilegovane zone organizma. Za jedan deo ovih antigena je nađeno da na njih postoji spontani humoralni i ćelijama-posredovani imunitet kod osoba obolelih od kancera, što ukazuje na to da u toku fetalnog razvoja nije došlo do brisanja klonova T i B limfocita, upravljenih protiv ovih antigena.

Funkcija većine ovih Ag je nepoznata, iako je verovatno da učestvuju u regulaciji genske ekspresije [28]. Utvrđeno je da regulatorni i kancer-pridruženi proteini imaju najmanje dvostruko više neuređenih struktura u udnosu na 10 drugih funkcionalnih kategorija ćelijskih proteina [8]. Kategorija kancer-pridruženih proteina (231 protein) u navedenom radu je imala ključne reči „oncogene“ „proto-oncogene“ ili „tumor“ i uglavnom je uključivala unutarćelijske regulatorne proteine. Može se pretpostaviti da bi i CT antigeni, kao pretežno regulatorni proteini, takođe imali veliki udeo neuređene strukture. Ukoliko bi se pretpostavka o vezi između epitopa i uređenih struktura proteina pokazala kao tačna za sve analizirane proteine i sve alele obuhvaćene

programima za predviđanje, na osnovu toga bi grupa tumorskih antigena mogla biti prva grupa antigena na kojoj bi se testirala tačnost metoda na eksperimentalnim rezultatima.

U ljudskom organizmu, kao što je prethodno rečeno, MHC molekuli su poznati kao HLA – eng. „Human Leukocyte Antigens“ i kodirani su sa HLA hromozom regionima.

1.9 Programi koji predviđaju antigene regione

Kako je već rečeno, uloga imunološkog sistema je odbrana od bolesti, virusa, infekcija, itd. Jedan pristup u proveri zašto i kada se indukuje imunološki odgovor je da se simulira unapređen model imunološkog sistema i da se analizira veza između domaćina i patogena. U zavisnosti od složenosti modela i datog ulaza, moguće je simulirati šta se dešava kada domaćin bude zaražen patogenom tj. uticaj patogena na imunološki sistem.

Jedan cilj modeliranja je pronalaženje delova proteina poznatih kao epitopi koje imunološki sistem prepoznaje, i na taj način indukuje odgovarajući imunološki odgovor. Poznavanje ovakvih reakcija je veoma važno za razvoj boljih vakcina i daje dobar uvid u prirodu kancerogenih oboljenja, alergija i autoimunih oboljenja.

Trenutno najpoznatiji programi za predviđanje epitopa su dati u tabeli 1.:

Serveri	Adrese	Cilj predviđanja
BIMAS	http://www-bimas.cit.nih.gov/molbio/hla_bind	MHC klasa I ligandi
MAPPP	http://www.mpiib-berlin.mpg.de/MAPPP	MHC klasa I ligandi i proteaze
NetChop	http://www.cbs.dtu.dk/services/NetChop	Proteaze
NetMHC	http://www.cbs.dtu.dk/services/NetMHC	HLA_A2 i H-2K
PAProC	http://www.paproc.de	proteaze
ProPred	http://www.imtech.res.in/raghava/propred	HLA-DR
ProPred I	http://www.imtech.res.in/raghava/propred I	MHC ligande klase I
SYFPEITHI	http://www.syfpeithi.de	MHC ligande klase I i II
RANKPEP	http://www.mifoundation.org/Tools/rankpep	MHC ligande klase I i II
SVMHC	http://www.sbc.su.se/~pierre/svmhc	MHC ligande klase I
Lib Score	http://www.ddbj.nig.ac.jp/analysesp-e	MHC ligande klase I
MHCPred	http://www.jenner.ac.uk/MHCPred	MHC ligande klase I
MULTIPRED	http://research.i2r.a-star.edu.sg/multipred	vezivanje sa MHC klasom I i II
TEPITOPE	http://www.vaccinome.com	vezivanje sa MHC klasom I i II
EpiMer	http://epivax.com	vezivanje sa MHC klasom I i II
IEDB	http://www.immuneepitope.org-tools.do	MHC ligande klase I i II, proteaze, vezivanje za MHC I

Tabela 1. Serveri za predviđanje T-ćelijskih epitopa

Postojeći programi za predviđanje se razlikuju u metodologiji predviđanja antigenih epitopa.

a) Najranija predviđana su zasnivana na izdvajanju motiva iz proteinskih sekvenci jer je utvrđeno da su peptidi koji se vezuju za određene MHC molekule funkcionalno srodni, i dele simbole (amino kiseline) sa sličnim osobinama na različitim pozicijama primarne sekvence. SYFPEITHI je primer programa za predviđanje epitopa, koji koristi ovu metodu. Program pronalazi peptide koji zadovoljaju osobine motiva koji se vezuju za neku od MHC klasa .
b) Prethodni način predviđanja je unapređen "Matricama povezanosti" (*engl*. "Binding matrices"). Konstruisane su matrice dimanzija *l* x 20, gde *l* predstavlja veličinu peptida a 20 je za simbol svake amino kiseline. Matrice su konstruisane izračunavanjem broja pojavljivanja svake amino kiseline na različitim pozicijama u peptidima već poznatim kao epitopi. Primer programa zasnovanog na ovoj metodologiji je: EpiMatrix, BIMAS.
c) Stabla odlučivanja: su modeli zasnovani na pravilima koja klasifikuju obrasce koristeći sekvence sa već poznatim, dobro ustanovljenim, pravilima. Stabla odlučivanja mogu da se primene i na linearne i nelinearne podatke, te se na ovoj metodologiji temelji veliki broj programa za predviđanje epitopa.
d) Veštačke neuronske mreže: modeli zasnovani na neuronskim mrežama su odgovarajući za klasifikaciju i prepoznavanje kompleksnih obrazaca. Mogu da kodiraju nelinearne podatke i iscrpno su korišćeni za predviđanje peptida koji se vezuju i za MHC klasu I i II. Peptidi su predstavljeni kao kompozicija simbola (amino kiselina). Simboli se koriste za treniranje mreže za klasifikovanje peptida na one koji se vezuju i one koji se ne vezuju sa molekulima neke od MHC klasa (eng. "binders", "nonbinders"). Metode veštačkih neuronskih mreža su pokazale znatno bolje rezultate nego sve ostale metode. Jedina mana ovog pristupa je što veštačke neuronske mreže zahtevaju ulaz fiksne dužine.
e) HMM (skraćeno od eng. "Hidden Markov models") – predstavlja grafički verovatnosni model, na kojem su zasnovani mnogi programi za predviđanje, koji sa velikom tačnošću prepoznaju statičke obrasce i klasifikuju statičke podatke. HMM modeli su razvijeni u cilju prevazilaženja nedostatka metoda zasnovanih na veštačkim neuronskim mrežama.
f) SVM (skraćeno od eng. "Support vector machine") modeli: su statističke metode zasnovane na principu minimizovanja strukturalnog rizika. Takođe pogodne i za linearne i nelinearne podatke. Svaki peptid se tretira kao vektor specifičnih stavki, kao što su: kompozicija amino kiselina, hidrofobnost, polarnost, itd. Parametri se treniraju mapiranjem ulaznih vektora u više-dimenzioni prostor stavki, zatim se maksimizira granica između epitopa i peptida koji to nisu sa optimalnom razdvajajućom hiper ravni. SVM modeli su prevazišli performanse modela zasnovanih na veštačkim neuronskim mrežama i stablima odlučivanja kada su podaci za treniranje manji.
g) Takođe postoje i metode zasnove na strukturi (eng. "Protein threading", "Homology modeling", "Docking").

Detaljnije objašnjenje svake od metoda, kao i lista programa napravljenih na osnovu tih metoda, se može naći u radovima [9,10], i nisu predmet ovog rada.

1.9.1 CBS Grupa i NetMHC programi

Grupa imunologa bioinformatičara (CBS skr. od eng. "Center of Biological Sequence Analysis" Tehničkog Univerziteta u Kopenhagenu, Danska) je razvila niz metoda u cilju pronalaženja epitopa, čija je svrha pronalaženje vakcine za HIV, malariju, tuberkulozu itd.

CBS grupa je razvila simulacioni model ljudskog imunološkog sistema i napravila bazu podataka sa svim ljudskim patogenima. Koristeći ovu bazu i bazu ljudskih genoma razvijene su metode i programi za predviđanje, koji simuliraju reakciju imunološkog sistema na patogene, i pronalaženje različitih epitopa imunološkog sistema. U većini projekata predviđeni epitopi su proveravani sa eksperimentalnim laboratorijskim rezultatima. Na ovaj način su razvijene metode za tri glavne grupe epitopa:

- B ćelijski epitopi koji se nalaze na proteinima, pretežno mikroorganizama, i koje prepoznaju B ćelije (limfociti) .
- Eitopi na pomoćnim ili regulatornim T limfocitima (skraćeno Th ili Tr). Ove ćelije luče supstance koje aktiviraju druge ćelije imunološkog sistema da unište, tolerišu ili daju alergijske odgovore na patogen.
- I epitope na citotoksičnim T limfocitima (skraćeno Tc). Ovi limfociti su zaduženi da pronađu i unište zaraženu ćeliju sopstvenog organizma.

Jedan od projekata ove grupe je Razvoj precizne metode za predviđanje vezivanja peptida za molekule MHC klasa I i II. Dva programa razvijena u tu svrhu su:

- NetMHCpan verzija 2.0 je metoda koja generiše kvantitativno predviđanje afiniteta bilo koje interakcije peptida sa MHC klasom I, zasnovana na metodi veštačkih neuronskih mreža. Omogućeno je predviđanje za sve peptide dužine od 8 do 11 amino kiselina, mada se za sve peptide koji nisu dužine 9 predviđanja dobijaju aproksimiranjem vrednosti dobijene za peptid veličine 9. Većina MHC molekula se pre vezuje za peptide upravo te veličine. Metoda je obučavana na velikom skupu dostupnih kvantitativnih MHC vezujućih podataka, i pokriva sve: HLA-A, HLA-B, HLA-C, HLA-G i HLA-E ljudske lokuse kao i šimpanze, majmuna i MHC klasu I miša. (http://www.cbs.dtu.dk/services/NetMHCpan/)
- NetMHCIIPan verzija 1.0 je metoda koja predviđa vezivanje peptida sa 517 različitih HLA-DR alela (MHC klase II) korišćenjem metoda veštačkih neuronskih mreža. (http://www.cbs.dtu.dk/services/NetMHCIIpan/)
- Tačnost programa za predviđanje epitopa je preko 86% za epitope MHC klase I i preko 81% za epitope MHC klase II. (Programi su testirani sa eksperimentalno dobijenim epitopima IEDB baze dostupne na adresi: http://www.immuneepitope.org/).

U ovom radu za predviđanje epitopa se koriste programi: NetMhcPan i NetMhcIIPan CBS grupe. Upravo ti programi su izabrani jer predviđaju epitope za sve postojeće ljudske alele. Takođe je od izuzetnog značaja činjenica da su obe metode pokazale odličnu tačnost predviđanja epitopa za različite grupe proteina.

1.10 **Primer rezultata programa NetMhcPan i NetMhciiPan:**

Oba programa se primenjuju na proteinsku sekvencu, zadatu u fasta formatu, gde se analiziraju svi mogući peptidi veličine 9 tako što se počinje od prvih 9 amino kiselina, a zatim klizno pokreće prozor sa leva na desno. Metode pri predviđanju uzimaju u obzir i peptide i HLA molekule. Za svaki peptid se daje kvantitavna ocena afiniteta vezivanja za određeni MHC molekul (peptid-HLA interakcija). Peptidi se klasifikuju u tri kategorije na osnovu unapred utvrđenih granica dobijenih eksperimentalnim putem. Podela se vrši na jake epitope, slabe epitope i one peptide koji nisu epitopi (ne vezuju se za MHC molekule). Osim mere afiniteta metode pridružuju i meru predviđanja koja se dobija kao **1-logk(aff),** a predstavlja skaliranu vrednost afiniteta na intervalu [0, 1]. Navedene metode su pokazale odličnu mogućnost razlikovanja epitopa i ne-epitopa. Predviđeni epitopi za prethodno ne testirane sekvence su testirani unakrsnim proverama i pokazali su veliku tačnost u predviđanju HIV imunoloških epitopa i endogenih peptida (95%). Kako metode uzimaju u obzir sve HLA molekule, pogodne su za globalno analiziranje imunoloških odgovora i epitopa koji nisu vezani samo za genome i patogene već sve HLA epitope.

Obe metode su dostupne i za interaktivan rad na predikcionom serveru na CBS-u. Podaci se unose u vidu veb forme, i dozvoljen je slobodan pristup svim akademskim korisnicima. Akademski korisnici mogu da dobiju programe i kao samostalne softverske pakete, za instaliranje i pokretanje na lokalnoj mašini. Uputsvo za instaliranje i korišćenje je dato na adresi: http://www.cbs.dtu.dk/services/. Obe metode su pisane za UNIX okruženje.

Pokretanjem programa za predviđanje epitopa NetMhcPan dobija se sledeći izveštaj (prikazan je samo deo rezultata):

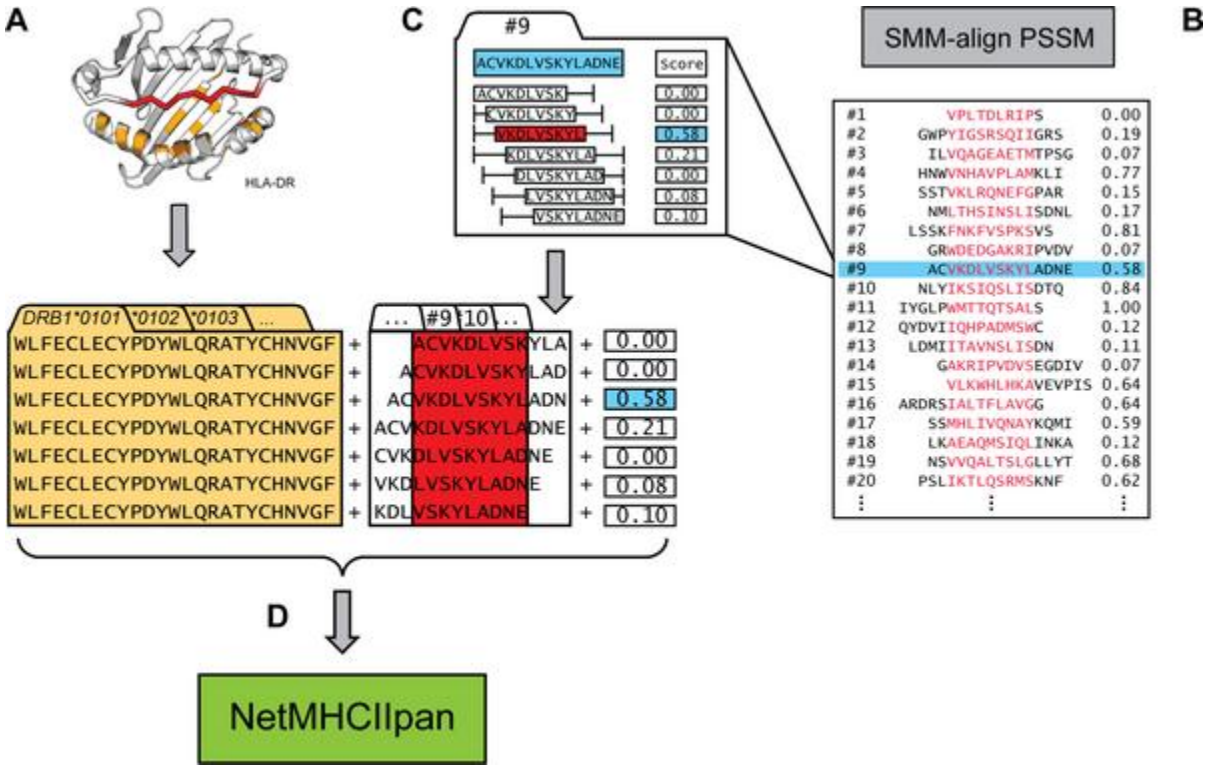
---

pos	HLA	peptide	Identity	1-log50k(aff)	Affinity(nM)	Bind Level
0	HLA-A*0201	TMDKSELVQ	143B_BOVIN_(P29	0.060	26191.54	
1	HLA-A*0201	MDKSELVQK	143B_BOVIN_(P29	0.011	44254.71	
2	HLA-A*0201	DKSELVQKA	143B_BOVIN_(P29	0.013	43661.29	
3	HLA-A*0201	KSELVQKAK	143B_BOVIN_(P29	0.018	41206.00	
4	HLA-A*0201	SELVQKAKL	143B_BOVIN_(P29	0.027	37222.38	
5	HLA-A*0201	ELVQKAKLA	143B_BOVIN_(P29	0.054	27789.56	

Izveštaj je u vidu tekstualne datoteka čija prva kolona predstavlja poziciju peptida u proteinu za koji se računa afinitet vezivanja sa molekulima MHC klase I. U ovom slučaj to je molekul

predstavljen alelom HLA-A*0201, i prikazan je u drugoj kolini. U trećoj koloni je sekvenca amino kiselina koja čini razmatrani peptide. Zatim sledi šifra proteina za koji se predviđaju epitopi. I na kraju izračunati afiniteti vezivanja kao i skalirana vrednost afiniteta. Rezultat pokretanja programa NetMhcIIPan daje sličan izveštaj, ima samo jednu kolonu više u kojoj su smeštene oznake SB – za jake epitope (kada je afinitet između 0 i 50), WB – za slabe epitope (50 > afinitet $\leq$ 500).

Na slici 9. je prikazan postupak izdvajanja peptida metodom NetMhcIIPan:



Slika 9. Šematski prikaz metode NetMHCiiPan

1.11 **Indeks hidropatije**

Hidropatija predstavlja ponašanje amino kiselina u vodenoj sredini. Amino kiseline su hidrofobne ili hidrofilne. U tabeli 2. su dati indeksi hidropatije za sve amino kiseline. Amino kiseline sa indeksom hidropatije većim od nule su hidrofobne, one sa indeksom manjim od nule su hidrofilne.

- Hidrofobnost predstvaljaju stepen ne-rastvorljivosti, tj. "odbojnosti" prema vodi.
- Hidrofilnost je stepen rastvorljivosti u vodi, tj. sposobnost vezivanja sa molekulima vode.

Ova osobina je važna kod funkcionisanja ćelijskih membrana, povezivanje sa drugim molekulima, itd.

Najčešće korišćene tablice za izračunavanje hidrofobnosti su:

- **Kajt Dulitl** (eng. "Kyte-Doolittle")
- **Hop Vuds** (eng. "Hopp-Woods")
- **Ajzenberg** (eng. "Eisenberg")

Za izračunavanje hidrofobnih i hidrofilnih regiona u proteinu, ovde je korišćena Kajt Dulitl skala. Skala hidropatije prema Kajt – Dulitlu je data sledećom tablicom:

Amino Acid	*Index*	*Amino Acid*	*Index*
R	-4.5	S	-0.8
K	-3.9	T	-0.7
D	-3.5	G	-0.4
Q	-3.5	A	1.8
N	-3.5	M	1.9
E	-3.5	C	2.5
H	-3.2	F	2.8
P	-1.6	L	3.8
Y	-1.3	V	4.2
W	-0.9	I	4.5

Tabela 2. Hidrofobnost / hidrofilnost svih amino kiselina

Hidrofobnost / hidrofilnost peptida se računa kao srednja vrednost hidrofobnosti svake amino kiseline koju sadrži. Izrazito hidrofilni regioni se nalaze na površini i vezuju se za molekule vode, regioni sa niskim koeficijentom hidrofilnosti su uglavnom u unutrašnjim regionima proteina i lako međusobno interaguju. Interakcija sa vodom je neophodna za pravilno savijanje i agregaciju proteina i formiranje membrana. Hidrofobni regioni stvaraju agregate radi smanjenja ukupne površine koja je izložena vodi.

2 Korelacija antigenih regiona i neuređenih delova proteina - opis problema

Mnogi proteinski regioni ili neki celi proteini nemaju definisanu 3D strukturu. Pretpostavlja se da su neuređeni regioni proteina, zbog svoje konformacione fleksibilnosti, maskirane nukleinskim kiselinama i drugim proteinima slabi antigeni, nevidljivi za imunološki sistem. Izrazito neuređeni proteini su veoma osetljivi na dejstvo proteaza, pa bi stoga, epitopi iz neuređenih regiona imali slab afinitet vezivanja za MHC molekule i ne bi bili efikasno prikazani kao T-ćelijski antigeni. Zbog slabog afiniteta za T-ćelijske receptore, T limfociti koji se vezuju za slabe epitope sopstvenih proteina imaju mogućnost da izbegnu imunološko brisanje u toku fetalnog razvoja, što bi bilo od značaja kod formiranja antitumorskih vakcina zasnovanim na koktelima antigenih peptida.

Predviđanje T-ćelijskih epitopa, koje je obrađeno u ovom radu, zasnovano je na analizi primarne strukture proteina i vrši se indirektno, preko određivanja peptida koji se vezuju za molekule MHC klasa I i II. Pretpostavka je da bi metode koje na osnovu sekvence proteina predviđaju strukturu proteina, mogle da pruže odgovore na neke od značajnih imunoloških pitanja kao što su raspodela i učestalost epitopa u različitim strukturalnim (i funkcionalnim) delovima antigena, jačina vezivanja epitopa za molekule MHC klasa I i II i fenomen širenja imunološkog odgovora sa jakih na slabe epitope, koji je od posebanog značaja za autoimuna oboljenja i izazivanje imunološkog odgovora na tumor –pridružene antigene.

Cilj ovog rada je da se poređenjem uređenih / neuređenih regiona proteina i antigenih regiona, dobijenih programima za predviđanje (VSL2, NetMhcPan za antigene regione klase HLA - 1 i NetMhciiPan za antigene regione klase HLA – 2), na materijalu veličine 654 analizirana proteina i sve postojeće ljudske alele (HLA-I 1469 alela i HLA-II 517 alela) :

- Ispita raspodela epitopa u uređenim i neuređenim regionima za sve poznate alele HLA-1 i HLA-2 klase, i svaku od 5 analiziranih funkcionalnih grupa proteina.
- Utvrdi da li isti odnosi važe u svakoj od 5 analiziranih funkcionalnih grupa proteina.
- Utvrdi odnos slabih i jakih epitopa i njihovu zastupljenost u neuređenim / uređenim delovima proteina.
- Tehnikama istraživanja podataka utvrdi ponašanje epitopa za obe klase MHC I i II, prema strukturi proteina, vrsti epitopa, alelima koje prepozanju te epitope, sekvencama amino kiselina koje predstavljaju epitope i hidrofobnoj vrednosti epitopa.
- Utvrdi da li najučestaliji aleli u populaciji prikazuju najveći broj epitopa i da li tvrđenje važi u svim strukturnim regionima proteina.

- Utvrdi da li dve analizirane strukturno i funkcionalno specifične podgrupe proteina (bakterijski proteini i kancer-testis tumor-pridruženi antigeni) imaju neke specifične karakteristike u odnosu na uređenost strukture, vezivanje za HLA-1 i HLA-2 alele i hidrofobnost.
- Utvrdi interval hidrofobnosti za epitope u neuređenim regionima, kao i alele koji su slične (tj. koji najčešće prepoznaju iste epitope tzv. promiskuitetne epitope)

Utvrđivanje korelacije između antigenih regiona i uređenih / neuređenih regiona u proteinu bi dalo značajan doprinos imunologiji.

3 Materijal i metode

Za potrebe skladištenja podataka korišćena je relaciona baza podataka implementirana u sistemu DB2 čija je struktura detaljno objašnjena u poglavlju 4. Proteini su prikupljani iz različitih funkcionalnih grupa i baza. Ukupan broj prikupljenih i analiziranih proteina je 654, preuzetih iz:

a) DISPROT baze (479 proteina): sadržaj ove baze su proteini za koje je eksperimentalno utvrđeno da su neuređeni. Za proteine DisProt baze je utvrđeno 7 različitih funkcionalnih osobina. A prema utvrđenim funkcijama i strukturama proteini su razvrstani u 17 kategorija. DisProt baza je detaljnije opisana u poglavlju 1.4;

b) PDB baza (21 protein): šesnaest proteina preuzetih is PDB baze sa 90% uređenom strukturom i pet proteina sa 90% neuređenom strukturom. Protein Data Bank (PDB) je javno dostupna baza svih poznatih prostornih struktura proteina. Strukture proteina su dobijene eksperimentalnim putem najčešće kristalografijom X – zracima i nuklearno magnetnom rezonantnom spektroskopijom. PDB baza podataka je osnovana 1971. godine u Brookhaven National laboratoriji i na početku je sadržala samo 7 struktura proteina. Danas sadrži preko 50 000 poznatih struktura. Baza je dostupna na adresi: http://www.pdb.org/pdb/home/home.do ;

c) SWISS-PROT (19 proteina). Swiss – Prot je baza podataka sa proteinskim sekvencama osnovana 1986. godine. Osim proteinske sekvence ova baza sadrži informacije o funkciji proteina, njegovoj domenskoj strukturi, post-translatornoj modifikaciji, itd. Iz ove baze je izdvojeno 19 kancer - testis antigenih proteina za koje je poznato da imaju izrazito neuređenu strukturu;

d) GenBank (134 proteina). Iz ove baze su preuzeta 4 proteina iz EBNA grupe, to su Epstein – Bar virusi koji odgovaraju različitim grupama maligniteta koje ovi virusi izazivaju. Za proteine ove grupe postoje eksperimentalni rezultati za antigene i neuređene regione i nekoliko objavljenih radova koji opisuju korelaciju uređenih / neuređenih regiona i antigenih regiona. Iz iste baze je preuzet i 131 bakterijski protein od kojih 81 sa kompletno uređenom strukturom po VSL2 programu i 50 sa kompletno neuređenom strukturom po VSL2 programu.

Navedene baze podataka su među najpoznatijim javno dostupnim bazama podataka sa proteinima i DNA sekvencama. Osnovane su i održavane u bioinformatičkim centrima kao što su

Evropski institut za bioinformatiku (eng. "European Bioinformatics Institute", EBI), Nacionalni centar za biotehnološke informacije (eng. “National Center for Biotechnology Information”, NCBI) i GenomeNet. Na internet strani NCBI centra se nalaze baze podataka sa proteinima. Najveća i najvažnija od tih baza podataka je “GenBank” osnovana 1992. godine. “GenBank” baza podataka čuva prikupljene sekvence proteina iz drugih međunarodnih baza podataka (EMBL i DDBJ) i pojedinačnih laboratorija.

Svi analizirani proteini su dužine do hiljadu amino kiselina, duži proteini nisu razmatrani zbog dužine trajanja obrade od strane izabranih programa za predviđanje.

U ovom radu su analizirani svi poznati ljudski aleli kojih ima 1469 za MHC klasu I i 517 alela za MHC klasu II.

3.1 **Priprema i obrada podataka**

Za obradu prikupljenih proteina, propuštanje kroz programe za predviđanje uređenih / neuređenih i antigenih regiona, vizuelni i uporedni prikaz navedenih regiona, automatizovano izvršavanje programa za predviđanje kao i skladištenje dobijenih rezultata u svrhu daljeg istraživanja, je napisana aplikacija nazvana EPDIS. EPDIS aplikacija je napisana u programskom jeziku Java, verzija 6. Podaci se skladište u relacionu bazu podataka InfoSphere Warehouse paketa. Prvobitno je to bila IBM DB2, nekomercijalna verzija Express-C 9.7.

EPDIS aplikacija je detaljno objašnjena u poglavlju 4.

Procenat svih amino kiselina (iz svih prikupljenih proteina) koje pripadaju neuređenim regionima je 49.13%. Broj razmatranih i uskladištenih peptida je preko 400 miliona. O svakom peptidu je sačuvan i podatak o afinitetu vezivanje za obe MHC klase. Na osnovu dobijenog afiniteta se peptid označava kao jak ili slab epitop ili ne-epitop i taj podatak se čuva. Takođe se uz svaki peptid čuva njegova hidrofobna vrednost, protein iz koga je dobijen, detaljan opis proteina i baze iz koje je preuzet kao i funkcionalna grupa kojoj protein pripada. Detaljan opis proteina se odnosi na pun naziv proteina, aminokiselinsku sekvencu kojom je predstavljen, dužinu proteina, intervale koji predstavljaju neuređene regione (eksperimentalne, ako postoje, i dobijene VSL2 programom) kao i sekvencu koja predstavlja neuređeni region, njihovu dužinu, broj i vrstu epitopa koji se nalaze u neuređenim regionima. Ukupan broj prepoznatih epitopa (jakih i slabih), kao i raspodela po uređenim / neuređenim regionima za obe MHC klase je data u tabeli 3. Ukupan broj epitopa (jakih i slabih) je: 1 073 894 za MHC klasu I i 3 608 750 za MHC klasu II.

Broj neuređenih regiona prema VSL2 je 2 446. Dužine pronađenih neuređenih regiona su od 1 do 799, a prosečna dužina neuređenih regiona je 30 amino kiselina.

MHC I	
ukupan broj epitopa	1073894
ukupan broj slabih epitopa	822261
ukupan broj jakih epitopa	251633
neuređeni regioni:	
ukupan broj epitopa	215457
broj slabih epitopa	169510
broj jakih epitopa	45947
uređeni regioni:	
ukupan broj epitopa	792991
broj slabih epitopa	601216
broj jakih epitopa	191775
na prelaznim regionima:	
ukupan broj epitopa	65446
broj slabih epitopa	51535
broj jakih epitopa	13911

MHC II	
ukupan broj epitopa	3608750
ukupan broj slabih epitopa	3409297
ukupan broj jakih epitopa	199453
neuređeni regioni:	
ukupan broj epitopa	678708
broj slabih epitopa	643001
broj jakih epitopa	35707
uređeni regioni:	
ukupan broj epitopa	2403925
broj slabih epitopa	2269419
broj jakih epitopa	134506
na prelaznim regionima:	
ukupan broj epitopa	526117
broj slabih epitopa	496877
broj jakih epitopa	29240

Tabela 3. Broj epitopa po strukturnim regionima za MHC klase I i II

Podaci su smešteni u bazu čija je velika preko 100GB, i koju nije moguće pretražiti u relativno kratkom vremenskom periodu. Kada se i dobije konačni odgovor na upit, obično je to izveštaj na velikom broju stranica i predstavlja selektivno prepisivanje podataka iz baze. Iz tog razloga je primena tehnika istraživanja podataka neophodna za dobijanje rezultata za ciljeve postavljene u prethodnom poglavlju. Izabrana je klaster analiza i tehnika pravila prudruživanja.

Alat koji je u radu korišćen za istraživanje podataka, primenu navedenih tehnika i vizelizaciju rezultata je IBM Intelligent Miner koji je deo paketa InfoSphere Warehouse. Ostali alati poput Veke (skr. od eng. „Waikato Environment for Knowledge Analysis“) nisu bili pogodni zbog vremenske i memorijske zahtevnosti algoritama koje koriste.

3.2 **Istraživanje podataka**

Napredak informacionih tehnologija doveo je do potrebe za obradom velikih količina podataka. Velike baze podataka mogu se danas naći kako u nauci (baze molekularnih podataka, baze medicinskih podataka itd.), tako i u raznim oblastima poslovanja (podaci o korišćenju kreditnih kartica, podaci vezani za poslovanje supermarketa i dr.). Sve veće količine podataka, koje je potrebno čuvati u bazama podataka i obrađivati su davno prevazišle sposobnost ljudskog razumevanja i analiziranja bez korišćenja dodatnih alata. Najveći izazov koji se postavlja je kako pronaći informacije skrivene u velikom broju podataka. Disciplina koja se bavi rešavanjem ovog izazova je poznata pod imenom Istraživanje podataka.

3.3 **Istraživanje podataka i otkrivanje znanja iz baza podataka**

Istraživanje podataka se često definiše kao poslednja faza obrade podataka. Da bi se pojam shvatio na adekvatan način mora da se krene od šireg razmatranja. Porast digitalnih podataka i tehnologija skladištenja je prouzrokovao ogroman porast količine podataka u bazama. Promene su zahvatile sve sfere ljudskog života - od uobičajenih (zapisi korišćenja kreditnih kartica, transakcioni podaci iz supermarketa, detalji iz telefonskih razgovora) do "neobičnih" (baze sa molekularnim, medicinskim podacima, slikama astronomskih tela). Otuda sledi činjenica da je naglo poraslo interesovanje za upravljenjem ovakvim podacima kao i pronalaženje znanja iz istih. Količine podataka toliko brzo rastu da je praktična korist od skladišta podataka ograničena. Javlja se potreba za razvijanjem nove generacije tehnologija i alata za otkrivanje kvalitetnih informacija. Upravo u tom cilju je razvijen koncept KDD – (skraćeno od eng. „Knowledge Discovery in Databases"), čiji je ključni deo upravo istraživanje podataka. Formalna definicija istraživanja podataka-a je:

"*Netrivijalan proces identifikovanja novih, tačnih, potencijalno korisnih i krajnje razumljivih obrazaca u podacima*".

Osnovu za uspeh KDD procesa čini pravilno izgrađeno skladište podataka.

Postoji nekoliko faza procesa otkrivanja znanja iz baza podataka, a to su:

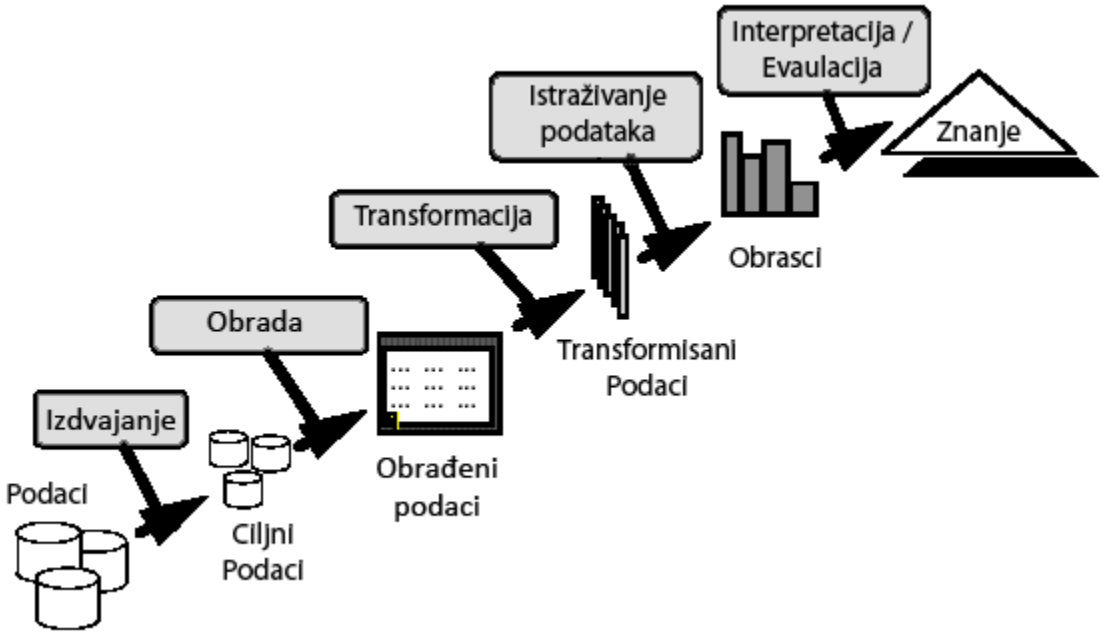
1. Utvrđivanje i analiza ciljeva, oblasti i opsega interesovanja.Analizira se priroda podataka u datom opsegu, a razmatraju se i ciljevi pronalaženja znanja. Ukoliko postoji bilo kakvo prethodno znanje o posmatranoj oblasti i ono se vrednuje.

2. Izdvajanje. U ovoj fazi se izdvajaju samo skupovi podataka nad kojima se traže pravila i obrasci.

3. Obrada i čišćenje. Ova faza podrazumeva pronalaženje ekstremnih vrednosti, obezbeđivanje konzistentnosti, grupisanje, standardizaciju, agregaciju podataka, upravljanje nedostajućim podacima.

4. Transformacija podataka tako da oni budu u skladu sa definisanim ciljevima. Podaci se analiziraju tako da se pronađu korisne karakteristike za prikaz podataka u zavisnosti od cilja istraživanja.
5. Utvrđivanje odgovarajuće tehnike istraživanja podataka. U skladu sa prvim korakom bira se model i parametri.
6. Istraživanje podataka. Algoritam za pronalaženje informacija se primenjuje na prethodno obrađene i transformisane podatke radi pronalaženja traženih pravila i obrazaca.
7. Interpretacija i vizualizacija. Tumače se otkriveni obrasci i bira način njihovog predstavljanja.
8. Eksploatacija znanja i ocenivanje. Dobijeni obrasci se stavljaju u upotrebu. Moguća upotreba uključuje unošenje znanja u druge sisteme radi daljeg istraživanja, dokumentovanje obrazaca i podnošenje izveštaja o njima. To podrazumeva čak i ponovnu upotrebu procesa otkrivanja znanja na istoj bazi podataka, koristeći nova predznanja.

Na slici 10. su prikazani opisani koraci otkrivanja znanja iz baza podataka.



Slika 10. Faze otkrivanja znanja iz baza podataka

Iz prethodno navedenih karakteristika je očigledno da otkrivanje znanja predstavlja multidisciplinarnu oblast i ima ulogu objedinjavanja i upravljanja različitim metodama i tehnologijama. Istraživanje podataka ima centralnu i ključnu ulogu u pronalaženju obrazaca, ali

KDD je taj koji obezbeđuje da nađeno znanje bude stvarno korisno i adekvatno. Bez svih faza otkrivanja znanja, istraživanje podataka uglavnom može doći do netačnih i beznačajnih obrazaca i znanja.

3.4 **Definisanje pojma istraživanja podataka**

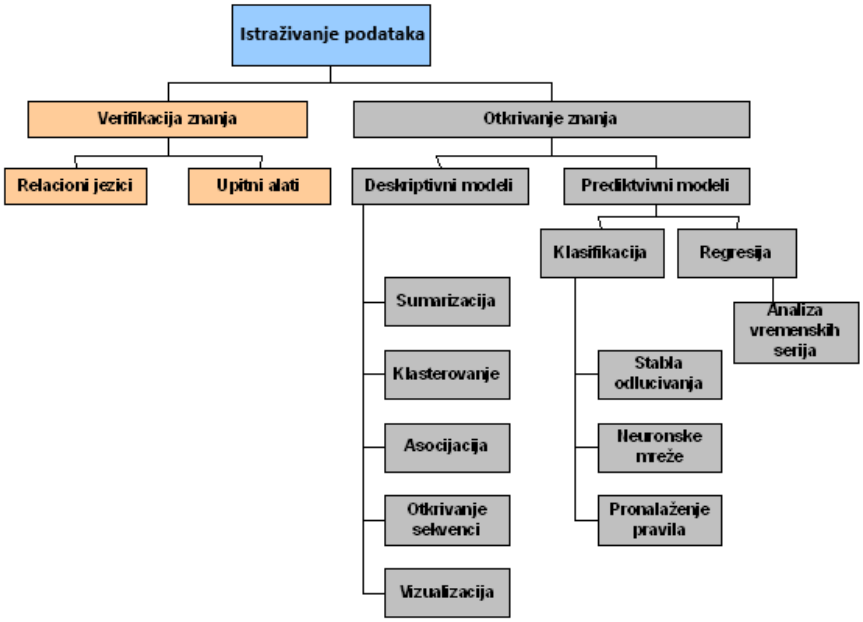
Istraživanje podataka je proces izdvajanja tačnih, prethodno nepoznatih i razumljivih informacija, naizgled nepovezanih u velikim bazama podataka, te njihovog korišćenja za donošenje odluka. Dobijene informacije se mogu iskoristiti za pravljenje modela za predviđanje, za utvrđivanje veza između slogova baze podataka, ili za pregled podataka baze iz koje se podaci izvlače. Relacije i sumiranja dobijena putem analize podataka se definišu kao modeli, ili obrasci. Otkriveni obrasci mogu da budu u vidu: linearnih jednačina, pravila, klastera (grupa), grafova, struktura tipa drveta i rekurentnih obrazaca u vremenskim serijama.

Istraživanje podataka nije pojedinačna tehnika, ili tehnologija, nego skup srodnih metoda i metodologija usmerenih ka pronalaženju i automatskom otkrivanju šablona, sličnosti, promena, anomalija i drugih karakterističnih stuktura iz podataka. Na slici 11 je prikazana jedna od mogućih taksonomija istraživanja podataka. U odnosu na ciljeve koji se postavljaju osnovna podela modela istraživanja podataka je na:

- **Verifikacione** – služe za potvrđivanje hipoteza. Upiti se postavljaju i pristupa se zapisima bitnim za nalaženje odgovora na unapred definisana pitanja. Traže se obrasci, ili informacije koje se mogu u tu svrhu iskoristiti. Prvi korak je formulisanje hipoteze. Zatim se ona odbacuje ili potvrđuje na osnovu rezultrata analize i upita. U prvom slučaju proces se završava, a u drugom se upiti preformulišu i ponovo se izvršavaju nad datim podacima. Očigledno, vrednost dobijenih zaključaka ne proizvodi novu, do tad neotkrivenu vrednost. Zahteva se prethodno znanje onoga ko donosi odluke, a kvalitet dobijene informacije zavisi od načina na koji ga analitičari interpretiraju.

- **Modele za otkrivanje znanja** - Zbog složenosti podataka koji se čuvaju i njihovih međusobnih veza, odlučivanje samo pomoću tehnologija zasnovanih na proveri nije efikasno. Ove tehnologije moraju da se prošire uključivanjem automatskog otkrivanja bitnih informacija, i pravila sakrivenih u podacima i njihovom adekvatnom prezentacijom. Modeli otkrivanja znanja dolaze do rezultata uz veoma malu pomoć korisnika. Međutim, ti modeli nisu rezultat slučajnosti. Naprotiv, alati za istraživanje podataka su dobro osmišljeni i izgrađeni, tako da dozvoljavaju obradu podataka na najjednostavniji i najbrži mogući način.

Na slici 11. prikazana je jedna od više mogućih taksonomija istraživanja podataka.



Slika 11. Taksonomija istraživanja podataka

Dalje, potrebno je praviti razliku između dva pravca otkrivanja znanja:

- **Predviđanja.** U ovom slučaju, cilj je da se pronađu korelacije između polja podataka, odnosno koristi se skup poznatih promenljivih da se predvide karakteristike i pravila vezana za druge nepoznate, ili buduće promenljive.
- **Opisivanja.** Pažnja je usmerena prevashodno na istraživanje opisanih podataka. Svrha njihovog istraživanja je da se identifikuju postojeći obrasci, u okviru podataka, koji opisuju same podatke, kako bi se izveli odgovarajući zaključci.

Postoji i podela modela istraživanja podataka na:

- **Nadgledane** ili ciljne koji zahtevaju skupove ciljnih podataka nad kojima uče.
- **Nenadgledane** ili usmerene koji ne zahtevaju podatke koji bi služili za učenje, nemaju unapred određen raspored i grupe, već se od tehnike istraživanja podataka očekuje formulisanje odgovarajućih struktura sa značenjem.

3.5 Zadaci i kategorije istraživanja podataka

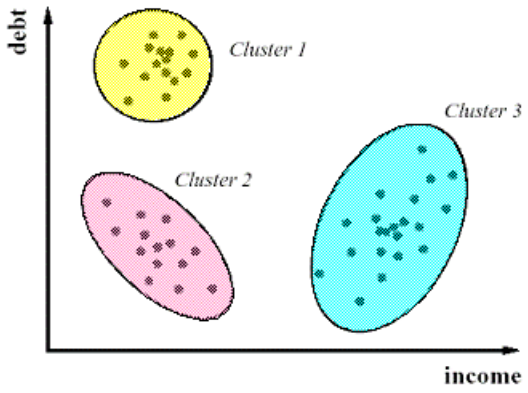
U odnosu na prirodu problema mogu se izdvojiti sledeće tehnike:

- **Klasifikacija***:* Jedna je od najzastupljenijih metoda istraživanja podataka. U tu grupu spadaju metode za svrstavanje entiteta u jednu od nekoliko prethodno definisanih grupa ili klasa. U postupku istraživanja formiraju se klasifikacioni modeli, ispitivanjem prethodno klasifikovanih podatka (slučajeva). Ovo je primer nadgledanog modela, jer zahteva

postojanje skupa podataka u kojem je za svaki ulazni slučaj definisana klasa kojoj pripada. Svaki slučaj sadrži niz atributa, od kojih je jedan specijalan atribut određen za oznaku klase. Suština klasifikacije je pronalaženje modela koji opisuje atribut koji označava klasu kao funkciju ulaznih atributa. Najčešći algoritmi klasifikacije su stabla odlučivanja, neuronske i Bajesove mreže.

- **Klasterovanje (grupisanje).** Ovom metodom se pronalazi prirodno grupisanje slučajeva na osnovu niza atributa, tako da atributi unutar jedne grupe imaju prilično slične vrednosti, a među grupama postoji značajna razlika. Logičke celine, odnosno dobijene grupe se nazivaju klasteri. Za razliku od klasifikacije gde postoje predefinisane klase, ovde to nije slučaj. Pošto ne zahteva skup podataka za treniranje, klasifikacija pripada nenadgledanim metodama istraživanja podataka. Svi ulazni atributi se podjednako tretiraju. Čak se od korisnika ne zahteva ni određivanje ulaznih atributa, niti izlaza, već samo eventualno, broj klastera. Većina algoritama klasterovanja se razvija kroz veći broj iteracija, dok se granice klastera ne stabilizuju. U skladu sa osnovnim definicijama istraživanja podataka, može da se kaže da je suština klasterovanja otkrivanje skrivene vrednosti i promenljivih koje precizno klasifikuju podatke. Metode klasterovanja imaju široku primenu, jer dosta efikasno rade sa različitim tipovima podataka (diskretne, numeričke, kategoričke vrednosti). Često predstavljaju početan korak u istraživanju podataka, koji prethodi klasifikaciji. Često je u upotrebi i naziv segmentacija.

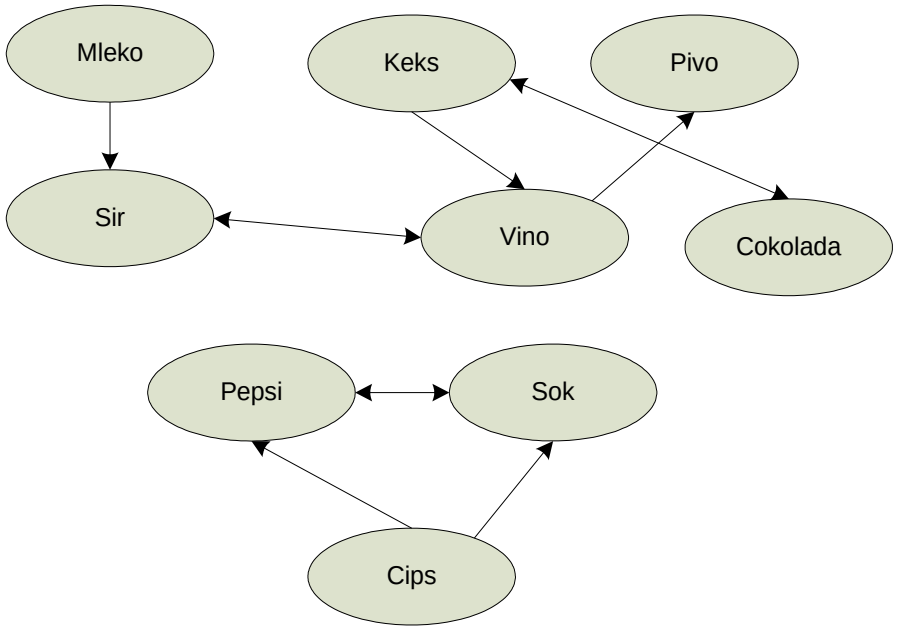
Na slici 12. je prikazan skup podataka koji sadrži dva atributa: „income“ i „debt“. Klaster 1 sadrži stariju populaciju sa niskim primanjima, Klaster 2 obuhvata potrošače srednjih godina i malih prihoda, a Klaster 3 grupiše mlađu populaciju sa nižim prihodom.



Slika 12. Klasterovanje na osnovu dva faktora

- **Pravila pridruživanja** (eng. „asociation rules“). Se opisuje i kao grupisanje po sličnosti. Može se posmatrati kao specijalna vrsta klasterovanja koja identifikuje simultane događaje i transakcije. Najpoznatiji primer pravila pridruživanja je analiza potrošačke korpe. Analiza potrošačke korpe je problem pronalaženja proizvoda koji se prodaju

zajedno. Beskorisno je, zbog velikog broja proizvoda, uzimati u obzir sve moguće kombinacije prodatih proizvoda. Treba izdvojiti samo značajne kombinacije, odnosno česte nizove proizvoda i pravila o povezanosti elemenata kupovine tj. pravila pridruživanja. Ova pravila su u formi A, B => C sa pridruženim verovatnoćama. Trgovački lanci koriste ovu metodu tako da mogu da planiraju raspored i aranžman proizvoda na rafovima, izlozima, katalozima i sajtovima. Na slici 13 je prikazan primer rezultata tehnike pravila pridruživanja na problem potrošačke korpe.



Slika 13. Pravila pridruživanja za problem "Potrošačke korpe"

Tipičan primer pravila na osnovu slike je:

Proizvod = "Pepsi", Proizvod = "Čips" => Proizvod = "Sok".

Interpretacija pravila glasi:

Ako se kupac odluči za Čips i Pepsi, kupac će najverovatnije kupiti i sok.

3.5.1 Istraživanje podataka i skladište podataka

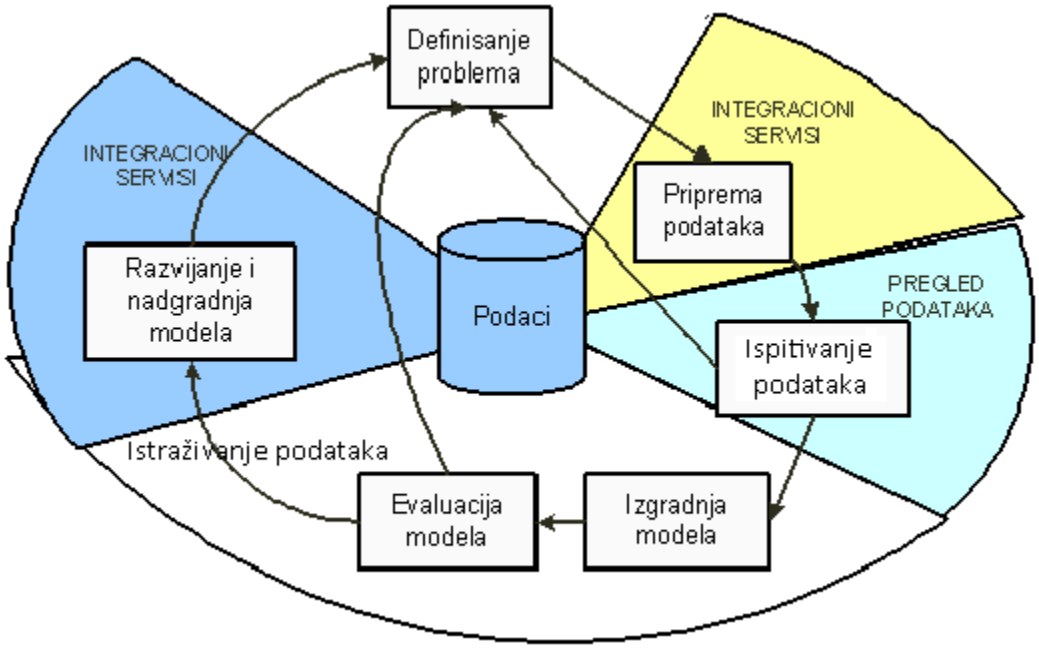
Podaci koji se koriste u procesu istraživanja podataka često potiču iz skladišta podataka. Samo organizovanje podataka za tehnike istraživanja podataka i čuvanje u skladištima podataka je vrlo slično, te u slučaju da su podaci organizovani po modelu skladišta podataka, nema potrebe za dodatnim preuređenjem podataka. Baza podataka za istraživanje podataka predstavlja logički (ne fizički) podskup baze skladišta podataka. Iako sama baza skladišta podataka nije neophodna za tehnike istraživanja podataka, ona ga u mnogome olakšava i potpomaže u ostvarivanju boljih rezultata.

3.6 Metodologija razvoja modela istraživanja podataka

Razvoj modela za istraživanje podataka je samo deo složenog procesa, koji se može definisati preko šest osnovnih koraka:

- Definisanje problema
- Priprema podataka
- Ispitivanje podataka
- Izgradnja modela
- Istraživanje i ocenjivanje modela
- Razvoj i nadgradnja modela

Na slici 14 su prikazani osnovni koraci u razvijanju modela za istraživanje podataka, kao i veze koje postoje između svih koraka.



Slika 14. Koraci u razvoju modela za istraživanje podataka

Treba naglasiti da, iako je na slici proces formiranja i implementacije modela prikazan kao kružni proces, svaki korak ne mora obavezno da vodi ka sledećem. Formiranje modela istraživanja podataka je dinamičan i iterativan proces koji zahteva da se pojedini koraci ponove onoliko puta koliko je potrebno da bi se dobio model odgovarajućeg kvaliteta.

3.6.1 Definisanje problema

Definisanje problema je inicijalna faza, koja se odnosi na razumevanje ciljeva istraživanja. Suština je da se prevedu ciljevi u odgovarajuće probleme istraživanja podataka. U ovom koraku vrši se analiza potreba i definisanje mera na osnovu kojih će se ocenjivati valjanost modela. Ovi zadaci mogu da se prevedu u niz odgovarajućih pitanja kao što su:

- Šta je željeni razultat analize?
- Koji su to atributi čije se vrednosti predviđaju?
- Koje vrste relacija se otkrivaju?
- Da li na osnovu modela treba da se vrši predviđanje ili se samo traže interesantni obrasci i pravila?
- Kako su podaci raspodeljeni?
- Kako su tabele povezane?

Da bi se dali odgovori na ova pitanja, potrebna je procena dostupnosti podataka, kako bi se utvrdilo da li su potrebe korisnika u skladu sa raspoloživim podacima.

3.6.2 Priprema podataka

Podaci koji su dobijeni iz različitih izvora mogu biti u različitim formatima i neretko sadrže nekonzistentnosti, kao što su netačne, ili nedostajuće vrednosti. Nakon formulisanja problema određuje se lista poželjinih podataka. Pri tom se postavljaju sledeća pitanja:

- *Koja količina podataka je dovoljna?* Odgovor zavisi od složenosti podataka, algoritma koji će biti primenjen, učestalosti mogućih izlaza (izlaznih promenljivih). Kada je skup podataka modela dovoljno veliki za izgradnju „dobrog“, stabilnog modela to može biti kontraproduktivno, jer će se vreme obrade povećati imajući u vidu da je proces straživanja podataka iterativan.
- *Koliki je broj promenljivih?* - Neke promenljive su značajnije od drugih. Analiza interpretacije je lakša ako je broj promenljivih manji, odnosno redukovan. Istraživanje podataka je proces kojim se podaci sami razvrstavaju na više i manje značajne. Konačni model se sastoji od samo nekoliko promenljih koje su izvedene kombinovanjem drugih promenljivih.

U vezi sa podacima javljaju se sledeći problemi:

- često se javljaju opisne promenljive sa velikim skupom vrednosti. Rešenje ovog problema je grupisanje u klase koje će sačuvati prvobitnu povezanost sa ciljnom promenljivom.

- numeričke promenljive sa velikim brojem različitih vrednosti ili elementima van granica prave probleme tehnikama koje koriste aritmetičke vrednosti. Problem ima više rešenja: isključivanje elemenata van granica iz analize, deljenje skupa vrednosti na intervale jednake dužine, kao i transformisanje promenljivih redukovanjem opsega tako da se svaka vrednost menja svojim logaritmom.
- javljaju se nedostajuće vrednosti nekog atributa. Neki algoritmi mogu da rade sa nepoznatim vrednostima dok drugi ne mogu.
- javljaju se vrednosti čije se značenje menja vremenom. Pošto se podaci uzimaju iz različitih perioda neretko se dešava da ista vrednost promenljive menja svoje značenje tokom vremena.
- razne nekonzistentnosti u različitim izvorima podataka uzrokovane nejednakim tretiranjem istih pojava.

Očigledno da se prikupljeni podaci moraju transformisati kako bi se prilagodili postavljenom problemu. Pronalaženje ekstremnih vrednosti, dijagnostika nedostajućih vrednosti i predviđanje istih, povezivanje relacionih ključeva iz različitih izvora podataka, postizanje jednoobraznosti (konzistentnosti) u podacima, uzorkovanje, kategorizacija vrednosti atributa, formiranje izvedenih atributa, sažimanje podataka, itd. su samo neke od potrebnih aktivnosti.

3.6.3 Ispitivanje podataka

Veoma je značajno pre formiranja modela dobro istražiti i razumeti podatke. Po završetku prethodno navedenih metodoloških postupaka pripreme podataka, u cilju još detaljnijeg istraživanja može se provesti i analiza relevantnosti atributa. Iako ova analiza nije preduslov za uspešno sprovođenje istraživanja, ali može da doprinese boljem razumevanju odnosa među atributima i izboru optimalne tehnike istraživanja podataka. Postavlja se pitanje da li je odabran pravi skup atributa koji jednoznačno opisuju problem koji treba rešiti, i da li su vrednosti tih atributa pravilno grupisane. Zadatak analize relevantnosti atributa svodi se na otkrivanje onih atributa koji imaju slab ili skoro nikakav uticaj na zadati cilj, što može rezultirati njihovim ne-uvrštavanjem u dalji procese analize. U praksi se obično skup podataka modela deli na tri dela:

- Skup podataka za učenje, koji se koristi za izradu inicijalnog modela,
- Skup podataka za ocenivanje, koji se koristi za proveru tačnosti modela,
- Skup podataka za testiranje, koji se koristi za merenje efikasnosti modela, kada se model primeni na nove podatke.

3.6.4 Izgradnja modela

Na osnovu podataka iz prethodnog koraka može se pristupiti projektovanju i izradi modela. Skup trening podataka se koristi za izradu modela, dok se skup podataka za testiranje

koristi za ocenu tačnosti modela. Nakon definisanja strukture modela, vrši se njegova primena. Reyultat primene je popunjavanje prazne strukture oblicima ponašanja koji opisuju dati model. Ovakav model naziva se ”trening model”. Izbor odgovarajuće tehnike je ključno i veoma kompleksno pitanje, jer zavisi od velikog broja specifičnih faktora, koje može da dovede i do vraćanja na neki od prethodnih koraka.

3.6.5 Ocenivanje i eksploatisanje modela

Posle izgradnje vrši se ispitivanje izrađenih modela i njihove efikasnosti. Ovaj korak je neophodan kako bi se proverilo koliko dobro funkcioniše model, ili ukoliko je izrađeno više različitih modela, koji od njih pokazuje najbolje performanse. Ako se utvrdi da model ne postiže zadovoljavajuće rezultate, potrebno je vratiti se na prethodne korake procesa i izvršiti odgovarajuće korekcije. Neka od pitanja koja se postavljaju su:

- Kolika je tačnost modela?
- Koliko model dobro opisuje i objašnjava posmatrane podatke?
- Sa kojom verovatnoćom i tačnošću model vrši predviđanje?
- Koliko je model razumljiv?

Za testiranje tačnosti i performansi modela se primenjuju različitite mere, kao što su lift koeficijent i klasifikaciona matrica.

3.6.6 Razvijanje i nadgradnja modela

Nakon uspešne izrade modela sledi njihova primena u praksi. Neke od mogućih primena su:

- Korišćenje modela za predviđanja, koja se zatim mogu iskoristiti za donošenje odluka.
- Klasifikacija ulaznih podataka
- Formiranje izveštaja koji omogućuju korisnicima da postavljaju direktne upite nad modelom.

Obično model koristi izvedene promenljive, formirane na osnovu ulaznih originalnih promenljivih. Rezultat je dodatno polje u tabeli podataka, koje može da predstavlja verovatnoću, ili nivo maksimalne verodostojnosti, ili naziv klase, ili klastera sa odgovarajućom verovatnoćom.

Pored navedenog, postoje i druge mogućnosti primene modela istraživanja podataka o čemu je već bilo reči u prethodnim poglavljima ovog rada. Treba, međutim naglasiti važnost veze između projektovanja i primene modela, obzirom da su vrsta modela i način izrade modela u velikoj meri određeni svrhom u koju će model biti upotrebljen.

Praćenje i nadogradnja modela je takođe značajan deo primene istraživanja podataka. Kako se u praksi količina podataka koje model obrađuje, stalno uvećava, neophodno je vršiti stalno praćenje funkcionisanja modela i njegovo prilagođavanje konkretnim uslovima primene.

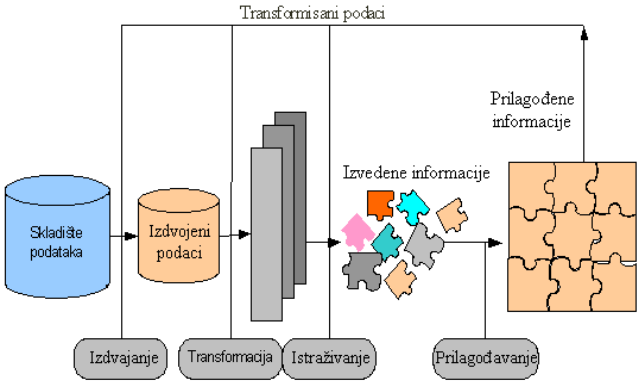
Treba istaći i da je uspostavljena standardna metodologija od strane *CRISP-a* (eng. „Cross Industry Standard Process“), prihvaćena od većeg broja proizvođača alata za istraživanje podataka. Sastoji se iz šest faza:

- Razumevanje poslovanja
- Razumevanje podataka
- Priprema podataka
- Modeliranje
- Provera
- Primena

3.7 Tok istraživanja podataka

Transformacija sadržaja iz skladišta podataka u informacije koje pomažu donošenju odluka je složen process, koji se prema IBM-ovoj metodologiji može organizovati u 4 osnovna koraka:

- izdvajanje,
- transformacija
- istraživanje podataka.
- predstavljanje i ocenjivanje



Slika 15. Tok istraživanja podataka prema IBM-ovoj metodologiji

a) Izdvajanje - skladište podataka sadrži veliku količinu različitih podataka, od kojih neki neće biti potrebni u procesu identifikovanja obrazaca. Shodno tome, vrši se izbor onih baza i podataka koji su adekvatni cilju istraživanja. Na primer, marketing baze podataka sadrže podatke o kupovinama, demografske podatke, podatke o strukturi kupaca i njihovoj kupovnoj moći. Kako bi prodavci u robnim kućama rasporedili proizvode na policama potrebno je da uporede samo demografske podatke i podatke o kupovinama. Nekad je neophodno izvršiti i spajanje među tabelama. Dešava se da u analizu ne mora

biti uključena čitava tabela, već samo pojedini delovi. Isto tako, podaci se u određenim slučajevima sakupljaju iz više izvora.

b) Transformacija - Nakon što su željene tabele odabrane i podaci za istraživanje podataka izabrani, obično je potrebno izvršiti određene transformacije podataka. Tip transformacije koju treba izvršiti, određuje vrsta operacije i tehniku istraživanja podataka koja se koristi:

- transformacija tipa podataka**:** najprostiji oblik transformacije, (npr. iz celobrojne u logičku vrednost), jer se neki algoritmi efikasnije i stabilnije ponašaju sa novodobijenim tipom,
- transformacija kontinualnih atributa (npr. atribut Godine i Prihod se diskretizuju u par grupa),
- grupisanje
- agregacija – koristi se kada su podaci suviše detaljni za zaključivanje, rešenje bi bilo agregirati ih u nove atribute.
- upravljanje nedostajućim vrednostima: dešava se da podaci nedostaju iz više razloga. Postoji veći broj metoda za otklanjnje ovih nedostataka.
- otklanjanje elemenata van granica: abnormalni slučajevi utiču na kvalitet rezultata i kada god je to moguće treba ih odstraniti.

c) Istraživanje podataka - Izbor optimalne tehnike, ili algoritma je suština procesa istraživanja podataka. Preciznost zavisi od prirode podataka, distribucije atributa, veza među atributima, itd.

d) Predstavljanje i ocenjivanje - Informacije dobijene primenom neke od tehnika istraživanja podataka se analiziraju u skladu sa potrebama korisnika. Vrši se izbor najbolje informacije i predstvaljanje preko sistema za podršku odlučivanju. Zadatak ove faze nije samo vizuelizacija (grafička i logička) rezultata, nego i izbor i prilagođavanje odgovarajuće informacije koja će biti predstavljena. Formiraju se optimizovani izveštaji, vrše se prognoze, a rezultati se koriste u različitim aplikacijama.

3.8 **Tehnike istraživanja podataka**

Sa razvojem koncepta istraživanja podataka pojavljuje se širok spektar analitičkih tehnika namenjenih ispunjavanju osnovnih zadataka u procesu otkrivanja znanja u podacima.

a) Stablo odlučivanja (eng. „decision tree“)

b) Pravila pridruživanja (eng. „association rules“)

c) Analiza povezivanja (eng. „link analysis“)

d) Klasterovanje (eng. „clustering“)

e) Konrolisana indukcija (eng. „controlled induction“)

f) Neuronske mreže (eng. „neural networks“)

g) Genetski algoritmi (eng. „genetic algorythms“)

h) Zaključivanje zasnovano na iskustvu (eng. „memory based reasoning“), itd.

Međutim, navedenu podelu je potrebno shvatiti uslovno, jer su opsezi tehnika veliki i međusobno su komplementarne. U nastavku će biti objašnjene tehnike korišćene u radu i tehnika stabla odlučivanja jer je najčešće korišćena tehnika.

3.8.1 Stablo odlučivanja

Jedna od najčešće korišćenih tehnika istraživanja podataka je *tehnika stabla odlučivanja.* Primenjuje se za razvrstavanje, predviđanje, procenu vrednosti, grupisanje, opisivanje podataka i vizualizaciju. Stablo ima svoju grafičku predstavu kao hijerarhijski uređen skup čvorova. Čvor koji je najviši u hijerarhiji se naziva koren (eng. „root“). Ostali čvorovi (eng. „nodes“), koji imaju svoje naslednike, nemaju specijalan naziv, dok se završni čvorovi nazivaju listovi. Svakom čvoru se dodeljuje nivo na kome se nalazi u odnosu na koren, kome se dodeljuje nivo nula. Svaka putanja od korena do lista predstavlja jedno pravilo. Stablo odlučivanja je struktura koja se koristi za rekurzivno deljenje velikih kolekcija objekata na manje skupove, dodeljivanjem niza jednostavnih pravila. Osnovna ideja da svaki podeljeni skup sadrži homogena stanja ciljne promenljive. Prilikom svakog deljenja ocenjuje se uticaj ulaznih faktora. Postoje tri slučaja deljenja polaznog skupa objekata, zavisno od karaktera ulaznih promenljivih:

- deljenje nad numeričkim ulazima,
- deljenje nad deskriptivnim ulazima,
- deljenje u prisustvu nedostajućih vrednosti.

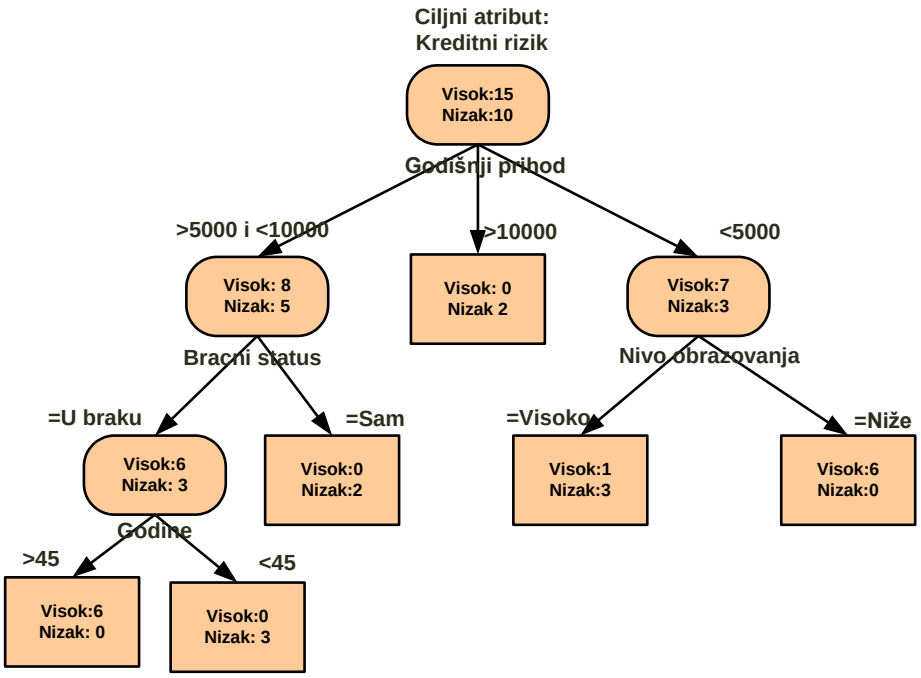
U slučaju numeričkih vrednosti promenljivih razvrstavanje objekata se zasniva na uslovu da vrednost promenljive X bude manja od postavljenog numeričkog praga N, odnosno $X < N$. Osnovni princip podrazumeva da se svi objekti čija je vrednost po kriterijumu X manja od neke konstante N šalju u jedan podčvor, a objekti čija je vrednost $X > N$ ili $X = N$ se šalju u drugi.

Najjednostavniji algoritam za deljenje nad deskriptivnim ulaznim promenljivim je formiranje nove grane za svaku klasu. Na primer, ako se deljenje objekata vrši prema njihovoj boji i skup mogućih vrednosti te promenljive je pet boja {crvena, narandžasta, žuta, zelena, plava}, tada će se formirati pet novih čvorova na nižem nivou stabla.

Češće korišćeni postupak je grupisanje objekata prema sličnim izlazima. Ako se distribucije dve klase ulaznih promenljivih ne razlikuju od distribucije izlaznih promenljivih onda se te dve klase grupišu. Jedan od kriterijuma za razvrstavanje stabala je tip ciljne promenljive prema kome se dele na regresiona stabla (promenljiva je kontinualna) i klasifikaciona stabla (promenljiva ima

diskretan skup vrednosti). Algoritmi korišćeni u procesu formiranja stabala su Hantov algoritam, CART, ID3 – C4.5, SLIQ, SPRINT.

Na slici 16. je prikazan primer upotrebe stabla odlučivanja u oblasti upravljanja rizikom, odnosno određivanja nivoa kredtinog rizika za različite klijente.



Slika 16. Stablo odlučivanja za problem analize kreditnog rizika

Na ovakvom modela se vrlo lako uočavaju pravila, na osnovu kojih se kasnije donose odluke. Ako klijent ima godišnji prihod između 5000 i 10000 dolara, u braku je i mlađi je od 45 god., onda je nivo kreditnog rizika nizak.

Isto tako se mogu formirati i pravila pridruživanja između atributa. Atraktivnost ove metode je u tome što stablo sadrži pravila koja su veoma čitljiva i razumljiva, koja se brzo i lako grade i prevode u poslovna pravila. Nedostaci ove tehnike su, pre svega nestabilnost, takva da mala promena ulaznih podataka pomoću kojih se trenira model, može da dovede do velikih promena topologije stabla.

3.8.2 Pravila pridruživanja

Tehnika pravila pridruživanja pronalazi interesantna pravila i/ili korelacije odnosa između različitih stavki ogromnih skupova podataka. Ova istraživačka tehnika je široko primenjena u mnogim sferama poslovne prakse i istraživanja – od analize potrošačkih navika,

preko upravljanja ljudskim resursima, do razvoja jezika. Omogućava otkrivanje skrivenih obrazaca u velikim skupovima podataka, kao što su na primer, otkrića da ”klijent koji naruči proizvod *A* često naruči i proizvod *B,* ili *C*” ili na primer ”klijenti koji imaju pozitivno mišljenje o usluzi X često se žale zbog problema *Y,* ali su srećni zbog koristi Z.”

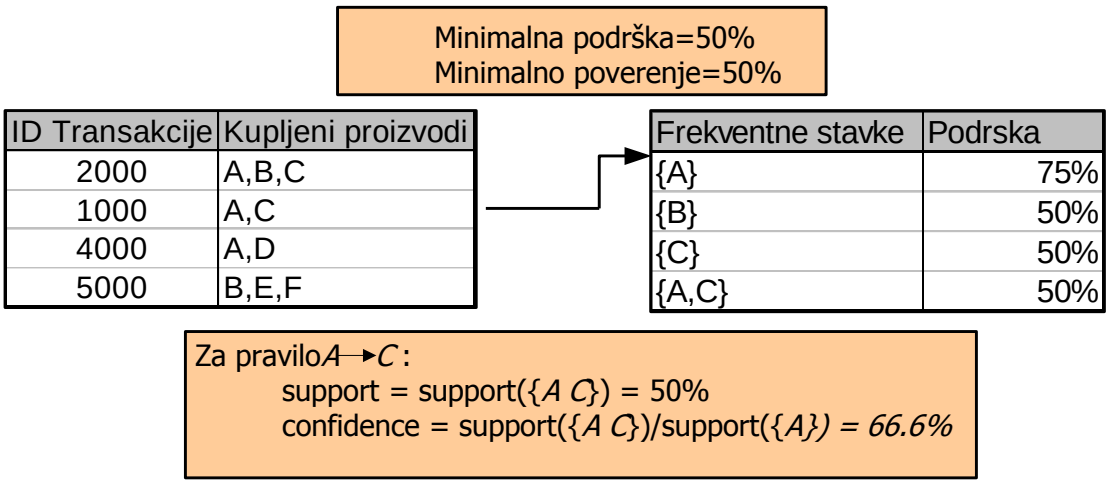
Tipičan i široko-rasprostranjen primer korišćenja pravila pridruživanja je *analiza potrošačke korpe*. Svi proizvodi koje kupac naruči ili kupi tokom određene aktivnosti predstavljaju jedan zapis (slog), odnosno čine jednu transakciju (eng. „itemset“). Svaki element transakcije ima određenu vrednost atributa. U procesu pronalaženja pravila pridruživanja postoje dve faze: pronalaženje čestih skupova i generisanje pravila pridruživanja na osnovu dobijenih rezultata. Mere koje se koriste su podrška (eng. „support“) i poverenje (eng. „confidence“).

Podrška odslikava učestalost sa kojom se skup određenih elemenata (transakcije) pojavljuje u skupu podataka. Računa se kao procenat transakcija (slogova) koji sadrže dati skup artikala (kao podskup) od ukupnog broja transakcija. Ako skup artikala ima podršku veću od specificiranog praga (eng. „minsup“), kažemo da je on podržan (eng. „supported“) ili čest (eng. „frequent“) ili veliki (eng. „large“).

Poverenje odslikava ”implikativnost” (uzročnost, povezanost) koje je prisutno u pravilu, odnosno uslovna verovatnoća da su artikli na desnoj strani pravila B prisutni ako su prisutni artikli na levoj strani pravila A:

$$\text{confidence}(A \rightarrow B) = \text{support}(A, B)/\text{support}(A)$$

Dakle, “poverenje” da je i B prisutno u podacima ukoliko je prisutno A jednako je odnosu nivoa podrške artikala A i B i nivoa podrške artikla A.



Slika 17. Primer određivanja nivoa podrške i poverenja u pravilima pridruživanja

Neki od tipičnih problema gde se pravila pridruživanja primenjuju su:

- *Analiza odlazaka (gubitka) klijenata* (eng. „Churn analysis“). Osnovno pitanje koje se postavlja je pronaći kupce sa najvećom verovatnoćom prelaska kod konkurencije. Utvrđivanje glavnih i skrivenih razlika dovodi do poboljšanja pruženih usluga i zadržavanja klijenata.

- *Ukrštena, dodatna prodaja* (eng. „Cross-selling“). Mnoge kompanije koje pružaju mogućnost online kupovine, obavljanja bankovnih transakcija pomoću platnih kartica, na osnovu utvrđenih pravila preporučuju klijentima sledeće aktivnosti, odnosno proizvode.

- *Otkrivanje prevara* (eng. „Fraud detection“). Kompanije dobijaju na hiljade zahteva za odobravanje kredita, osiguranja, itd. Nije lako utvrditi stepen rizika i njegovu zavisnost od velikog broja parametara.

- *Upravljanje marketinškim aktivnostima* Politika cena (npr. ne nuditi popust na one proizvode koji se ionako kupuju zajedno), politika ponude i promocija, dizajn kataloga, raspored proizvoda u prodavnici, planiranje i optimizacija asortimana proizvoda.*Otkrivanje nepoznatih lidera prodaje* (eng. „loss-leader analysis“)

Utvrditi proizvode i usluge (na kojima se inače ne zarađuje mnogo) koje posredno navode klijente na one na kojima se dosta zarađuje..Na kraju treba istaći da je od velikog značaja izabrati bitna i korisna pravila od mnoštva generisanih, a neodgovarajuća i trivijalna izbaciti iz dalje analize.

U ovom radu tehnika pravila pridruživanja je primenjena za grupisanje „sličnih“ alela u različitim strukturama proteina. To su aleli koji u najvećem broju slučajeva prepoznaju iste epitope. Zatim za pronalaženje epitopa koji se najčešće javljaju zajedno u uređenim i neuređenim strukturama proteina, kao i utvrđivanje intervala hidrofobnosti za epitope u neuređenim regionima.

3.8.3 Neuronske mreže

Oblast neuronskih mreža, odnosno „veštačkih neuronskih mreža”, je vrlo složena i široka.

Stvarne neuronske mreže su biološki sistemi koji imaju sposobnost da otkrivaju šablone, uče i predviđaju. Veštačke neuronske mreže su računarske tehnike koje implementiraju mašinske algoritme učenja i sofisticirano otkrivaju obrasce, u cilju izgradnje modela za predviđanje. Kao što je ljudski mozak sposoban da posle učenja izvlači pretpostavke na osnovu ranijih opažanja, tako su i neuronske mreže sposobne da u projektovanoj sferi predvide promene i dešavanja u sistemu.

Još 40-ih godina prošlog veka se javila ideja da bi jednostavne jedinice za obradu (isto kao pojedinačni neuroni u ljudskom mozgu) mogle da se povežu u velike mreže i tako formiraju sistem koji bi bio u stanju da rešava teške probleme i otkriva ponašanja, koja ne mogu da se otkriju na drugi način. Čini se da su pronalazak algoritma za propagaciju unatrag i povećanje moći računarskih procesora najviše doprineli uspešnoj realizaciji ove ideje. Suština neuronskih mreža je u paralelnoj obradi, što se razlikuje od drugih računarskih programa koji izvršavaju komande sekvencijalno. Proces učenja se obavlja tako što se mreža balansira na osnovu odnosa koji postoje između elemenata u primerima. Na osnovu važnosti uzroka i posledica između određenih podataka formiraju se jače, ili slabije veze između ”neurona”. Tako formirana mreža spremna je za rad na nepoznatim podacima i reagovaće na osnovu prethodno naučenog.

Primena neuronskih mreža, odnosno nelinearnih modela predviđanja je važna, jer omogućuje modeliranje velikih i složenih problema koji mogu da sadrže na stotine promenljivih, sa mnogo interakcija. Zbog sposobnosti da otkriju skrivene veze, nepoznate obrasce i inteligentno generišu izlaze u zavisnosti od ulaza, neuronske mreže se primenjuju u tehnikama istraživanja podataka, pre svega u klasterovanju, klasifikaciji i predviđanju.

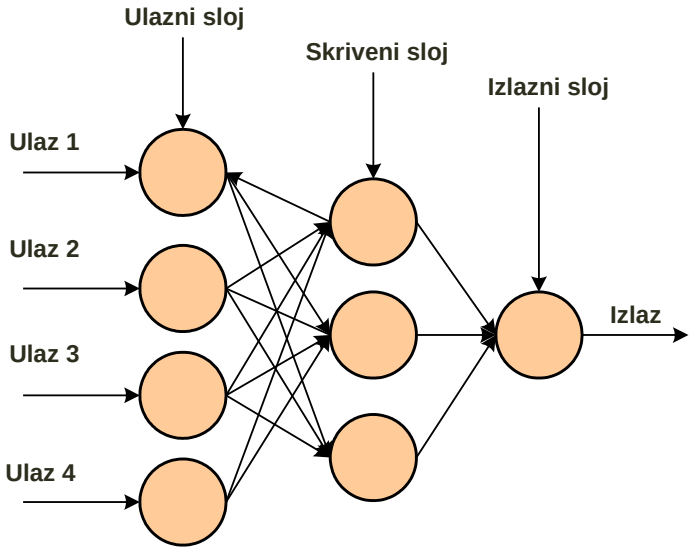
Svaki čvor u mreži predstavlja jedinicu obrade. Između ”neurona” postoje veze sa određenom težinom, analogno sinapsama. Smer veze određuje protok podataka tokom procesa predviđanja. Postoje tri vrste čvorova:

- ulazni (eng. „input“),
- kriveni (eng. „hidden“) i
- izlazni (eng. „output“).

Ulazni čvorovi formiraju početni sloj. U većini mreža su povezani sa jednim od ulaznih atributa iz tabele podataka, koji se normalizuju (najčešće između –1 i 1) i pretvaraju u odgovarajući izlaz.

Jedinice skrivenog sloja računaju izlaze tako što se vrednost svakog ulaza množi odgovarajućim koeficijentom, posle čega se vrši sabiranje i prosleđuje do sledećeg sloja. Neuronska mreža može da ima i više skrivenih slojeva. Širenjem skrivenog sloja se povećava kapacitet mreže za prepoznavanje obrazaca, ali je potrebno voditi računa da se ne formira prevelik broj divergentnih zaključaka.

Izlazni sloj prikazuje vrednost predviđajućeg atributa (obično između 0 i 1). Na slici 18. je prikazan primer topologije neurpnske mreže.



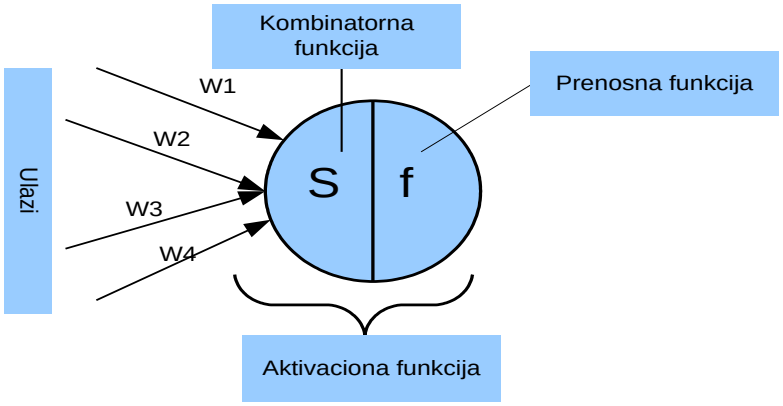
Slika 18. Topologija neuronske mreže

Postoje različite varijacije topologije mreža. Ponekad je ulazni sloj direktno povezan sa izlaznim slojem. U tom slučaju direktna veza se ponaša kao standardna regresija (linearna ili logistička, zavisno od aktivacione funkcije izlaznog sloja). Osnovne jedinice mreže prevode skup ulaznih vrednosti u jednu koju zatim transformišu u izlaz. Takva transformacija je opisana aktivacionom funkcijom. Aktivacione funkcije su najčešće zasnovane na biološkom modelu, čiji izlaz uzima veoma male vrednosti dok kombinacija ulaza ne dostigne prag nadražaja. Kada se dostigne prag, jedinice se aktiviraju i izlaz se povećava. Male promene ulaza mogu da dovedu do velikih promena izlaznih vrednosti. Takođe je moguće da velike promene ulaza uslove male promene izlaza. Takvo ponašanje se naziva nelinearno. Aktivacione funkcije se sastoje iz dva dela:

- kombinaciona funkcija
- prenosna funkcija

Kombinaciona funkcija preslikava sve ulaze u jednu vrednost i svaki ulaz ima sopstvenu težinu. Najčešći oblik je ponderisana suma, gde se svaki ulaz množi težinskim faktorom, a zatim se svi sabiraju. Postoji velika fleksibilnost pri izboru kombinacione funkcije. Ona nije jedinstvena, ali se ponderisana suma pokazala kao veoma dobar oblik u praksi.

Prenosna funkcija je dobila naziv po prenosu vrednosti kombinacione funkcije do izlazne jedinice. Najčešći oblici prenosne funkcije su sigmoidna (logistička), linearna i hiperbolički tangens.



Slika 19. Aktivaciona funkcija u čvoru

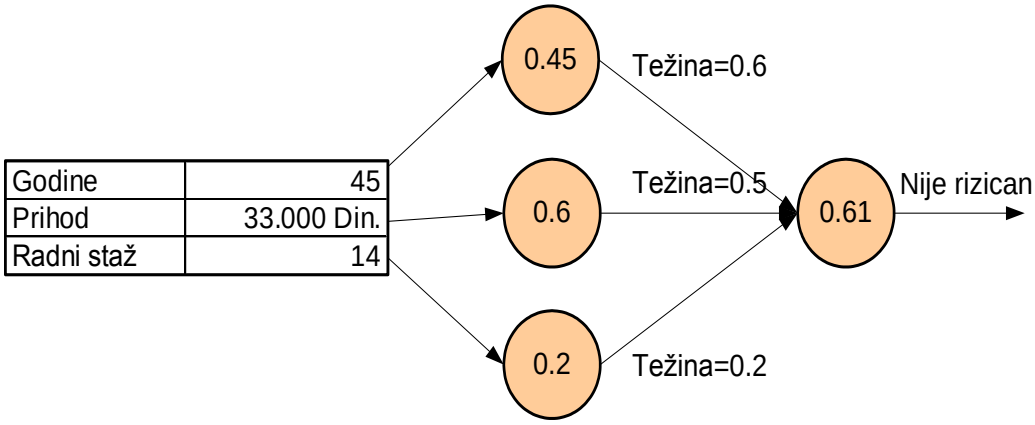
Neuronske mreže su dovoljno dobre samo ako su modeli pravljeni na osnovu trenirajućeg skupa podataka. Treniranje neuronske mreže je proces korigovanja i izbora najboljih težina grana koje povezuju sve jedinice mreže. Cilj se realizuje korišćenjem trenirajućeg skupa za računanje težina, pri čemu se teži da izlaz mreže bude blizak željenom izlazu za većinu objekata trenirajućeg skupa.

Mreža uzima trenirajući skup, koristi postojeće težine i izračunava izlaze. Kretanjem unazad se računa greška kao razlika između proračunatih i očekivanih vrednosti. Greška je povratna informacija mreže. Izračunavaju se nove težine kojima se minimizuje greška. Težine se polako menjaju, težeći optimalnim vrednostima, tako da se greška smanjuje. Cilj je generalizacija i identifikacija obrazaca ulaza. Ako se pri postupku podešavanja težine ne menjaju značajno i greška ne smanjuje potrebno je zaustaviti postupak poboljšavanja. Mreža je naučila da prepoznaje obrasce. Ova tehnika podešavanja se naziva *uopšteno delta pravilo*. Za njega se vezuju dva parametra:

- momentum (kretanje) i
- koeficijent učenja.

Momentum pokazuje tendenciju promene težina, a drugi parametar kontroliše brzinu promene težina. Najbolji slučaj je kada je koeficijent učenja veliki, a zatim se smanjuje sporo tokom treniranja mreže. Kako mreža ide u pravcu optimalnog rešenja tako se koeficijent učenja smanjuje, odnosno mreža pronalazi optimalne težine. Objavljen je veliki broj različitih modela neuronskih mreža. Svaki od njih ima prednosti i nedostatke, a kao osnovni kriterijum se uzima brzina pronalaženja optimalnog rešenja. Opasnost kod svake trening tehnike je pronalaženje lokalnog optimuma. To se dešava kada mreža daje dobre rezultate na trenirajućem skupu, a podešavanjem težina dolazi do pada performansi mreže.

Nakon utvrđivanja topologije mreže, aktivacionih funkcija i realizacije procesa učenja, mreža može relativno brzo i efikasno da rešava i probleme vezane za velike skupove podataka.



Slika 20. Model neuronskih mreže za prognozu rizika izdavanja kredita

Mreža na prethodnoj slici je trenirana tako da vrednost izlaza 1 znači da će klijent vratit dug, a vrednost 0 znači da klijent vrlo verovatno neće vratiti kredit. Dobijena vrednost od 0,61 je bliža jedinici, pa otuda i zaključak da klijent nije rizičan.

Kohonen neuronske mreže se značajno razlikuju od prethodno objašnjenih neuronskih mreže, kako u načinu treniranja tako i u prepoznavanju obrazaca. Kohonen neuronske mreže ne koriste aktivacionu funkciju niti težine. Nema skrivenih slojeva, samo ulazni i izlazni sloj. Mreža se trenira (uči) u nenadgledanom modu, tj. ne postoji definisan skup izlaznih podataka. Ograničenje Kohonen neuronske mreže je što se može primeniti samo na linearno razdvojivom skupu podataka, gde se ulazni podaci klasifikuju u grupe. Prednosti Kohonen neuronskih mreža je što se jednostavno prave i vrlo brzo treniraju, što je na velikom skupu podataka od prevashodnog interesa.

U radu se koristi klaster analiza Kohonen neuronskim mrežama zasnovanim na samoorganizujućim mapama (eng. „Self Organizing Maps"). Ova tehnika je složenija od K – Sredina i hijerarhijskog klasterovanja, ali i značajnija.

3.9 Alat za istraživanje podataka „Infosphere DataWarehouse"

Prethodno poznat kao „DB2 Warehouse", „InfoSphere Warehouse" je najsveobuhvatniji alat za skladištenje i obradu podataka na tržištu. Obezbeđuje pristup strukturalnim i nestrukturalnim podacima, kao i operacionim i transakcionim podacima. Neke od karakteristika su:

- na jednostavan način omogućava izvršavanje standardnih tehnika istraživanja podataka prostim prevlačenjem (eng. „drag and drop“). Takođe, bogatim prezentacionim komponentama omogućena je vizuelna analiza rezultata.
- sadrži „DB2 Warehouse Design Studio”, okruženje zasnovano na Eklipsu, koje uključuje alate za modeliranje, pravljenje i obrnuti inžinjering šema baze.
- Omogućava analizu nestruktuiranih podataka

Besplatna probna verzija korišćena u ovom radu za otkrivanje znanja iz baze podataka i primenu tehnika istraživanja podataka je dostupna na: http://www.ibm.com/developerworks/downloads/im/infospherewarehouse/learn.html

4 EPDIS – „EPitopes in DISorder”

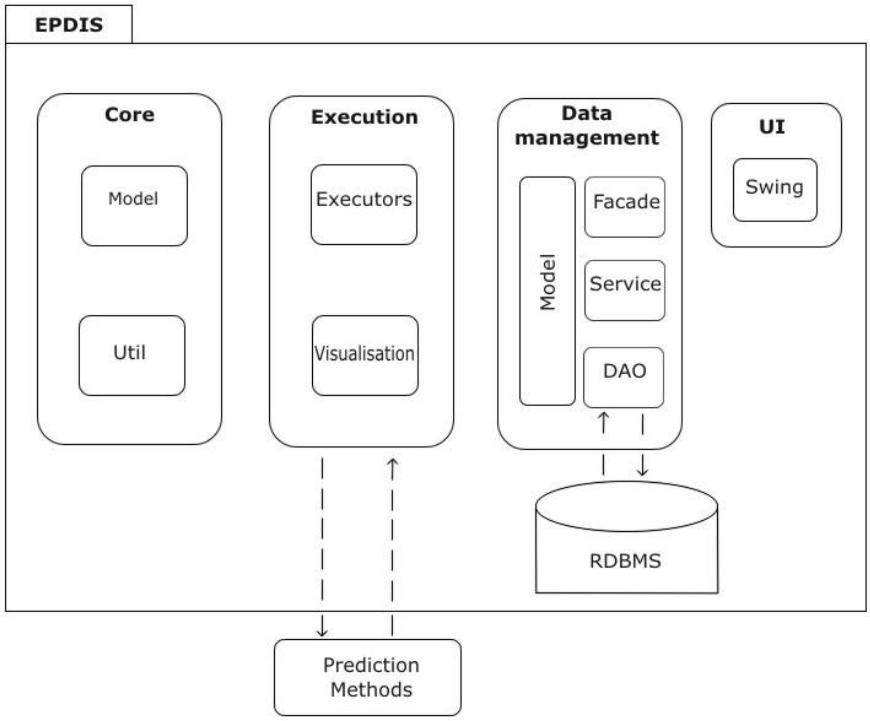
Ključnu ulogu u istraživanju podataka, kako je već objašnjeno u prethodnom poglavlju, igra njihova priprema. Priprema podataka je uključivala izvršavanje programa: NetMhcPan, NetMhcIIPan i VSL2 prediktor, kao i obradu dobijenih rezultata. Programi su detaljno objašnjeni u uvodu. Zatim je trebalo izračunati indeks hidropatije za sve razmatrane peptide u navedenim programima. Programi za predviđanje NetMhcPan i NetMhcIIPan predviđaju antigene regione (epitope) koji se vezuju za molekule klase MHC I predstavljene lokusima HLA 1 i epitope koji se vezuju za molekule MHC klase II predstavljene lokusima HLA 2 (ljudskih alela), respektivno. EPDIS aplikacija je razvijena za pripremu i obradu podataka. Da omogući automatizovano izvršavanje pomenutih programa, izračunavanje i obradu dobijenih rezultata, njihovo čuvanje i vizuelizaciju. Obrađeno je 654 proteina za sve postojeće ljudske alele (ukupno 1986 alela).

4.1 Arhitektura EPDIS aplikacije

EPDIS aplikacija je implementirana u programskom jeziku Java, verziji 6. Sastoji se iz četiri modula (sloja):

- **Core** – čine klase koje predstavljaju model koji simulira i predstavljaju jezgro aplikacije, od kojeg zavise svi ostali moduli. Klase poput Protein, AminoAcid, ClosedInterval, su neke od klasa koje čine pomenuti modul. Takođe ovom sloju pripadaju i pomoćne klase “zadužene” za obradu datoteka u Fasta formatu, itd.
- **Execution** – modul čiji je zadatak izvršavanje eksternih metoda za predviđanje, obrada rezultata i njihovo prevođenje u klase modela radi dalje obrade i vizuelizacije.
- **Data Management** - modul zadužen za obradu rezultata Execution modula, kao i njihovo čuvanje.
- **User Interface** – implementiran u Sving – u . Zbog slojevite implementacije aplikacije lako može biti zamenjen odgovarajućim veb interfejsom.

Arhitektura EPDIS aplikacije je prikazana na slici 21.



Slika 21. Arhitektura EPDIS - a

4.2 **Tehnologije korišćenje u izradi aplikacije**

Sloj za upravljanje podacima (eng. „Data Management“) EPDIS aplikacije sastoji se od tri pod-sloja:

- **DAO** – (skr. od eng. „Data Access Object“) sloja koji se sastoji od klasa za pristup i obradu podataka, u ovom slučaju rezultatima programa za predviđanje. Čine ga interfejsi ProteinDAO i DisorderDAO (kao i generički DAO interfejs koji nije relevatan za sam opis aplikacije), i odgovarajuće implementacije u Hajbernetu . Hajbernet je vodeća ORM (skr. od eng. „Object Relational Mapping“) biblioteka visokih performansi, koja uprošćava objektno-relaciono mapiranje između Java-inih klasa i tabela izabrane baze podataka. Neke od karakteristika Hajberneta su:
    - pruža transparentno čuvanje podataka (klase domenskog modela, ne moraju da naslede nikakvu specificnu klasu niti da implementiraju bilo kakav specifican interfejs da bi se podaci sačuvali).
    - osim mogućnosti zadavanja upita na čistom SQL-u, Hajbernet nudi mogućnost pisanja upita na tzv. HQLu (skr. od eng. „Hibernate Query Language“) koji je nezavisan od tipa relacione baze nad kojom operiše, kao i upite kriterijumima pogodne za formiranje dinamičkih upita,
    - otvorenog je koda, skalabilan i pouzdan, fleksibilan i stabilan.

Glavni razlog izbora Hajberneta u aplikaciji je taj što znatno olakšava pisanje upita nad bazom, a još važniji to što omogućava automatsku proveru sinhronizovanosti klasa koje čine model aplikacije sa tabelema baze. Svaka promena u modelu, dodavanje atributa klase, brisanje ili menjanje, uslovljava automatske promene u bazi prilikom pokretanja aplikacije.

- **Servisni** sloj sadrži funkcionalnost celog modula, i u njemu je sadržana sva logika obrade rezultata programa za predviđanje i njihovo čuvanje. Metode servisnog sloja su transakcione: poziv svake metode servisa će biti ili u celosti izvršena ili ne. Servisni sloj se oslanja na Execution modul za dohvatanje rezultata metoda za predviđanje i DAO sloj za samo manipulisanje podacima, npr. jedna od metoda ProteinService interfejsa

```
void saveProgramForProteinsAndAllele(List<PersistentProtein> proteins,
Allele allele, PredictionMethod method) throws ValidationException;
```

za zadate proteine i alelu izvršava ciljnu metodu za predviđanje određenu trećim argumentom, nad peptidima izdvojenim iz sekvenci proteina i skladišti u bazu. Za kontrolisanje zavisnosti, i opisno upravljanje transakcijama aplikacija koristi **Spring** biblioteku, čije je jezgro Springov IoC (skr. od eng. „Inversion Of Control“) kontejner. Spring smanjuje zavisnost između slojeva aplikacije i na taj način omogućava bolje testiranje aplikacije – svaki deo aplikacije može da se testira nezavisno od ostalih delova. Transakcionalnost servisa moguće je definisati na nekoliko načina:

  - kroz anotacije u samom kodu
  - ili u XML konfiguracionim fajlovima.

- **Facade** dodatni sloj preko servisnog sloja koji je sam po sebi ne-transakcion a omogućava „oslanjanje” na transakcioni servisni sloj, u ovom slučaju zgodan kada se prethodno navedena metoda poziva za višestruke alele. Tada u slučaju da izvršavanje metode za predviđanje ili snimanje rezultata ne prođe kako treba za neku od alela, neće biti poništeno celo izvršavanje metode (pošto metoda fasade nije transakciona), već samo ta neuspešna koja je izvršena pomoću transakcionog servisa. Odgovarajuća poruka o grešci će biti prosleđena korisniku u tom slučaju.

Relaciona baza podataka korišćena za skladištenje podataka u aplikaciji je IBM DB2, nekomercijalna verzija Express-C 9.7.

4.3 Priprema okruženja

Aplikacija koristi konfiguracionu datoteku **disorder.properties** u kojem su navedene apsolutne putanje do instaliranih programa za predviđanje. Zbog lakšeg prilagođavanja lokalnog okruženja prilikom pokretanja aplikacije vrši se pretraga korisničkog direktorijuma u cilju pronalaženja pomenute datoteke. U suprotnom se učitava konfiguracija putanje klase (eng. „class

path“). U slučaju bilo kakve greške i loše konfiguracije korisniku se prikazuje odgovarajuća poruka o grešci.

4.4 Tok pokretanja programa za predviđanje i obrada dobijenih rezultata

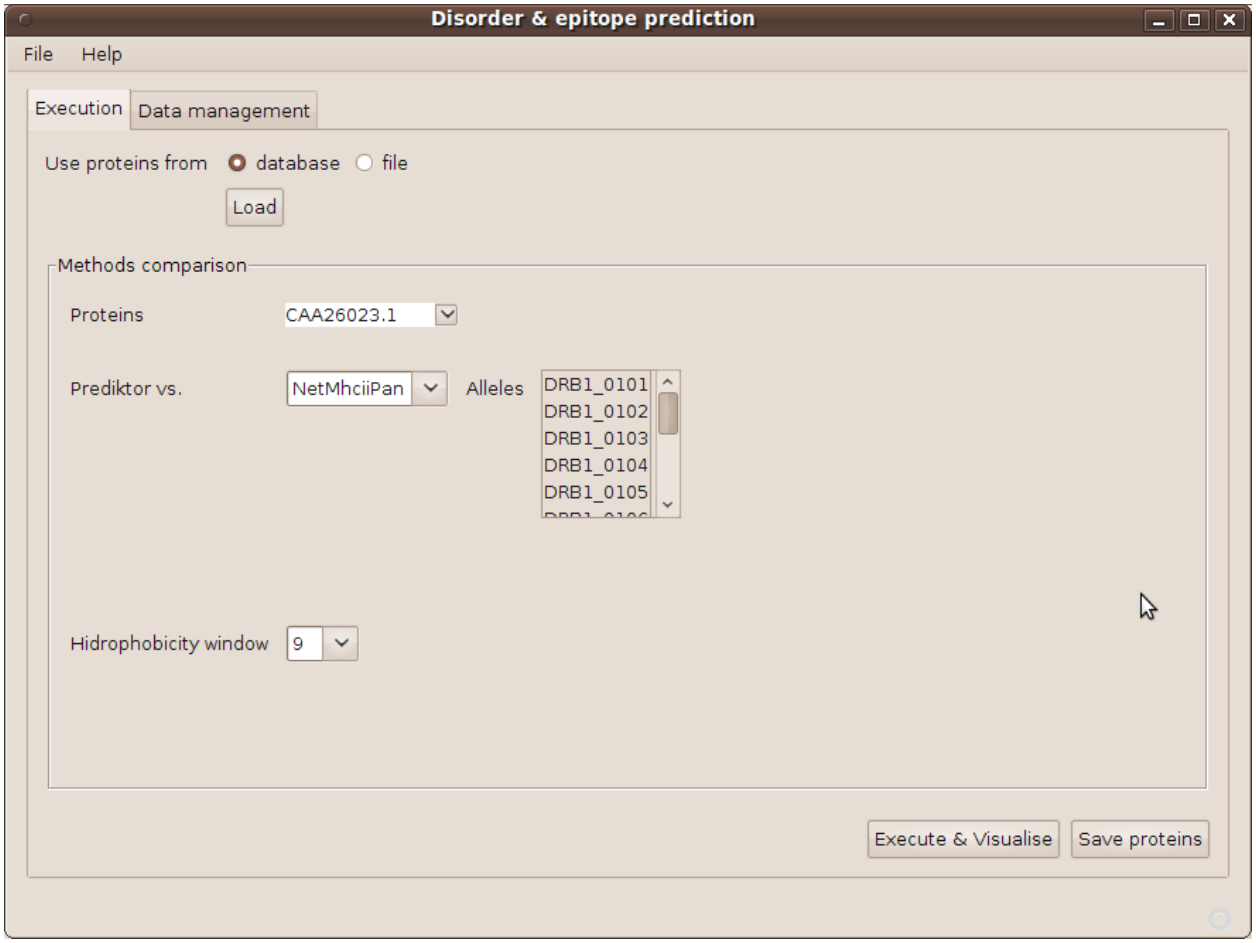
EPDIS aplikacija omogućava učitavanje sekvenci proteina iz tekstualnih datoteka u Fasta formatu, na osnovu kojih se izdvajaju peptidi (u ovom slučaju 9-torke), za koje se vrši predviđanje. FastaUtils klasa sadrži statičke metode za obradu datoteka u Fasta formatu, kao što su: metoda za izdvajanje naziva proteina iz zaglavlja, izdvajanje rednog broja proteina u disprot bazi, itd. Rezultat uspešno obrađenog sadržaja datoteke predstavlja lista objekata klase Protein, koja je ujedno centralna klasa modela. Sadrži odgovarajuću sekvencu aminokiselina, i identifikator koji je u ovom slučaju celo Fasta zaglavlje, iz kojeg se kasnije izdvajaju dodatne informacije. Sledi primer FASTA datoteke koja sadrži jedan protein „MAGE3“ (protein iz grupe kancer-testis antigena).

U zaglavlju datoteke je dat detaljan opis proteina, kao i baza iz koje je protein preuzet. Zaglavlje fasta datoteke se čuva odvojeno u bazi u cilju obezbeđivanja detaljnih informacija o proteinu za tumačenje rezultata. Šifra koja se izdvaja je jedinstvena identifikacija proteina, u ovom slučaju to je „NP_005353.1“. Svakom proteinu u bazi je pridružena jedinstvena šifra. Na taj način je obezbeđena jedinstvenost proteina u bazi jer većina proteina ima veliki broj različitih izomorfnih sekvenci.

```
>gi|4885467|ref|NP_005353.1| melanoma-associated antigen 3 [Homo sapiens]
MPLEQRSQHCKPEEGLEARGEALGLVGAQAPATEEQEAASSSSTLVEVTLGEVPAAESPDPPQSPQGASSLP
TTMNYPLWSQSYEDSSNQEEEGPSTFPDLESEFQAALSRKVAELVHFLLLKYRAREPVTKAEMLGSVVGNW
QYFFPVIFSKASSSLQLVFGIELMEVDPIGHLYIFATCLGLSYDGLLGDNQIMPKAGLLIIVLAIIAREGDCAPEEKI
WEELSVLEVFEGREDSILGDPKKLLTQHFVQENYLEYRQVPGSDPACYEFLWGPRALVETSYVKVLHHMVKIS
GGPHISYPPLHEWVLREGEE
```

Posle izbora proteina, iz čijih sekvenci se izdvajaju peptidi za koje se predviđanje izvršava, potrebno je izabrati neki od programa za predviđanje. Od izabranog programa NetMhcPan ili NetMhciiPan zavisi i izbor raspoloživih alela za koje se predviđanje može izvršiti. U slučaju izbora metode NetMhcPan vrši se predviđanje afiniteta peptida sa kojim se vezuju za molekule klase MHC I, i tada je na raspolaganju 1469 ljudskih alela. Kada se izabere program NetMhciiPan na raspolaganju je 517 ljudskih alela.

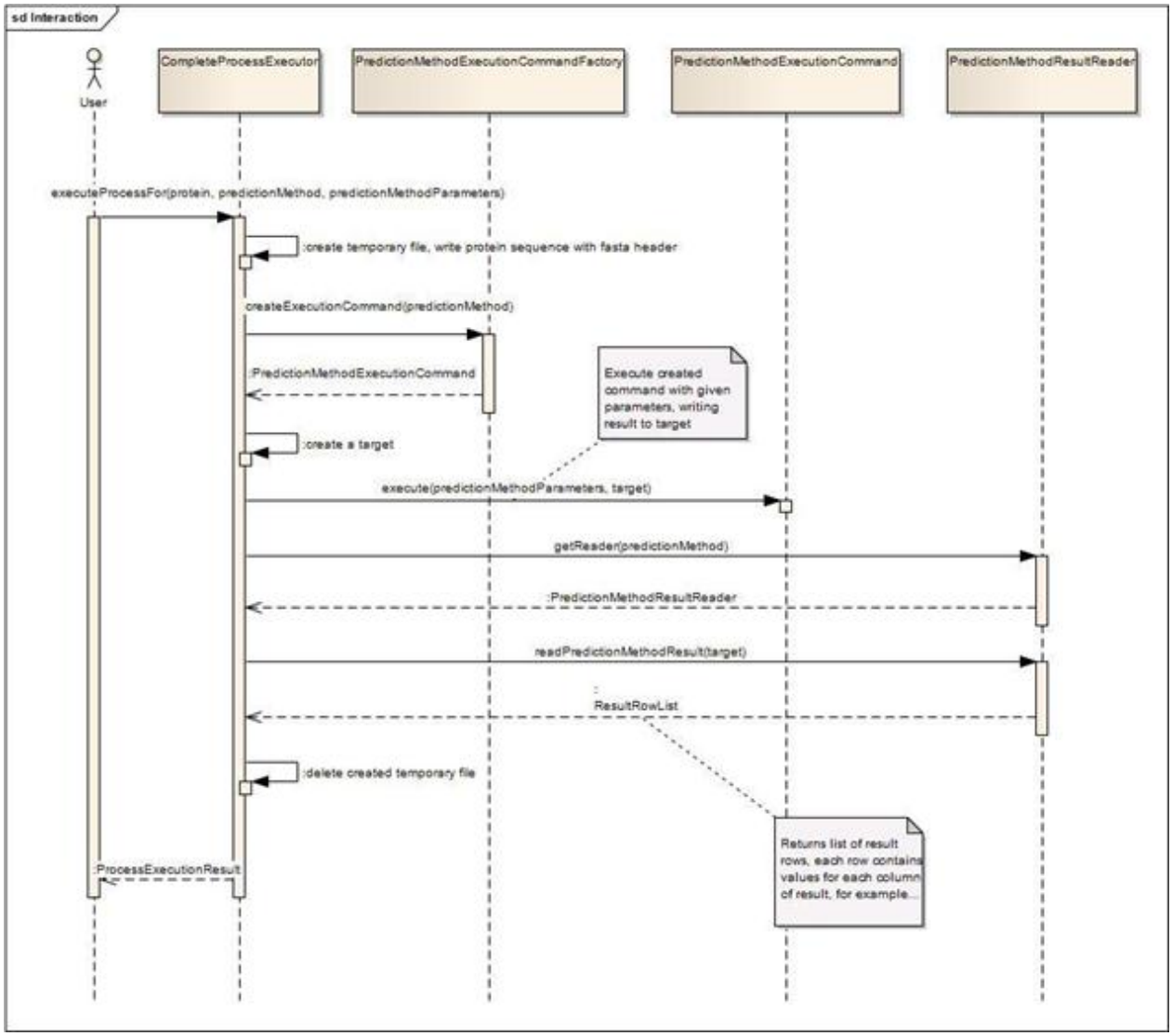
Osnovni interfejs aplikacije prikazan je na slici 22.



Slika 22. Osnovni interfejs aplikacije

Treba napomenitu da prva metoda daje odlične rezultate i za predviđanje antigenih regiona - epitopa drugih vrsta: miša, svinje, majmuna, ali one u ovom radu nisu razmatrane.

Sam tok pokretanja programa za predviđanje i dobijanje rezultata prikazan je sledećim dijagramom toka:



Slika 23. Dijagram toka - EPDIS aplikacije

Korisnik pokreće izabranu metodu (program) za jedan protein sa odgovarajućim parametrima, npr. –a HLA-A0201 za prosleđivanje izabranih alela, -p 9 za dužinu peptida, itd. što je definisano samom metodom.

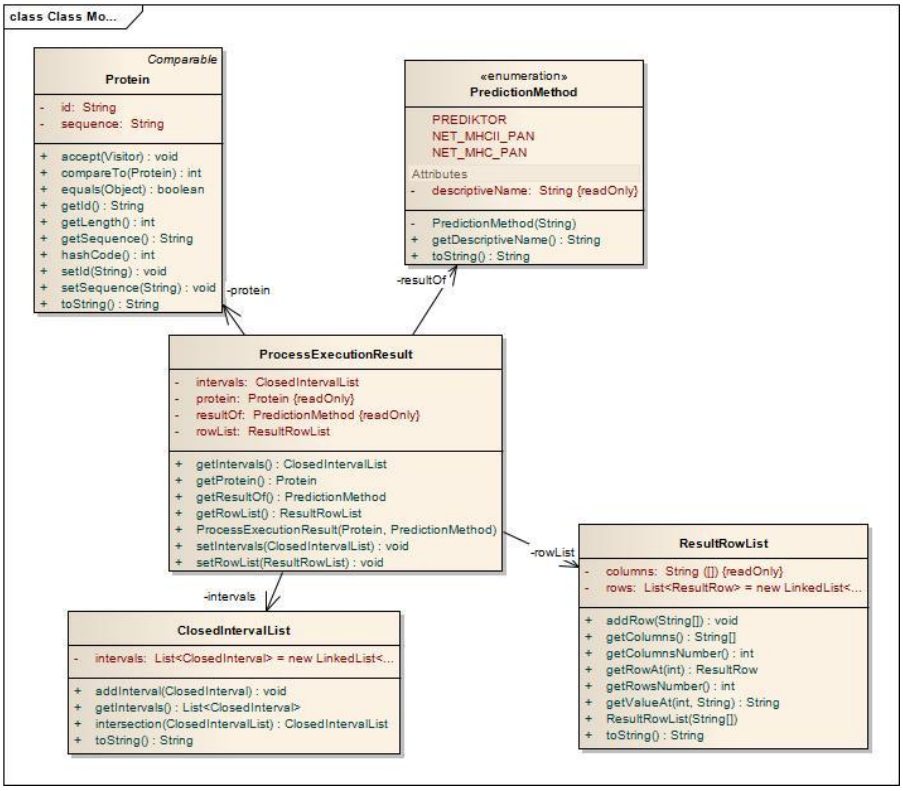
Kako sve izabrane metode predviđanja imaju mogućnost zadavanja parametra koji predstavlja putanju do ulazne datoteke u Fasta formatu, to se prvo formira privremena datoteka u koju se upisuje zaglavlje Fasta datoteke i sekvenca izabranog proteina. Ta datoteka se daje kao parametar programu, i zatim se sama metoda izvršava.

Rezultat programa se zatim obrađuje i iz njega se formira lista rezultata (instanca klase ResultRowList), u kojoj svaki rezultat sadrži odgovarajući izdvojeni peptid, izračunatu vrednost afiniteta vezivanja za molekule neke od klasa MHC I ili II, početnu i krajnju poziciju peptida u sekvenci, itd. Ovakva lista je zatim pogodna za dalju obradu i analizu: iz nje se mogu izdvojiti

zatvoreni intervali u kojima se nalaze antigeni regioni - epitopi, ili neuređeni regionu u slučaju izbora programa VSL2 prediktor.

Takođe je omogućeno preklapanje dobijenih intervala sa intervalima koji se nalaze u samom zaglavlju Fasta datoteke, ako postoje (odnosi se samo na proteine iz DisProt baze). Intervali zadati u samom zaglavlja fasta datoteke predstavljaju eksperimantalno dobijene neuređene regione i nisu poznati za sve razmatrane proteine.

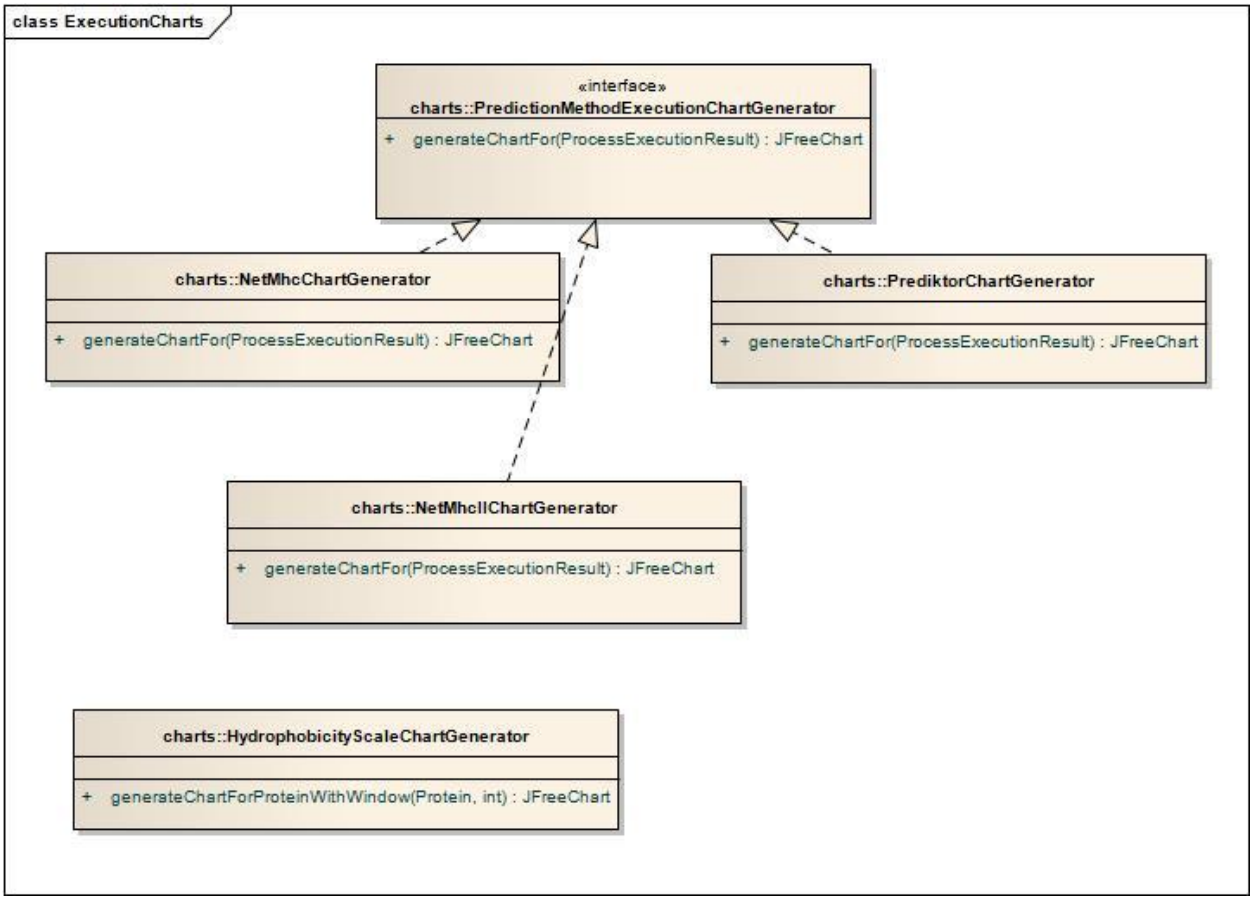
Kada pokrenuti program za predviđanje završi sa radom privremeno formirana datoteka se briše.Rezultat koji generiše metoda za predviđanje i analize dobijenih rezultata predstavlja instancu klase ProcessExecutionResult, koja sadrži informacije o tome za koji protein je određena metoda pokrenuta, rezultat metode, kao i pomenutu listu zatvorenih intervala. Treba napomenuti da su oba programa za predviđanje antigenih regiona vremenski dosta složena. Tako predviđanje za 400 proteina dužine manje od 1000 amino kiselina traje oko 30 minuta za NetMhcPan i 45 minuta za NetMhciiPan za jedan alel. Složenost je dosta veća ako se izabere prozor veličine preko 9 amino kiselina, ali kako se najbolji rezultati dobijaju upravo za tu veličinu to su metode za potrebe ovog rada pokretane isključivo za prozor veličine 9.



Slika 24. Dijagram klasa centralnog modela Execution modula

4.5 **Vizuelizacija**

Poseban deo Execution modula čini deo koji vizuelno prikazuje dobijene rezultate. Za svaki program za predviđanje postoji klasa koja „tumači” rezultate dobijene njenim izvršavanjem (ProcessExecutionResult), i koja zatim iscrtava odgovarajući grafik. U slučaju NetMhcPan i NetMhciiPan prikazuju se rezultati predviđanja za sve peptide dužine 9, gde je osnovna vrednost koja se razmatra 1 – log50k(aff) (normalizovana vrednost afiniteta) sa kojim se vezuje za molekule MHC klasa I i II. Programi rade predviđanje vezivanja i za peptide drugih veličina, kao što je objašnjeno u uvodnom delu, ali ovde su razmatrani samo peptidi veličine 9. U zavisnosti od dobijene mere peptidi se klasifikuju kao ne-epitopi, slabi ili jaki epitopi. U slučaju VSL2 programa prikazuje se koja od amino kiselina iz sekvence proteina pripada neuređenom regionu, a koja uređenom. Takođe je na istom grafiku dodata linija koja pokazuje na osnovu prethodno dobijenih rezultata da li peptid (9 uzastopnih amino kiselina) pripada uređenom odnosno neuređenom regionu, ili se delimično nalazi u uređenom a delimično u neuređenom regionu.



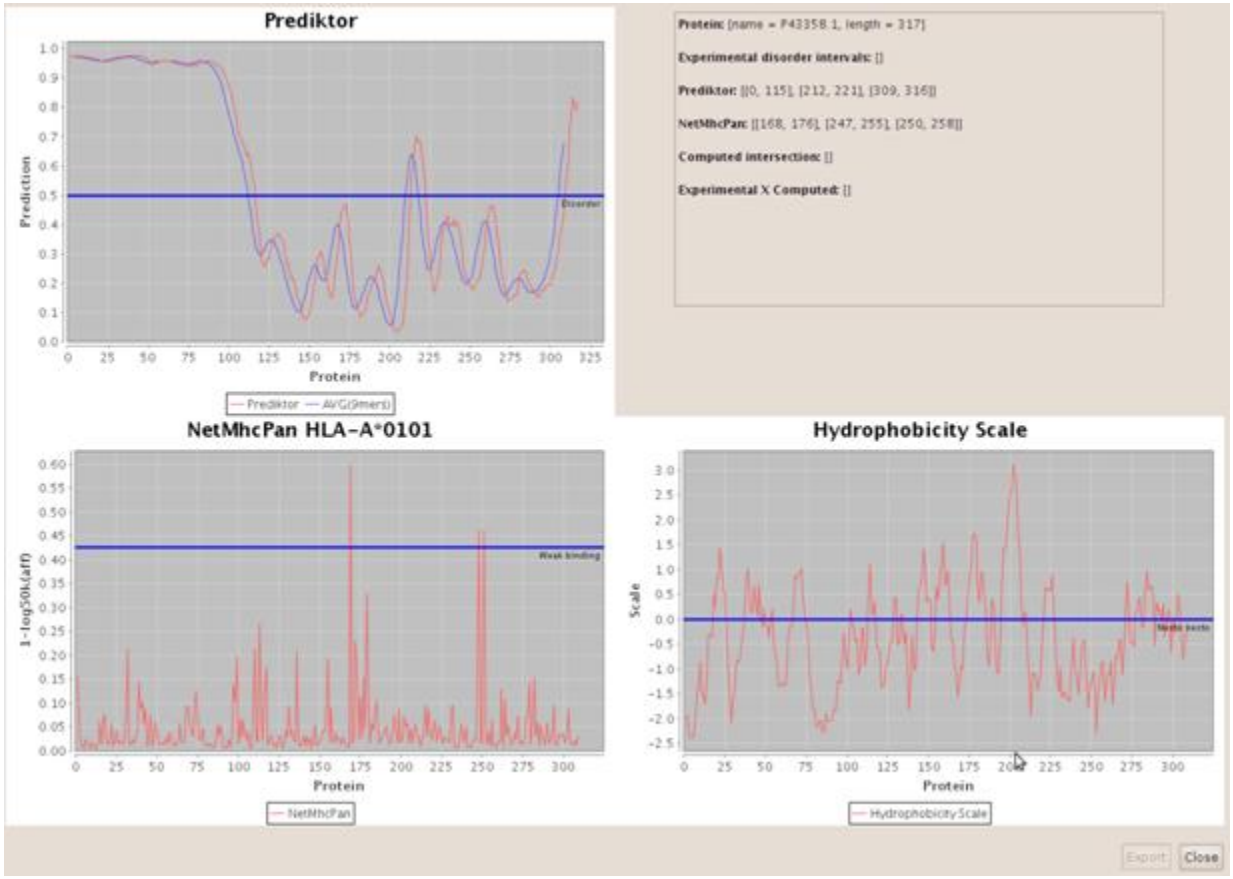
Slika 25. Dijagram klasa generatora grafikona

Crta se i grafik koji predstvalja hidrofobnost svakog peptida u sekvenci, koji se generiše pomoću klase **HydrophobicityScaleChartGenerator**-a. Grafik hidrofobnosti se formira na osnovu Kajt Dulitl skale, objašnjne u uvodu.

Hidrofobnost peptida se računa kao srednja vrednost hidrofobnosti amino kiselina koje čine peptid, i na grafiku svaka tačka predstavlja peptid i vrednost hidrofobnosti za dati peptid.

Osim grafičkog prikaza rezultata svakog od programa za predviđanje koji su integrisani u EPDIS aplikaciji, uz grafik se prikazuju i preseci dobijenih intervala, odnosno preseci neuređenih regiona i regiona koji sadrže antigene regione (epitope), kao i presek ovako dobijenih intervala sa eksperimentalnim rezultatima ako postoje za odgovarajući protein.

Na slici 26. je prikazan primer grafičkog prikaza rezultata aplikacije za jedan protein iz grupe kancer-testis antigenih proteina, za koje ne postoje eksperimantlni rezultati za neuređene regione:



Slika 26. >gi|1170858|sp|P43358.1|MAGA4_HUMAN RecName: Full=Melanoma-associated antigen 4; AltName: Full=MAGE-4 antigen; AltName: Full=MAGE-X2 antigen; AltName: Full=MAGE-41 antigen; AltName: Full=Cancer/testis antigen 1.4; Short=CT1.4

Svi grafici u aplikaciji generisani su uz pomoć javno dostupnog koda, besplatne „JFreeChart“ biblioteke.

4.6 Priprema podataka za istraživanje i njihovo čuvanje

Osim grafičkog prikaza rezultata programa za predviđanje strukture proteina i antigenih regiona, EPDIS aplikacija omogućava skladištenje dobijenih podataka. Već je pomenuto da se pokretanjem aplikacije automatski formiraju tabele u bazi, ukoliko ne postoje.

Struktura baze podataka je sledeća:

Tabela **PROTEIN** sadrži jedinstvenu identifikaciju proteina, sekvencu i dužinu proteina.

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

Tabela 4. Struktura tabele PROTEIN

Tabela **PROTEIN_DETAILS** sadrži detaljnije informacije o proteinu, nazivu proteina koji se nalazi u fasta hederu, i grupi proteina kojoj pripada. Za proteine iz DisProt baze postoji i interna oznaka proteina, koja je značajna jer se prema njoj klasifikuju proteini u različite funkcionalne ili strukturalne kategorije.

Key	Name	Data type	Length	Nullable
🔑	PROTEIN_ID	VARCHAR	20	No
	FASTA_HEADER	VARCHAR	500	No
	ORD_NUM_IN_DISPROT	VARCHAR	20	No
	GRUPA	VARCHAR	20	No

Tabela 5. Struktura tabele PROTEIN_DETAILS

Tabela **ALLELE** sadrži informacije o svim (ljudskim) alelima za koje su pokretani programi za predviđanje antigenih regiona, (analizirane su sve postojeće ljudske alele kojih ima 1469 MHC klase I i 517 MHC klase II).

Key	Name	Data type	Length	Nullable
🔑	CODE	VARCHAR	20	No
	PSEUDO_SEQUENCE	VARCHAR	50	No

Tabela 6. Struktura tabele ALLELE

Tabela **PREDIKTOR** sadrži informacije dobijene pokretanjem programa za predviđanje neuređenih regiona VSL2:

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

Tabela 7. Struktura tabele PREDIKTOR

Tabela **HIDROPHOBIC** sadrži izračunate vrednosti za hidrofobnost svakog peptida koji se može dobiti iz proteina tabele PROTEIN:

Key	Name	Data type	Length	Nullable
	ID	INTEGER	4	No
	PEPTIDE	VARCHAR	9	No
	HYDROPHOBIC_VALUE	DECIMAL	9	Yes
	PROTEIN_ID	VARCHAR	20	No

Tabela 8. Struktura tabele HIDROPHOBIC

Tabela **DISORDER** sadrži informacije o svim neuređenim regionima po proteinu. Ova tabela se koristi za analizu neuređenih regiona, izbor proteina koji sadrže bar jedan neuređeni region željene dužine, zanemarivanje proteina koji imaju samo kratke neuređene regione itd.

Key	Name	Data type	Length	Nullable
	ID	INTEGER	4	No
	SEQUENCE	VARCHAR	9000	No
	START_POS	SMALLINT	2	No
	END_POS	SMALLINT	2	No
	PROTEIN_ID	VARCHAR	20	No

Tabela 9. Struktura tabele DISORDER

Tabela **PROGRAMI** čuva sve rezultate dobijene pokretanjem programa za predviđanje antigenih regiona NetMhcPan i NetMhcIIPan:

Key	Name	Data type	Length	Nullable
	ID	INTEGER	4	No
	"PROGRAM"	VARCHAR	11	No
	PEPTIDE	VARCHAR	9	No
	START_POS	SMALLINT	2	No
	END_POS	SMALLINT	2	No
	POS	SMALLINT	2	No
	PREDICTION	DECIMAL	5	Yes
	AFFINITY	DECIMAL	9	Yes
	BIND_LEVEL	VARCHAR	4	No
	ORDER_LEVEL	CHARACTER	1	Yes
	PROTEIN_ID	VARCHAR	20	No
	ALLELE_CODE	VARCHAR	20	No

Tabela 10. Struktura tabele PROGRAMI

Tabela **EPITOPI_IN_DISORDER** predstavlja vezu između neuređenih regiona i epitopa koji pripadaju odgovarajućim neuređenim regionima. Koristi se u analizi neuređenih regiona prema frekventnosti pojavljivanja epitopa za svaki alel.

Key	Name	Data type	Length	Nullable
	ID	INTEGER	4	No
	PROGRAM_ID	INTEGER	4	No
	DISORDER_ID	INTEGER	4	No

Tabela 11. Struktura tabele EPITOPE_IN_DISORDER

Tabela **PREDIKTOR_PREDICTION** čuva informacije o pripadnosti peptida neuređenom odnosno uređenom regionu.

Key	Name	Data type	Length	Nullable
	ID	INTEGER	4	No
	PEPTIDE	VARCHAR	9	No
	PREDICTION_AVG_VALUE	DECIMAL	9	Yes
	START_POS	SMALLINT	2	No
	PROTEIN_ID	VARCHAR	20	No

Tabela 12. Struktura tabele PREDIKTOR_PREDICTION

5 Rezultati

5.1 Grafički prikazi i analize rezultata

Prvi deo ovog poglavlja sadrži primere grafičkog prikaza rezultata programa za predviđanje za nekoliko proteina iz grupe kancer-testis antigena, dobijenih kao rezultat EPDIS aplikacije. Sa grafika se jasno vidi korelacija uređenih / neuređenih regiona i antigenih regiona (epitopa) u proteinu za obe MHC (HLA) klase alela.

Grafički prikaz je pogodan za analizu i prikaz rezultata za jedan alel, dok su rezultati dobijeni za sve alele, tehnikama istraživanja podataka, prikazani dalje u radu.

Za demonstraciju su izabrani proteini iz kancer-testis grupe tumor-asociranih antigena upravo iz razloga što predstavljaju izrazito neuređene proteine. To su najčešće proteini uključeni u brojne ćelijske regulatorne procese koji zahtevaju adaptibilnost karakterističnu za neuređene regione. Sa druge strane, za kancer vakcine poželjni su epitopi u uređenim strukturama koje prepoznaje što veći broj alela radi imunizacije većeg broja pacijenata. To su tzv. promiskuitetni epitopi. Kako su kancer pridruženi proteini najčešće normalni proteini, samo previše ili pogrešno ispoljeni, imunološki odgovor na njih je eliminisan u prenatalnom razvoju. Izuzetak može biti imunološki odgovor zasnovan na slabim epitopima. Stvaranje imuniteta (tumorske vakcine) na ove proteine uporedivo je sa autoimunitetom, jer se stvara imunološki odgovor na sopstvene proteine. Kod autoimunih proteina je, kao i u odgovoru na neke strane proteine, primećeno "širenje" imunološkog odgovora koji počinje od uređenog epitopa pa ide ka epitopima u neuređenim regionima ili prelazi na različite antigene. Izuzetak mogu, takođe, biti kancer-testis tumor-pridruženi antigeni, proteini koji se nalaze samo na tumorskim ćelijama, dok se u normalnom tkivu nalaze samo u tzv. imunološki-zaštićenim (privilegovanim) zonama u organizmu, kao što su testisi, placenta ili fetalni ovarijum. Imunološki odgovor na ovu grupu proteina nije eliminisan u toku fetalnog razvoja i imunološki odgovor se može javiti i na jake epitope, zbog čega su ovi antigeni dobar cilj za imunoterapiju tumora.

Na slici 27. su prikazani rezultati dobijeni za protein poznat kao MAGE 4. Plava linija na slici (prvi grafik) predstavlja granicu između uređenih i neuređenih regiona, predviđenih programom VSL2.

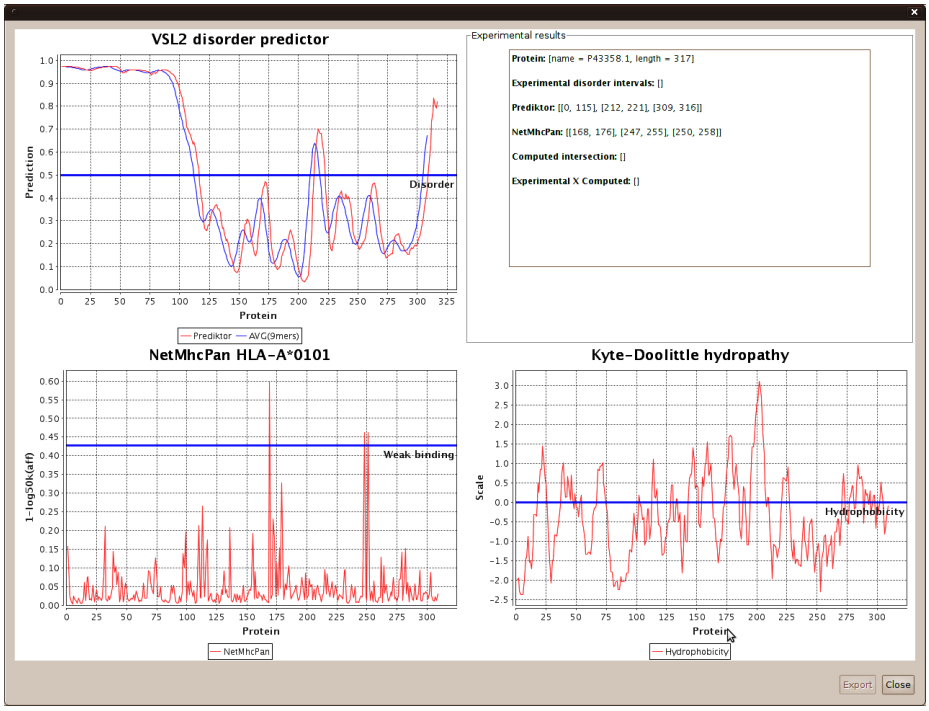
Grafik ispod prikazuje potencijalne epitope, koji se nalaze iznad plave linije (peptide čiji je afinitet vezivanja veći od unapred utvrđene granice, a dobijen je nekim od programa netmhcpan ili netmhciipan). Ako je na grafiku prikazana i zelena linija onda je njome označen prelaz između slabih i jakih epitopa: sve iznad zelene linije predstavlja predviđene jake epitope koji se vezuju za molekule izabranog alela.

Treći grafik na slici predstavlja hidrofobnost peptida u proteinu. Hidrofobnost je izračunata kao srednja vrednost hidrofobnosti svake amino kiseline u peptidu po Kajt-Dulitl skali. Grafik prikazuje hidrofobni (hidrofilni) karakter peptida, koji je koristan za predviđanje širenja membranskog domena, potencijalnih anigeničnih mesta i površina koje su verovatne za prikazivanje na površini proteina.

Kajt-Dulitl skala se najčešće koristi za određivanje hidrofobnog karaktera proteina. Sve vrednosti (amino kiselina ili peptida) iznad nule se tretiraju kao hidrofobne, a ispod nule kao hidrofilne (na grafiku plava linija označava granicu). Veličina prozora (broj uzastopnih amino kiselina) koja je zgodna za pronalaženje hidrofilnih regiona je najčešće između 5 i 7. Ovako dobijeni hidrofilni regioni se verovatno eksponiraju na površini proteina i predstavljaju potencijalne antigene regione. Za pronalaženje hidrofobnih regiona je najbolja veličina prozora između 19 i 21 i tada se uzima kao granica 1.6. U tom slučaju peptidi koji imaju vrednost ove mere ispod 1.6 su hidrofilni, a preko 1.6 se smatraju hidrofobnim. Kako su peptidi, razmatrani kao potencijalni epitopi, (ovde veličine 9), to je i hidrofobnost računata za prozor veličine 9. Primenom metoda istraživanja podataka, koje su opisane u prethodnom poglavlju a čiji su rezultati dati u nastvaku, je utvrđeno da bi granica hidropatije za peptide veličine 9 mogla biti 1.1.

Analizom grafika, prikazanog na slici 27, za protein MAGE4 i alel HLA*A0101 (MHC klase I) se vidi sledeće: nisu prepoznati jaki epitopi, već samo slabi i to mali broj. Svi prepoznati epitopi se nalaze u uređenim regionima. Iako protein ima duže neuređene regione, što se na grafiku označenom kao „VSL2 disorder predictor“ dobro vidi, afinitet vezivanja u tim regionima je jako mali (grafik označen kao NetMhcPan).

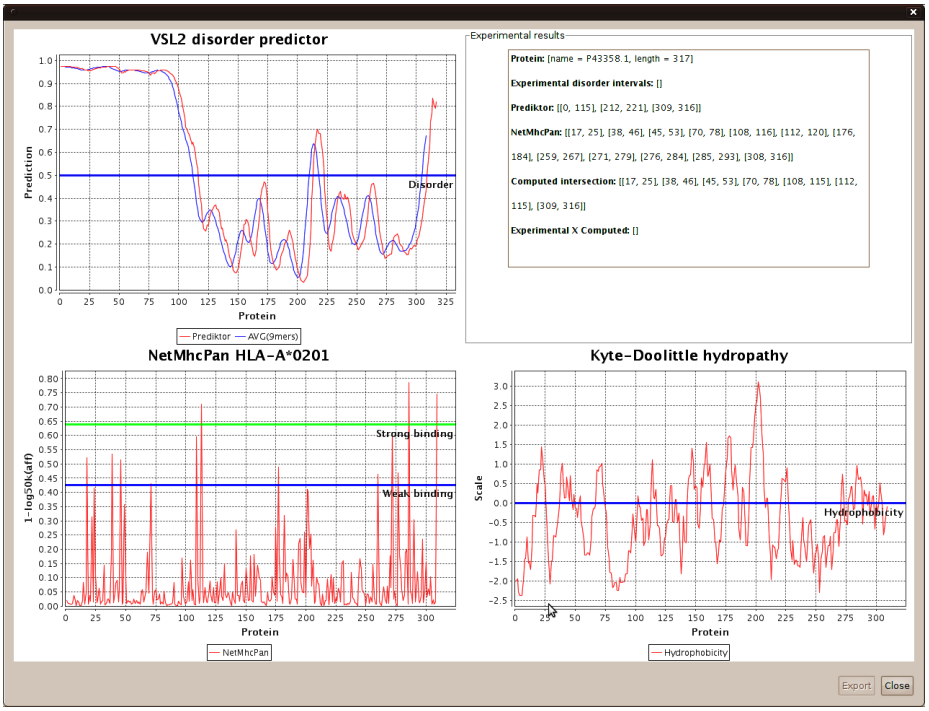
Grafik hidrofobnosti za protein MAGE 4 pokazuje da je uređeni region pretežno hidrofoban. Poznato je da su uređeni regioni uglavnom hidrofobni a neuređeni hidrofilni.



Slika 27. Protein MAGE4 MHC klasa I (alel HLA*A0101)

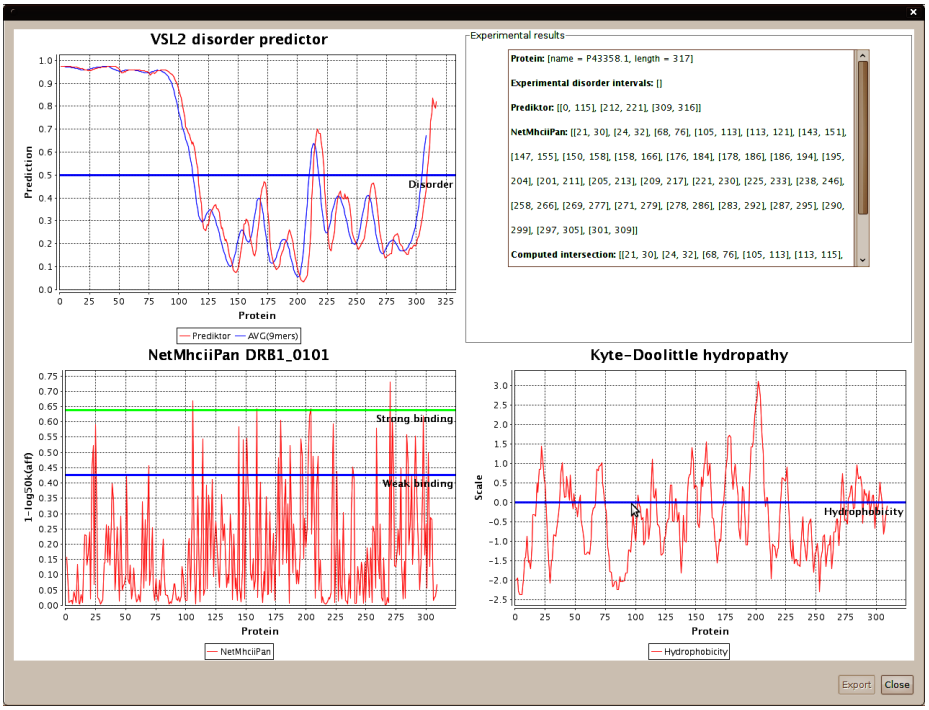
Sledeći grafik (slika 28) prikazuje drugačije ponašanje proteina MAGE 4 kada je u pitanju vezivanje za alel HLA*A0201, za koji je eksperimentalno potvrđeno da "pedstavlja" MAGE4 epitop, koji se već koristi u kliničkim ispitivanjima u terapiji melanoma. U tom slučaju broj prepoznatih epitopa je nešto veći. Među prepoznatim epitopima ima i jakih.

Na ovom primeru se vidi da se epitopi (u ovom slučaju slabi) javljaju i u neuređenim regionima, dok svi prepoznati jaki epitopi pripadaju uređenom delu proteina. Epitopi koji pripadaju neuređenim regionima su označeni na slici u panelu „Experimental results" i polju „Computed intersection". U intervalu početna vrednost predstavlja poziciju u proteinu na kojoj počinje epitop, a krajnja vrednost je pozicija u proteinu do koje epitop pripada neuređenom regionu.



Slika 28. Protein: MAGE4 MHC klasa I (alel HLA*A0202)

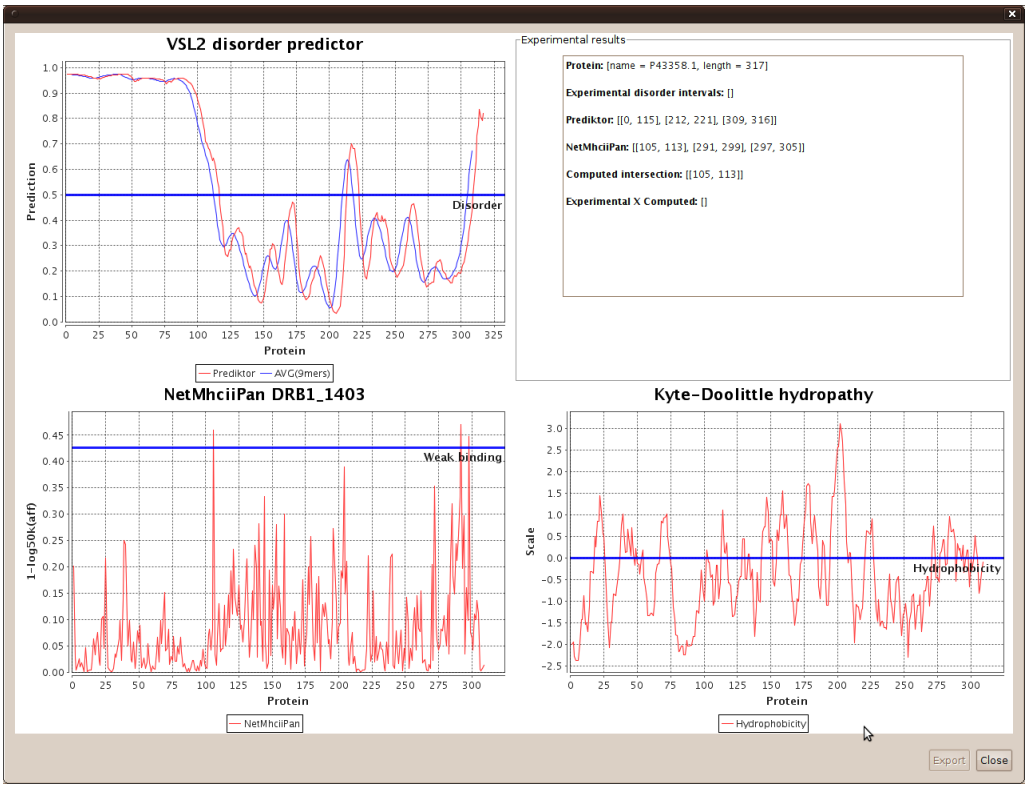
Broj prepoznatih epitopa (predviđenih programom NetMhciiPan) koji se vezuju za molekule MHC klase II je znatno veći. Na slici 29. prikazani su epitopi proteina MAGE 4 za alel DRB1_0101 MHC klase II:



Slika 29. MAGE4 MHC klasa II (alel DRB1*0101)

Prepoznati su i jaki i slabi epitopi. Broj jakih epitopa je mali i svi se nalaze u uređenim regionima. Broj slabih epitopa je veliki, i oni su skoncetrisani u uređenim regionima sa malim brojem izuzetaka (tačno 5) koji se nalaze u neuređenim regionima.

Broj epitopa koji se vezuju za molekule MHC klasa I i II zavisi od izabranog alela. Ovde su prikazani rezultati samo za po dva alela za obe klase. Aleli HLA*A0101 i DRB1*0101 su najčešći ljudski aleli. Na slici 30 je predstavljen rezultat za protein MAGE 4 i alel klase MHC II DRB1*1403:



Slika 30. MAGE4 alel DRB1*1403 MHC II klasa

Za ovaj alel se vezuje manji broj epitopa i to samo slabih. Svi epitopi se nalaze u uređenim regionima. Ono što se još može zaključiti analizom sa grafika je to da je jedan broj epitopa skoncetrisan na prelazima između uređenih i neuređenih regiona.

Korelacija sa eksperimentannim rezultatima za *in vitro* indukovani T4 pomažući imunološki odgovor (nisu prikazani rezultati afiniteta za odgovarajuće MHC II alele) za ovaj antigen pokazuje slaganje, gde se za alel DRB1*0101 dobija najjači imunološki odgovor i to u regionu od 250-280 A.K. [25]. Programom je predviđen veliki broj epitopa, od kojih i jedan jak. Za alel klase MHC II DRB1*1403, u istom radu je eksperimentalno nađeno da vezuje peptid 290-300, što bi na osnovu predviđanja bila 2 slaba epitopa u istom peptidu.

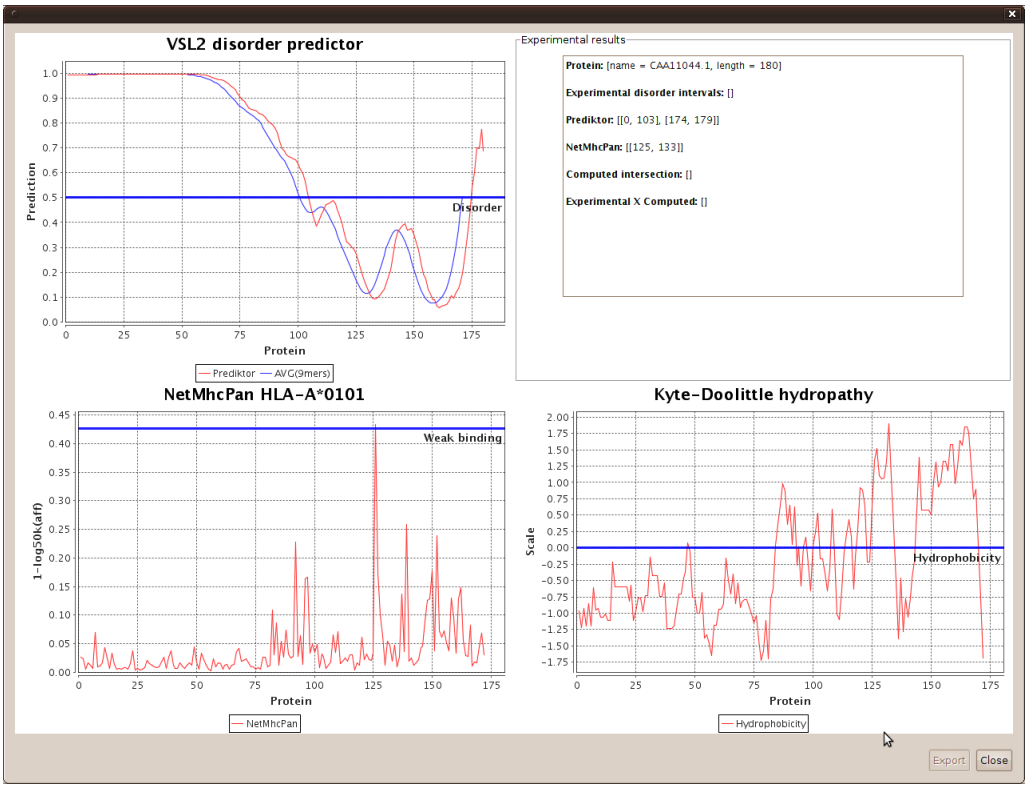
Za ostale kancer-testis antigene prikazani su, u daljem tekstu, prvi aleli svake od klasa MHC I i II, koji i predstavljaju najčešće ljudske alele. Preostala dva alela su izdvojena jer za njih postoje poznati rezultati.

LAGE1 je takođe protein iz grupe kancer-testis antigena. Rezultati dobijeni za protein LAGE1 su prikazani na slici 31.

Za najčešći ljudski alel HLA*A0101 MHC klase I je prepoznat samo jedan slab epitop, koji pripada uređenom delu proteina.

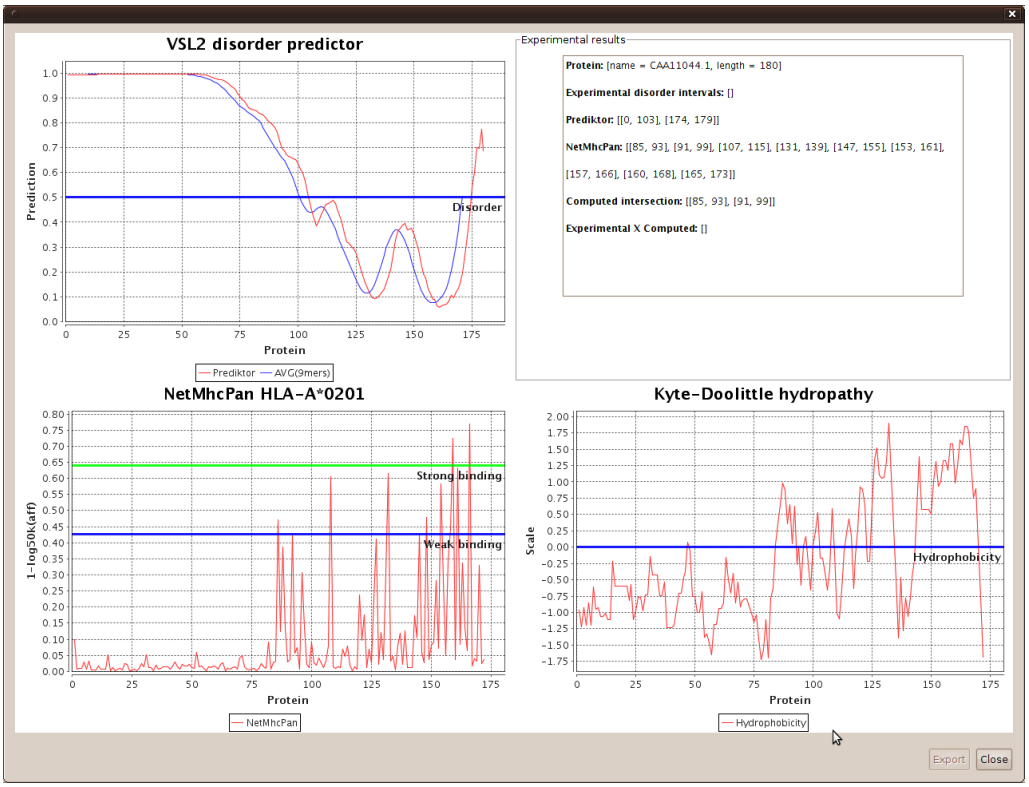
Bez obzira što epitopa nema, može se primetiti da afinitet vezivanja peptida u uređenim regionima raste što je „veća mera uređenosti“ (verovatnoća da amino kiselina ili peptid pripadaju uređenom regionu), kao i da afinitet opada što je veći stepen neuređenosti.

LAGE1 protein ima dva neuređena regiona (prema programu VSL2). Prvi (duži) je pretežno hidrofilan, dok drugi (kraći) skroz pripada hidrofilnom regionu. Uređen region je pretežno hidrofoban (sa izuzecima koji pripadaju hidrofilnom regionu).



Slika 31. LAGE1 MHC klasa I (alel HLA*A0101)

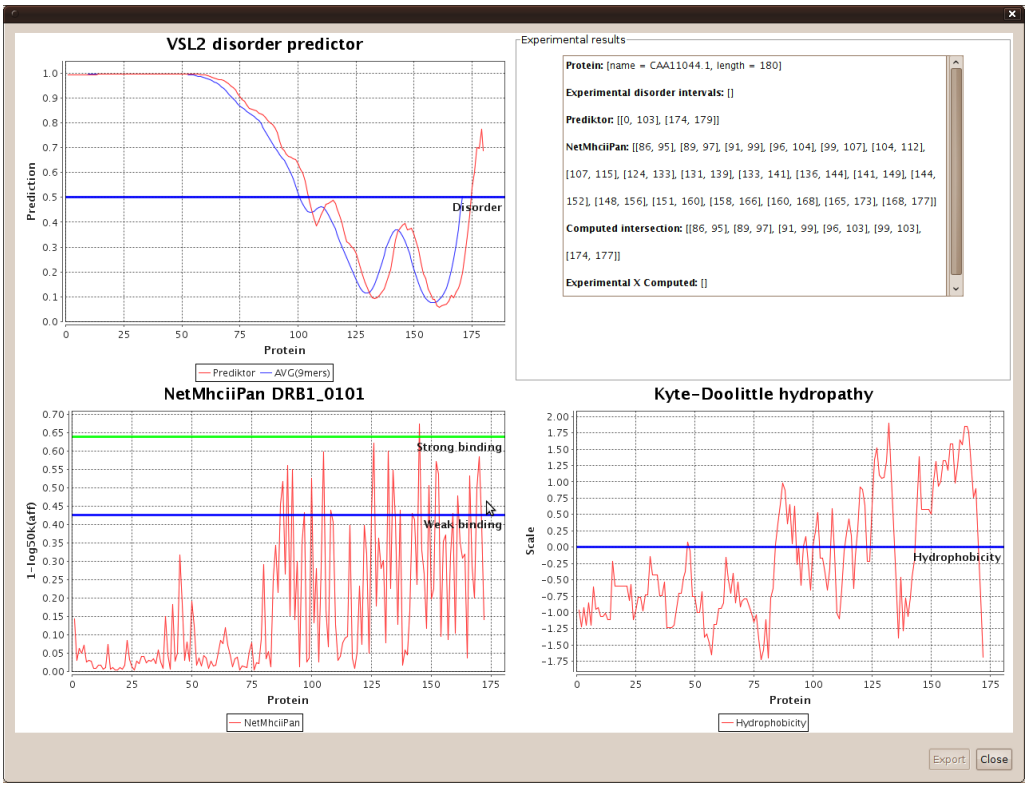
Mnogo veći broj epitopa proteina LAGE1 se vezuje za alel HLA_A0201, što je prikazano na slici 32:



Slika 32. LAGE1 MHC klasa I (alel HLA*A0201)

Prepoznati su i jaki i slabi epitopi za ovaj alel. Svi jaki epitopi se nalaze u uređenom regionu, kao i svi slabi epitopi uz dva izuzetka, koji se nalaze u neuređenom regionu. Oba ovakva epitopa se nalaze u blizini prelaza iz neuređenog regiona u uređeni.

Broj epitopa proteina LAGE1 koji se vezuju za molekule MHC klase II je znatno veći nego što je to slučaj za molekule MHC klase I (slika 33). Epitopi su najčešći u uređenim regionima sa malim brojem izuzetaka, koji su uglavnom slabi epitopi, u neuređenim regionima. Anologno prethodnim slučajevima, i ovde epitopi koji predstavljaju izuzetke pripadaju neuređenim regionima upravo na prelazu iz neuređenog u uređeni region.

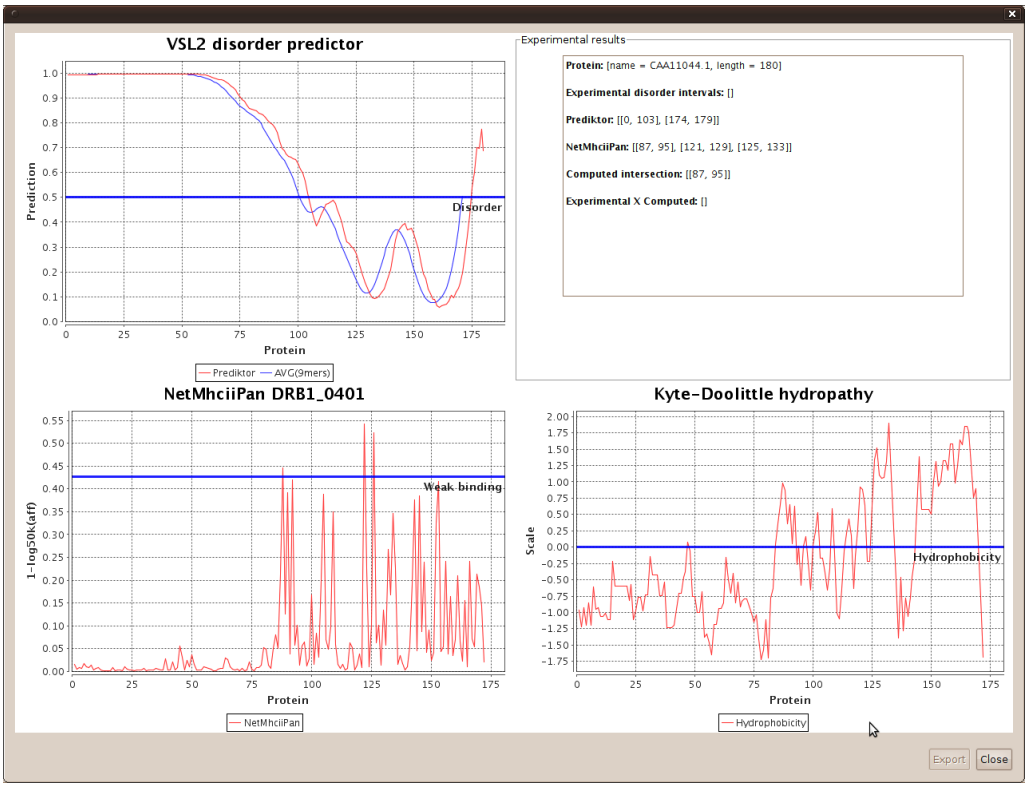


Slika 33. LAGE1 MHC klasa II (alel DRB1*0101)

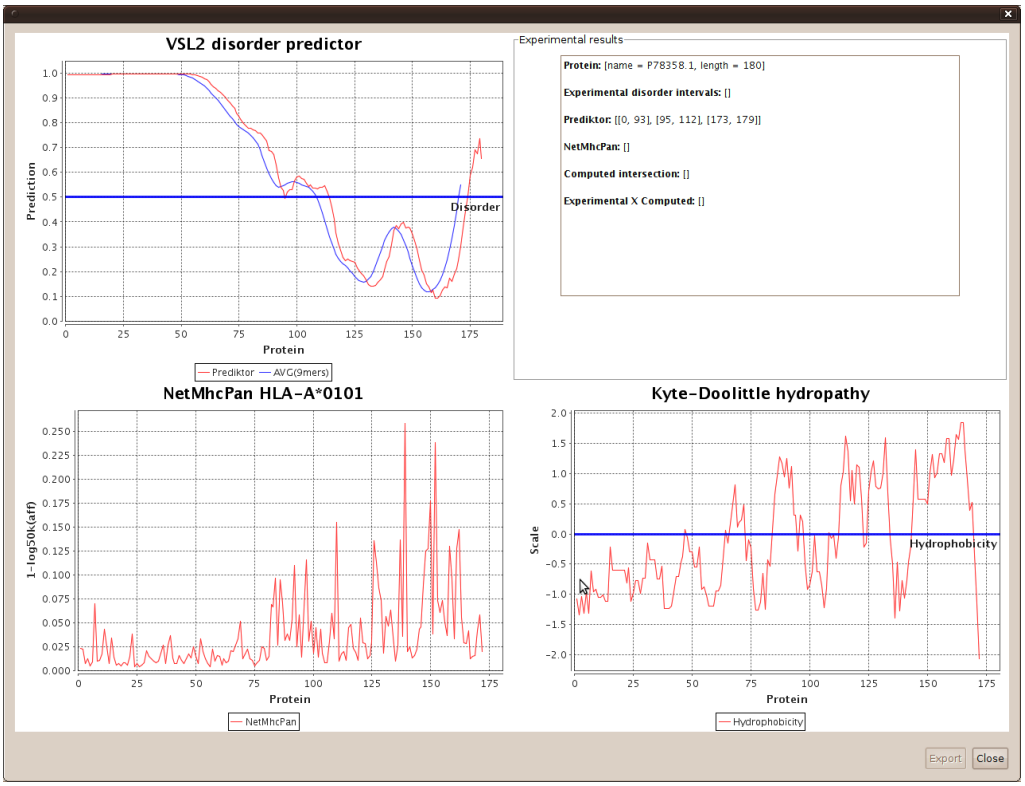
Naredni primer je dat za alel DRB1_0401 (slika 34). Prepoznat je jako mali broj epitopa, koji se nalaze u uređenom regionu.

Za protein LAGE2 nisi prepoznati epitopi koji se vezuju za molekul MHC klase I alela HLA_A0101 (slika 35). Međutim, sa grafika se vidi porast afiniteta vezivanja u uređenim regionima. Afinitet vezivanja opada što je stepen „neuređenosti" veći.

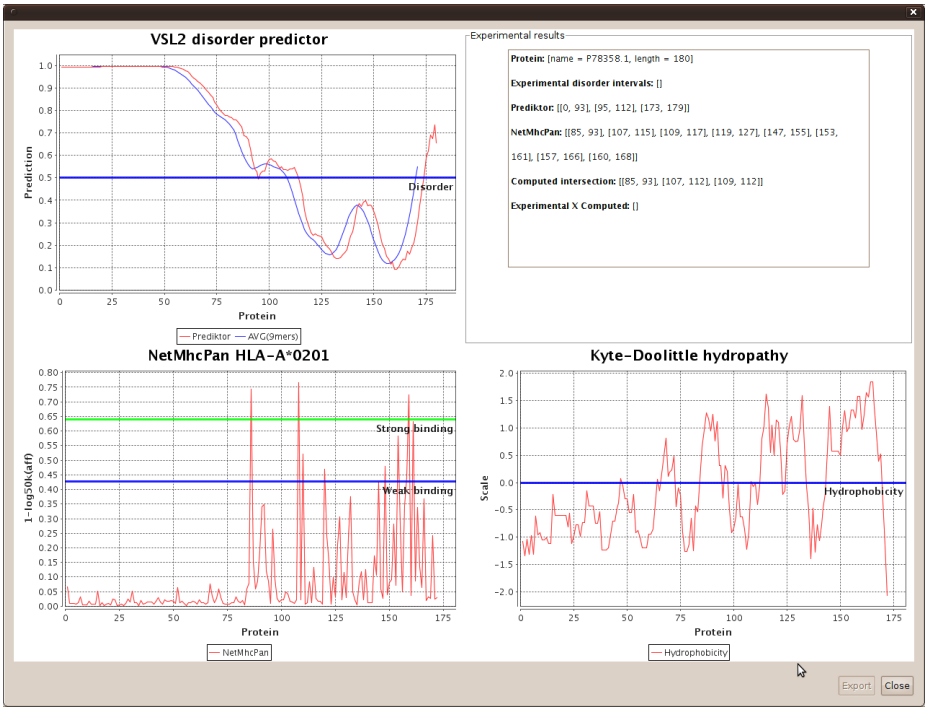
Za alel HLA_A0201 je prepoznat znatno veći broj epitopa (slika 36). Među prepoznatim epitopima ima i jakih. Epitopi se pretežno poklapaju sa uređenim regionima, izuzetak su tri epitopa koja pripadaju neuređenim strukturama. Takođe se vidi da je jedan od njih prepoznat kao jak epitop, što pokazuje da se i jaki epitopi nekada javljaju u neuređenim regionima.



Slika 34. LAGE1 MHC klasa II (alel DRB1*0401)

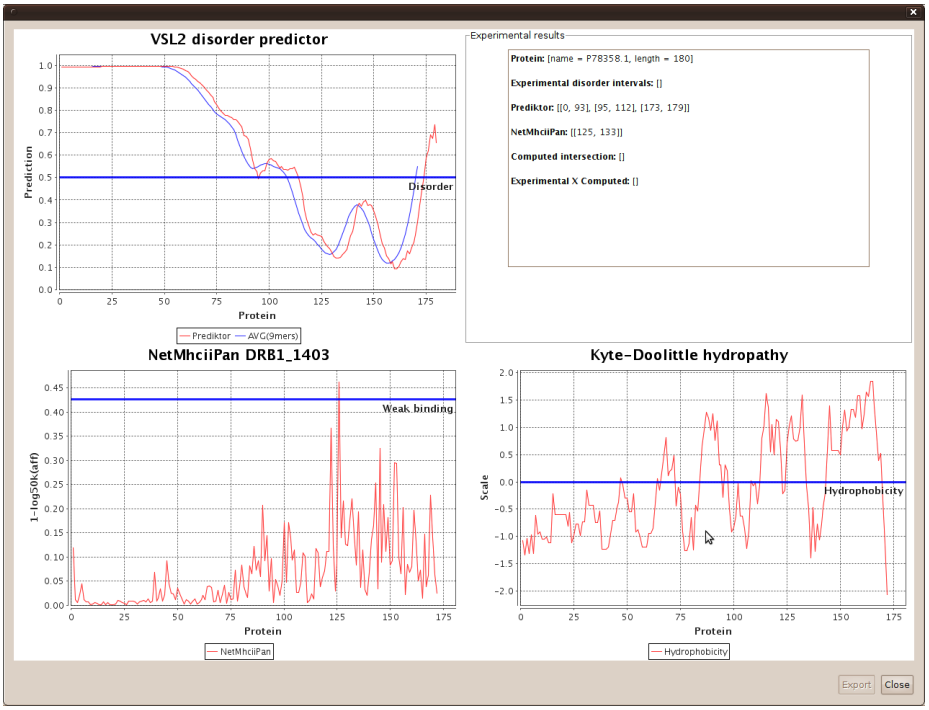


Slika 35. LAGE2 MHC klasa I (alel HLA*A0101)



Slika 36. LAGE 2

Na slici 37. je prikazano ponašanje proteina LAGE2 u interakciji sa molekulima MHC klase II (alel DRB1*1403). Za ovaj alel se vezuju samo slabi epitopi. Njihovo prisustvo je utvrđeno samo u uređenim strukturama proteina. Porast afiniteta vezivanja odgovara padu mere „neuređenosti”.



Slika 37. LAGE2 MHC klasa II (alel DRB1*1403)

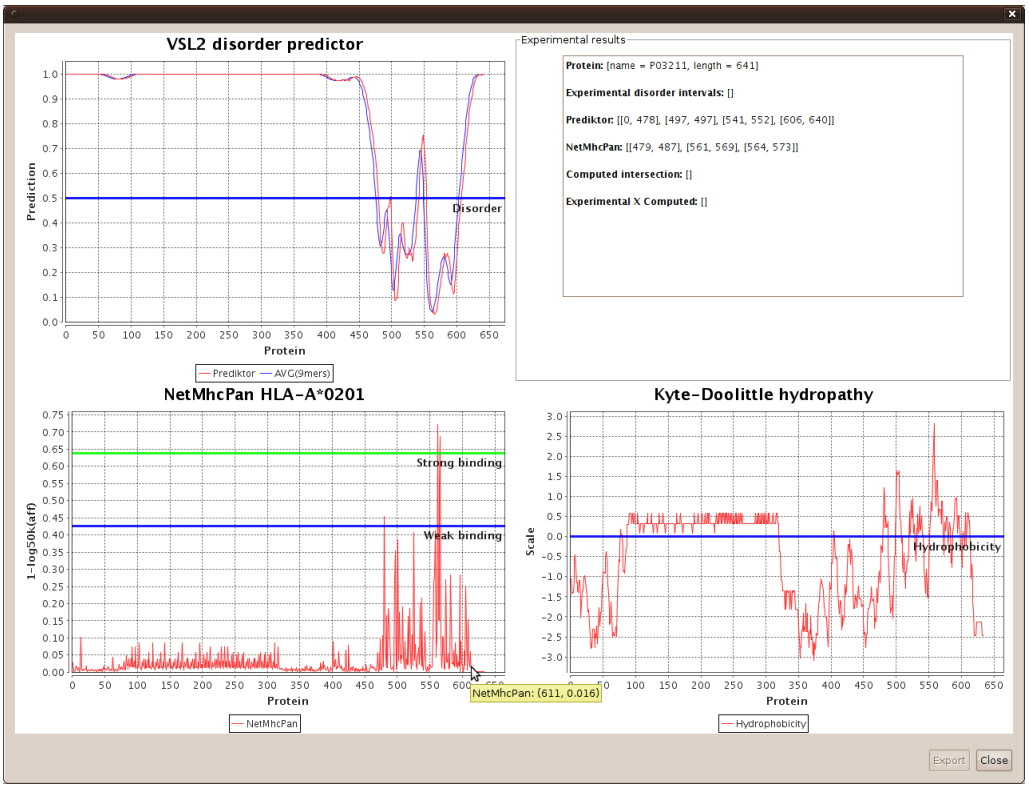
Za proteine MAGE 4, LAGE 1 i LAGE 2 iz funkcionalne grupe kancer-testis antigena važi da se epitopi retko javljaju u neuređenim regionima. Ako takvih izuzetaka ima onda su to slabi epitopi skoncetrisani na prelazima iz neuređenih u uređene regione.

Protein EBNA1 (Epstein – Bar virus)

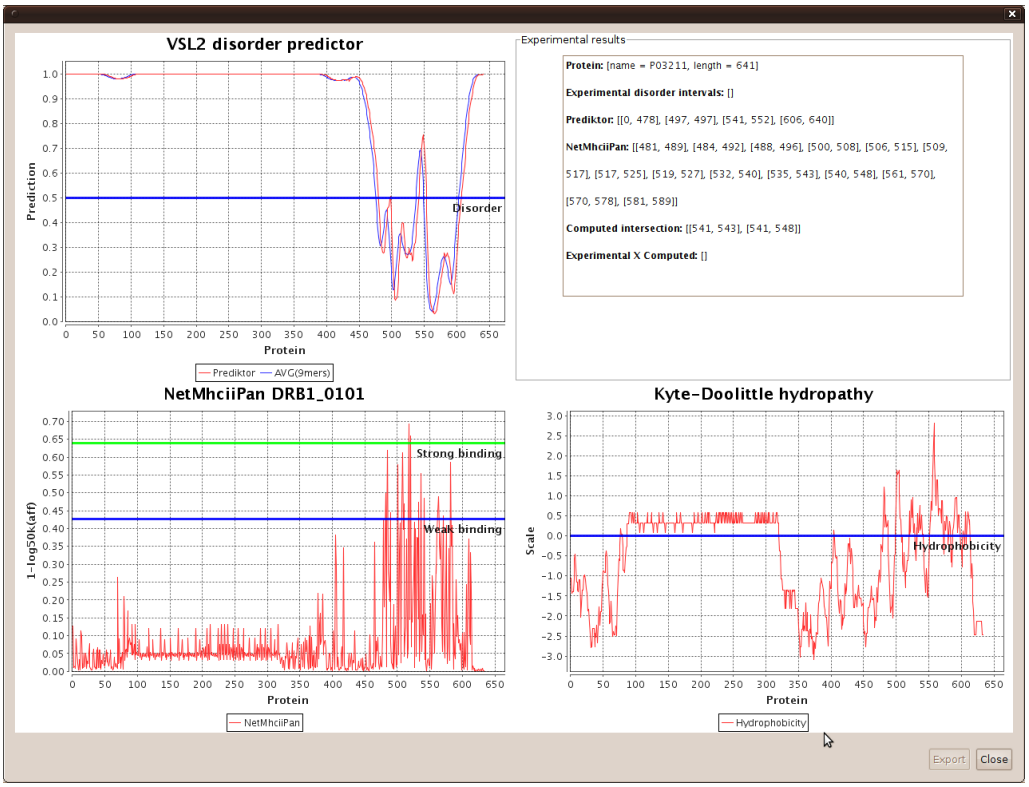
Proteini iz grupe Epstein Bar virusa odgovaraju različitim tipovima maligniteta, koje ovaj virus izaziva, ili post-transplatornim oboljenjima [6]. EBNA1 može izazvati i autoimuni odgovor kros-reaktivnošću sa SmB/B[1] autoantigenom [1]. Rezultati dobijeni za protein EBNA1 su prikazani na slikama 38 i 39.

Za protein EBNA1 su prepoznati i jaki i slabi epitopi koji se vezuju za alel HLA_A0101. Svi epitopi se nalaze u uređenim regionima proteina, iako se sa grafika može videti da protein ima jedan duži neuređeni region (preko 450 amino kiselina) i dva kraća.

Isto pravilo važi za MHC klasu II. Dat je prikaz alela DRB1_0101, koji je najčešći u ljudskoj populaciji. Prepoznati broj epitopa je mnogo veći nego za MHC klasu I, ali se svi preopoznati epitopi nalaze isključivo u uređenim regionima.



Slika 38. EBNA1 MHC klasa I (alel HLA*A0101)



Slika 39. EBNA1 MHC klasa II (alel DRB1*0101)

Nađeno je slaganje predikcije sa eksperimentalno dobijenim podacima za peptid-vezujući kapacitet više DRB1 alela, koje se za DRB1_0101 alel nalazi upravo u pretpostavljenom regionu jakog vezivanja (uređena struktura), tj od 475 do 552 A.K. [6].

Na osnovu predviđanja VSL2 programom, dobijeno je da je raspodela amino kiselina u neuređenim / uređenim regionima za proteine iz grupe kancer – testis antigena je približno jednaka: 53% amino kiselina pripada uređenim strukturama i 47% amino kiselina je u neuređenim strukturama.

5.2 **Rezultati za sve proteine**

Prikupljeno je 654 proteina iz različitih funkcionalnih grupa. Raspodela amino kiselina po uređenim odnosno neuređenim regionima je sledeća:

	Uređeni regioni	**Neuređeni regoni**
Broj AA (u %)	49.137	50.863

Urađeno je predviđanje epitopa za sve postojeće ljudske alele klasa HLA-1 i HLA-2. Za prikupljene proteinske sekvence razmatrano je preko 400 miliona peptida. Od kojih je dobijeno 4 682 644 epitopa za obe klase MHC I i MHC II. Epitopi su razvrstani po alelima kao:

1. ukupni (u uređenim i neuređenim sekvencama).
2. jaki (u uređenim i neuređenim sekvencama) i
3. slabi (u uređenim i neuređenim sekvencama).

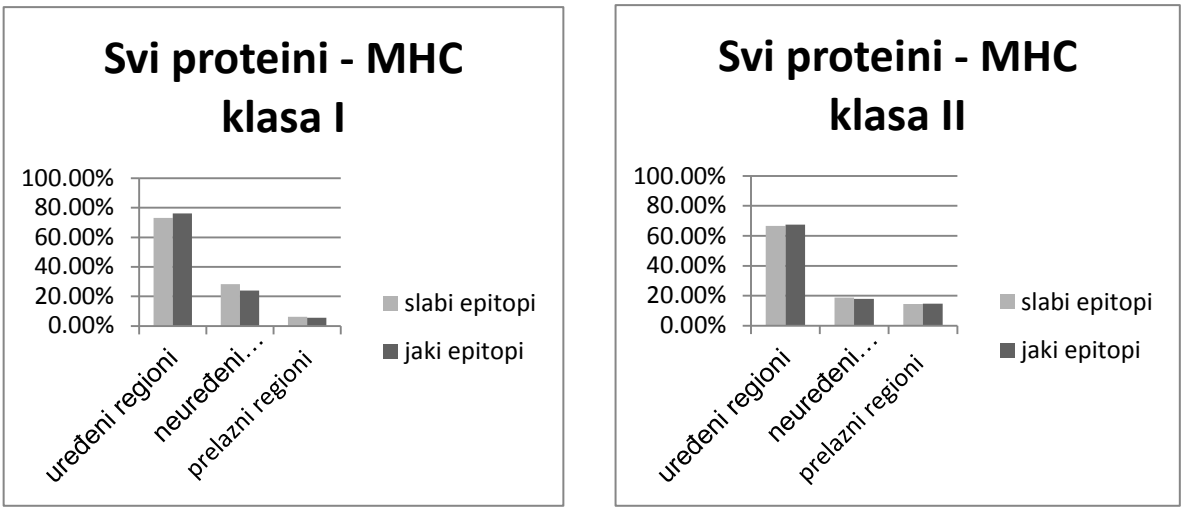
Rezultati su prikazani sledećom tabelom:

MHC I	
ukupan broj epitopa	1073894
ukupan broj slabih epitopa	822261
ukupan broj jakih epitopa	251633
neuređeni regioni:	
ukupan broj epitopa	215457
broj slabih epitopa	169510
broj jakih epitopa	45947
uređeni regioni:	
ukupan broj epitopa	792991
broj slabih epitopa	601216
broj jakih epitopa	191775
na prelaznim regionima:	
ukupan broj epitopa	65446
broj slabih epitopa	51535
broj jakih epitopa	13911

MHC II	
ukupan broj epitopa	3608750
ukupan broj slabih epitopa	3409297
ukupan broj jakih epitopa	199453
neuređeni regioni:	
ukupan broj epitopa	678708
broj slabih epitopa	643001
broj jakih epitopa	35707
uređeni regioni:	
ukupan broj epitopa	2403925
broj slabih epitopa	2269419
broj jakih epitopa	134506
na prelaznim regionima:	
ukupan broj epitopa	526117
broj slabih epitopa	496877
broj jakih epitopa	29240

Tabela 13. Broj i raspodela epitopa po regionima MHC klase I i MHC klase II

Epitopi koji se vezuju za molekule MHC klase II su brojniji, ali tvrđenje da se više epitopa nalazi u uređenim regionima važi za obe klase. Na slici 40 je prikazan odnos epitopa (jakih i slabih) u uređenim, neuređenim i prelaznim regionima. Grafici su dati za sve analizirane proteine, kao i za izdvojenu funkcionalnu grupu kancer-testis antigenih proteina (slike 40 i 41):

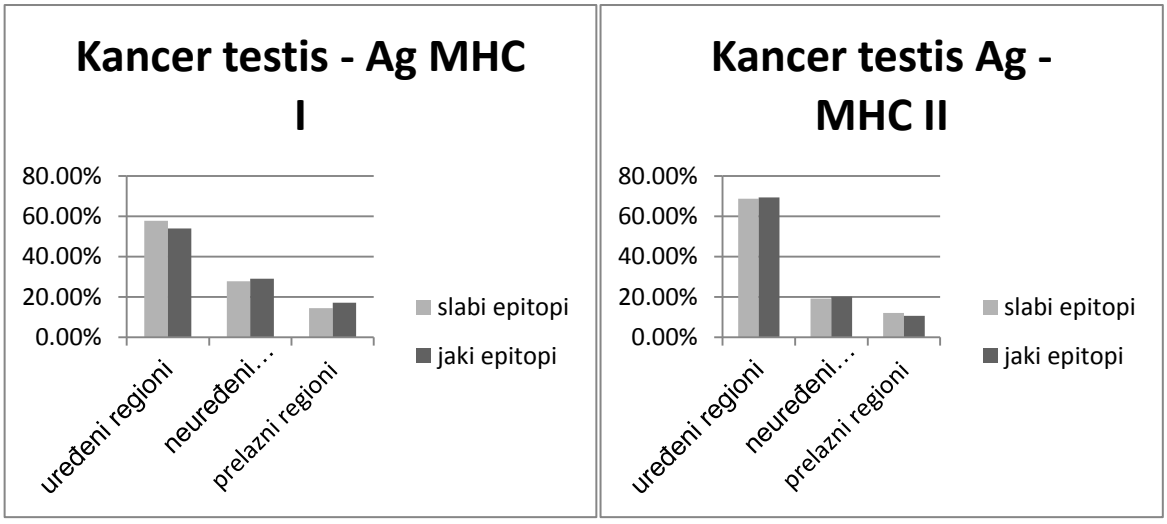


Slika 40. Raspodela epitopa za MHC klasu I i II u različitim strukturama proteina

Stubići na dijagramu predstavljaju odnos epitopa (jakih, slabih) i ukupan broj epitopa (jakih, slabih) po odgovarajućim regionima.

Odnos epitopa u uređenim / neuređenim regionima je zadržan u obe klase, iako je prepoznati broj epitopa koji se vezuju za molekule MHC klase II veći. Razlog tome je što je u drugom slučaju prepoznat veći broj epitopa na prelaznim regionima.

Rezultati dobijeni za kancer-testis antigene proteine su prikazani na slici 41:



Slika 41. Raspodela epitopa za kancer - testis antigene proteine u različitim strukturama proteina

Kod grupe Kancer-testis antigenih proteina je za MHC klasu II odnos epitopa u uređenim i neuređenim regionima veći nego što je to slučaj za sveukupne proteine, dok je taj odnos lošiji za MHC klasu I.

5.2.1 Rezultati dobijeni klaster analizom

Rezultati za MHC klasu I:

Za MHC klasu I epitopi su grupisani u pet klastera metodom neuronskog klasterovanja. U tabeli 14 su prikazani rezultati klasterovanja. Model je kreiran na slučajnom uzorku od 70% ukupnih podataka za ovu klasu, a testiran na preostalom skupu podataka.

HLA 1 - rezultati dobijeni neuronskim klasterovanjem epitopa					
Veličina klastera	**Struktura proteina**	**Vrsta epitopa**	**Hidrofobnost**	**Najdominantniji epitop**	**Najdominantniji alel**
43.58%	Skroz uređena	Slabi epitopi	[-1.5, 2]	ITTQSTLPY	HLA*A0276
22.38%	Skroz neuređena	Slabi epitopi	[-2, 1.5]	TSFESMIEY	HLA*A0269
15.39%	Prelazni regioni	Svi epitopi	[-2, 1.5]	MSLPMNSLY	HLA*A0269
12.55%	Skroz uređena	Jaki epitopi	[-2, 3.5]	IWEEGTFNI	HLA*A0269
6.10%	Skroz neuređena	Jaki epitopi	[-2, 1.5]	SSNSSFLSF	HLA*A0211

Tabela 14. Rezultati dobijeni neuronskim klasterovanje za HLA 1 klasu

Tačnost modela je 0.978, homogenost po klasterima i sličnost između klastera je prikazana u tabeli 15.

▼ Cluster Statistics

Statistic

ID	Abs. Size	Size (%)	Homogeneity
[2] 2	55,924	43.58%	0.981
[3] 3	28,715	22.38%	0.984
[1] 1	19,747	15.39%	0.966
[5] 5	16,108	12.55%	0.974
[4] 4	7,834	6.10%	0.977

Similarity Between Clusters

Cluster	Cluster	Similarity
[2] 2	[3] 3	0.0254
[2] 2	[1] 1	0.0699
[2] 2	[5] 5	0.0977
[2] 2	[4] 4	0.0641
[3] 3	[1] 1	0.0952
[3] 3	[5] 5	0.1049
[3] 3	[4] 4	0.0833
[1] 1	[5] 5	0.1086
[1] 1	[4] 4	0.0566
[5] 5	[4] 4	0.0784

Tabela 15. Statistike neuronskog klasterovanja epitopa klase HLA 1

U tabeli 16 su prikazani najčešći epitopi u prvom klasteru koga čine samo slabi epitopi u uređenim regionima, kao i najčešći aleli koji prepoznaju epitope ove grupe. Isti skup alela je najdominantniji u svim strukturama proteina (uređeni, neuređeni i prelazni regioni). Hidrofobnost slabih epitopa u uređenim regionima je skoncetrisana u intervalu [-1.5, 2].

Category	▼ Cluster	Whole population
ITTQSTLPY	84	111
RMVYPQPKV	68	76
RLQNTTIGL	64	86
RTAGINGLY	63	91
FSAHGSPYY	58	88
QSIDEVWAY	56	71
GSNTSIHAY	52	55
RTASCALAF	51	89
VVTPWLAPI	50	78
STAEGIQLY	50	69
FVSENVTQY	48	69
AINSKQLTY	48	53
AVLMWVFTY	45	62
FLKLFLETA	42	52
IVQPENLEY	39	44

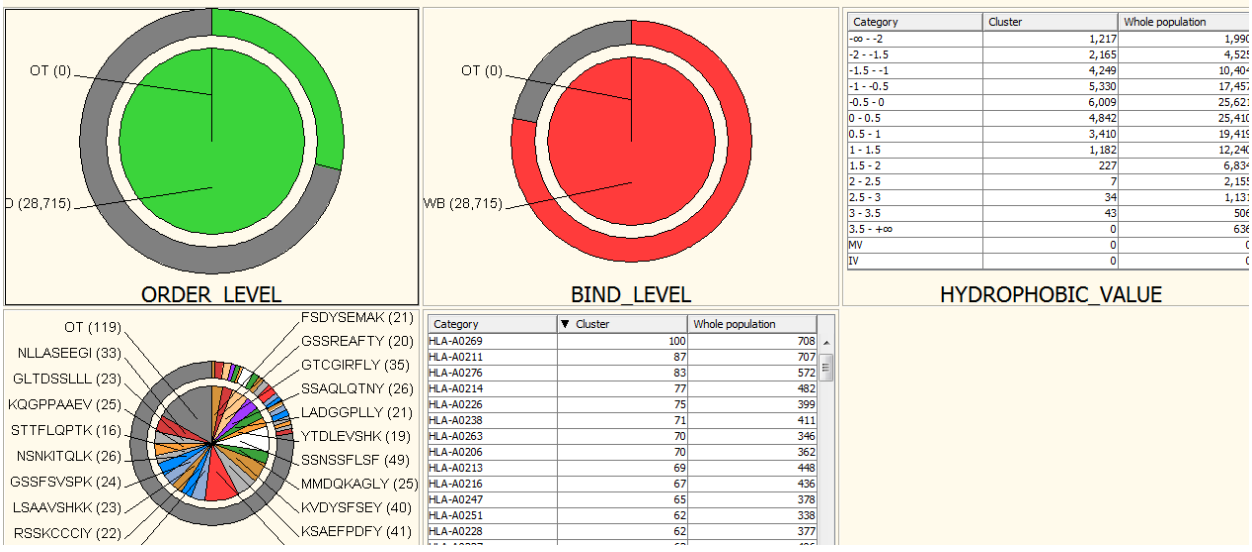
PEPTIDE

Category	▼ Cluster	Whole population
HLA-A0276	269	572
HLA-A0211	224	707
HLA-A0269	217	708
HLA-A0214	206	482
HLA-A0248	175	376
HLA-A0272	167	354
HLA-A0247	166	378
HLA-A0228	165	377
HLA-A0244	164	407
HLA-A0206	162	362
HLA-A0213	159	448
HLA-A0279	157	352
HLA-A0221	155	350
HLA-A0257	153	270
HLA-A0216	153	436

ALLELE_CODE

Tabela 16. a) Najdominantniji epitopi u prvom klasteru; b) najdominantniji aleli u uređenim regionima

Drugu grupu (klaster) po veličini čine epitopi u neuređenim regionima (22.38% epitopa). Epitopi koji pripadaju ovoj grupi su uglavnom hidrofilnog karaktera. Rezultati za ovu grupu su prikazani na slici 42:



Slika 42. Klaster br. 2 sa slabim epitopima u neuređenim regionima

Analizom dobijenih grupa vidi se da atributi „order_level“ i „bind_level“ (koji predstavljaju strukturu proteina i vrstu epitopa, redom) određuju podelu po klasterima, dok su aleli

ravnomerno zastupljeni u svim klasterima. Odavde sledi zaključak da su za razmatranu grupu proteina (koju čine svi prikupljeni proteini osim bakterijskih) dominantni isti aleli, koji prepoznaju epitope u svim regionima proteina (nezavisno od strukture).

Rezultati koji su do sada prikazani se odnose na sve analizirane proteine osim bakterijskih (131 protein). Epitope u bakterijskim proteinima prepoznaje druga grupa alela, koja nema preseka sa prethodnom. Najdominantniji aleli za bakterijske proteine su dati u tabeli 17.

Category	▼ Cluster	Whole population
HLA-B1503	1,612	2,616
HLA-B1516	1,476	2,473
HLA-B1517	1,144	2,125
HLA-B1529	1,032	1,476
HLA-B1525	901	1,263
HLA-B0813	867	1,201
HLA-B1528	722	976
HLA-B1501	722	976
HLA-B1505	671	832
HLA-B1524	659	779
HLA-B1506	650	844
HLA-B1502	467	599
HLA-B0750	453	626
HLA-B1507	448	556
HLA-B0737	441	834

Tabela 17. Najdominantniji aleli u prepoznavanju epitopa bakterijskih proteina

Rezultati za MHC klasu II:

Za epitope MHC klase II demografskim klasterovanjem dobijamo osnovne statistike o raspodeli epitopa po strukturnim regionima proteina. Rezultati demografskog klasterovanja su prikazani u tabeli 18:

HLA 2 – Rasprostiranje eptopa po uređenim/neuređenim regionima					
Veličina klastera	**Struktura proteina**	**Vrsta epitopa**	**Hidrofobnost**	**Najdominantniji epitop**	**Najdominantniji alel**
62.89%	Skroz uređena	Slabi epitopi	[-1.5, 4]	LQSMRALDF	DRB1*0101
17.82%	Skroz neuređena	Slabi epitopi	[-1.5, 1.5]	FPRMSNLRL	DRB1*0101
13.77%	Prelazni regioni	Slabi epitopi	[-1.5, 2]	MNKLKKMAL	DRB1*0101
3.73%	Skroz uređena	Jaki epitopi	[-1.5, 3.5]	VEVLQSMRA	DRB1*0101
0.99%	Skroz neuređena	Jaki epitopi	[-1.5, 2]	FPRMSNLRL	DRB1*0101
0.81%	Prelazni regioni	Jaki epitopi	[-1.5, 2.5]	FKMIDTDNS	DRB1*0101

Tabela 18. Demografsko klasterovanje epitopa MHC klase II

Demografskim klasterovanjem se epitopi grupišu u zavisnosti od toga da li pripadaju uređenim / neuređenim regionima i tipu epitopa (jaki ili slabi). Tačnost dobijenog modela je 0.733. Neuronskim klasterovanjem se dobija bolja podela koja uzima u obzir peptid koji predstavlja epitop, hidrofobnu vrednost epitopa i alele koje ga prepoznaju. Dobijeni model grupiše epitope nešto drugačije po strukturnim regionima proteina. Model je tačnosti 0.967, sličnost slogova u klasteru je veća kao i „udaljenost“ klastera.

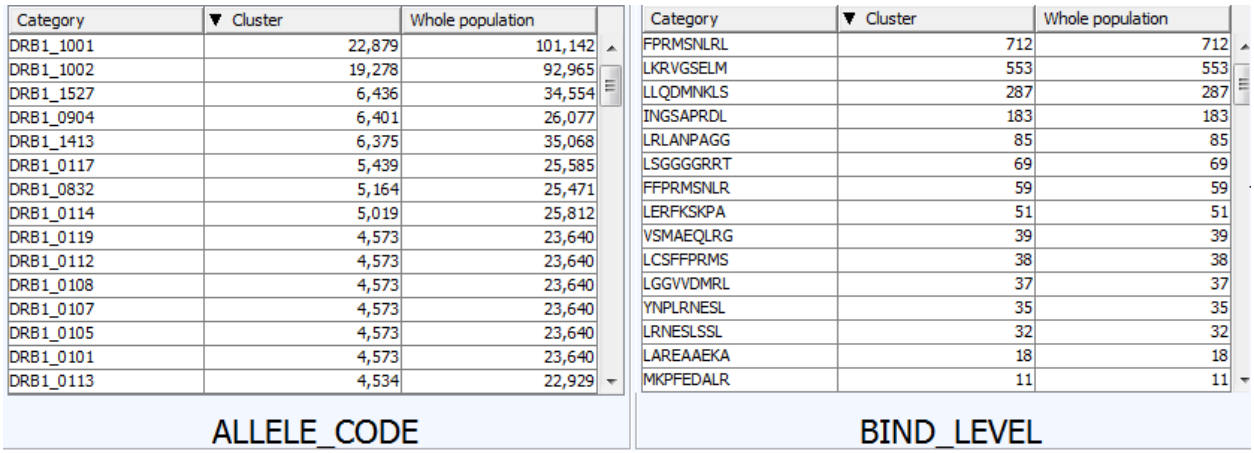
U tabeli 19 su prikazani rezultati neuronskog klasterovanja epitopa MHC klase II za sve proteine.

HLA 2 – rezultati dobijeni neuronskim klasterovanjem epitopa					
Veličina klastera	**Struktura proteina**	**Vrsta epitopa**	**Hidrofobnost**	**Najdominantniji epitop**	**Najdominantniji alel**
32.77%	Skroz uređena	Slabi epitopi	[0, 4]	VSYLVRYMG	DRB1*1527
30.11%	Skroz uređena	Slabi epitopi	[-1.5, 1]	LQSMRALDF	DRB1*0101
18.81%	Skroz neuređena	Svi epitopi	[-1.5, 2]	FPRMSNLRL	DRB1*0101
14.58%	Prelazni regioni	Svi epitopi	[-1.5, 2]	MNKLKKMAL	DRB1*0101
3.73%	Skroz uređena	Jaki epitopi	[-1.5, 3.5]	VEVLQSMRA	DRB1*0101

Tabela 19. Grupisanje epitopa MHC klase II neuronskim klasterovanjem

Prva dva klastera čine slabi epitopi u uređenim strukturama. Razlika između ova dva klastera je što prvu grupe čine izuzetno hidrofobni epitopi, a drugu uglavnom hidrofilni. Epitopi koji se nalaze u neuređenim regionima su grupisani zajedno, nezavisno od toga da li su jaki ili slabi. Bez obzira na tip epitopa (jaki / slabi) hidrofobnost u neuređenim regionima je u intervalu [-1.5, 2]. Najdominantniji alel za ove epitope je DRB1*0101, koji je i najčešći ljudski alel. Epitopi MHC klase II retko imaju hidrofobnu vrednost ispod -1.5.

Izdvojeni su i najdominantniji aleli u neuređenim regionima, kao i najdominantniji epitopi u istim. Izdvojeni epitopi se javljaju samo u neuređenim regionima. U tabeli 20 su prikazani najdominantniji aleli u neuređenim regionima, (oni su najdominantniji i u drugim strukturama proteina), i najdominantniji epitpi u neuređenim regioninima.



Category	Cluster	Whole population
DRB1_1001	22,879	101,142
DRB1_1002	19,278	92,965
DRB1_1527	6,436	34,554
DRB1_0904	6,401	26,077
DRB1_1413	6,375	35,068
DRB1_0117	5,439	25,585
DRB1_0832	5,164	25,471
DRB1_0114	5,019	25,812
DRB1_0119	4,573	23,640
DRB1_0112	4,573	23,640
DRB1_0108	4,573	23,640
DRB1_0107	4,573	23,640
DRB1_0105	4,573	23,640
DRB1_0101	4,573	23,640
DRB1_0113	4,534	22,929

ALLELE_CODE

Category	Cluster	Whole population
FPRMSNLRL	712	712
LKRVGSELM	553	553
LLQDMNKLS	287	287
INGSAPRDL	183	183
LRLANPAGG	85	85
LSGGGGRRT	69	69
FFPRMSNLR	59	59
LERFKSKPA	51	51
VSMAEQLRG	39	39
LCSFFPRMS	38	38
LGGVVDMRL	37	37
YNPLRNESL	35	35
LRNESLSSL	32	32
LAREAAEKA	18	18
MKPFEDALR	11	11

BIND_LEVEL

Tabela 20. levo: Najdominatniji aleli MHC klase II u neuređenim regionima (a takođe i u ostalim strukturama); desno: Najdominantniji epitopi u neuređenim regionima (klaster 3)

5.2.2 Pravila pridruživanja – epitopi i aleli

Primenom tehnike pravila pridruživanja za epitope MHC klase I, izdvojena su pravila sa najvećom podrškom i nivoom poverenja 100%. Značajna pravila dobijena na ovaj način su:

- Slabi epitopi koji se nalaze u neuređenim regionima imaju hidrofobnost iz intervala [-1.9, 1.1].
- Kod bakterijskih proteina i jaki i slabi epitopi uzimaju vrednosti iz intervala [-1.9, 1.1] u neuređenim regionima. U uređenim regionim iz intervala [-0.6, 2.4], sa retkim izuzecima preko 2.4.
- Izdvojeni su i epitopi koji se zajedno javljaju u neuređenim regionima. Rezultat je prikazan u tabeli 21.
- Pronađeni su aleli koji najčešće prepoznaju iste epitope (tzv. promiskuitetne epitope). U tabeli 22 su prikazani „srodni aleli" (prepoznaju iste epitope) u neuređenim regionima.

Visible rules:

ID	Rule	Support	▼ Confidence	Lift
6,887	[WQAPMFDAI] ==> [WQIPHSYAI]	26.6571%	100.0000%	2.58
6,872	[YQVFSVLSL] ==> [WQIPHSYAI]	25.4323%	100.0000%	2.58
6,821	[YQVFSVLSL] +[RQFMWTTLL] ==> [WQIPHSYAI]	23.9914%	100.0000%	2.58
6,816	[LPLPLTIPL] ==> [LPMSLPLPL]	23.7752%	100.0000%	3.48
6,810	[RQWFLDLPL] +[WQAPMFDAI] ==> [WQIPHSYAI]	23.6311%	100.0000%	2.58
6,802	[RQWFLDLPL] +[YQVFSVLSL] ==> [WQIPHSYAI]	23.4870%	100.0000%	2.58
6,673	[MQQAYAAPM] +[YQVFSVLSL] ==> [WQIPHSYAI]	22.4784%	100.0000%	2.58
6,657	[WQAPMFDAI] +[RQFMWTTLL] ==> [WQIPHSYAI]	22.4784%	100.0000%	2.58
6,670	[YQVFSVLSL] +[WQAPMFDAI] ==> [WQIPHSYAI]	22.4784%	100.0000%	2.58
6,676	[RQWFLDLPL] +[YQVFSVLSL] +[RQFMWTTLL] ==> [WQIPHSYAI]	22.4784%	100.0000%	2.58
6,637	[RQWFLDLPL] +[LQFWLNILL] ==> [WQIPHSYAI]	22.4063%	100.0000%	2.58
6,635	[MQQAYAAPM] +[WQAPMFDAI] ==> [WQIPHSYAI]	22.4063%	100.0000%	2.58
6,598	[YQVFSVLSL] +[RQYDRVAEL] ==> [WQIPHSYAI]	22.2622%	100.0000%	2.58
6,583	[LPTTMNYPL] +[LPYALRIEL] ==> [LPMSLPLPL]	22.1902%	100.0000%	3.48

Tabela 21. Epitopi koji se zajedno javljaju u neuređenim regionima

Analogno su izdvojeni aleli koji u neuređenim regonima prepoznaju isti epitop, prikazani su u tabeli 22.

Rule	Support	▼ Confidence
[ALLELE_CODE=HLA-B1569] +[ALLELE_CODE=HLA-B1598] +[ALLELE_CODE=HLA-B9503] ==> [ALLELE_CODE=HLA-B1503]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B1598] ==> [ALLELE_CODE=HLA-B1503]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B1569] +[ALLELE_CODE=HLA-B9527] +[ALLELE_CODE=HLA-B1503] +[ALLELE_CODE=HLA-B9503] ==> [ALLELE_CODE=HLA-B9532]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B9532] +[ALLELE_CODE=HLA-B1598] +[ALLELE_CODE=HLA-B1503] ==> [ALLELE_CODE=HLA-B1569]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B9532] +[ALLELE_CODE=HLA-B1569] +[ALLELE_CODE=HLA-B1503] ==> [ALLELE_CODE=HLA-B9527]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B1569] +[ALLELE_CODE=HLA-B9527] ==> [ALLELE_CODE=HLA-B9532]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B9527] +[ALLELE_CODE=HLA-B1503] +[ALLELE_CODE=HLA-B9503] ==> [ALLELE_CODE=HLA-B1569]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B1569] +[ALLELE_CODE=HLA-B1598] ==> [ALLELE_CODE=HLA-B9532]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B9532] ==> [ALLELE_CODE=HLA-B1598]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B9527] ==> [ALLELE_CODE=HLA-B1503]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B1569] +[ALLELE_CODE=HLA-B9527] ==> [ALLELE_CODE=HLA-B1598]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B9532] +[ALLELE_CODE=HLA-B1503] +[ALLELE_CODE=HLA-B9503] ==> [ALLELE_CODE=HLA-B1569]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B1569] +[ALLELE_CODE=HLA-B9527] +[ALLELE_CODE=HLA-B1503] ==> [ALLELE_CODE=HLA-B9532]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B1569] +[ALLELE_CODE=HLA-B9503] ==> [ALLELE_CODE=HLA-B1598]	19.5395%	100.0000%
[ALLELE_CODE=HLA-B9532] +[ALLELE_CODE=HLA-B1569] +[ALLELE_CODE=HLA-B1598] +[ALLELE_CODE=HLA-B1503] ==> [ALLELE_CODE=HLA-B9503]	19.5395%	100.0000%

Tabela 22. Aleli koji se ponašaju slično u neuređenim regionima

Izdvojena je grupa alela koja je najdominantnija u svim strukturama proteina, tj. prepoznaje najveći broj epitopa i to su aleli HLA_A grupe. Takođe je dobijena i prikazana grupa alela koja prepoznaje promiskuitetne epitope u neuređenim regionima, a to su aleli HLA_B grupe. Sledi da pored toga što prva grupa alela prepoznaje najviše epitopa u svim strukturama proteina (tako i u neuređenim), u neuređenim regionima prepoznati epitopi su retko promiskuitetni.

Aleli MHC klase II koji se najčešće javljaju zajedno (prepoznaju iste epitope), su prikazani u tabeli 23 (Dat je samo delimičan prikaz rezultata). Ustanovljeno ponašanje važi i u uređenim i neuređenim regionima.:

ID	Rule	Support	▼ Confidence	Lift
9,986	[DRB1_0119]+[DRB1_0105]+[DRB1_0108]+[DRB1_0101] ==> [DRB1_0112]	51.2847%	100.0000%	1.9
9,951	[DRB1_0119]+[DRB1_0108]+[DRB1_0101]+[DRB1_0112] ==> [DRB1_0105]	51.2847%	100.0000%	1.9
9,826	[DRB1_0107] ==> [DRB1_0105]	51.2847%	100.0000%	1.9
9,877	[DRB1_0108]+[DRB1_0107] ==> [DRB1_0112]	51.2847%	100.0000%	1.9
9,940	[DRB1_0119]+[DRB1_0107]+[DRB1_0101] ==> [DRB1_0112]	51.2847%	100.0000%	1.9
9,870	[DRB1_0108]+[DRB1_0107]+[DRB1_0112] ==> [DRB1_0105]	51.2847%	100.0000%	1.9
9,884	[DRB1_0105]+[DRB1_0112] ==> [DRB1_0119]	51.2847%	100.0000%	1.9
9,959	[DRB1_0119]+[DRB1_0108]+[DRB1_0107]+[DRB1_0101] ==> [DRB1_0105]	51.2847%	100.0000%	1.9
9,882	[DRB1_0105]+[DRB1_0112] ==> [DRB1_0107]	51.2847%	100.0000%	1.9
9,993	[DRB1_0119]+[DRB1_0105]+[DRB1_0108] ==> [DRB1_0101]	51.2847%	100.0000%	1.9
9,921	[DRB1_0105]+[DRB1_0108] ==> [DRB1_0101]	51.2847%	100.0000%	1.9
9,946	[DRB1_0119]+[DRB1_0107] ==> [DRB1_0105]	51.2847%	100.0000%	1.9
9,966	[DRB1_0119]+[DRB1_0108] ==> [DRB1_0105]	51.2847%	100.0000%	1.9
9,938	[DRB1_0119]+[DRB1_0107]+[DRB1_0101]+[DRB1_0112] ==> [DRB1_0108]	51.2847%	100.0000%	1.9
9,853	[DRB1_0107]+[DRB1_0101]+[DRB1_0112] ==> [DRB1_0119]	51.2847%	100.0000%	1.9
9,822	[DRB1_0101] ==> [DRB1_0119]	51.2847%	100.0000%	1.9
9,819	[DRB1_0101] ==> [DRB1_0107]	51.2847%	100.0000%	1.9
9,868	[DRB1_0108]+[DRB1_0101] ==> [DRB1_0119]	51.2847%	100.0000%	1.9
9,961	[DRB1_0119]+[DRB1_0108]+[DRB1_0107] ==> [DRB1_0101]	51.2847%	100.0000%	1.9

Tabela 23. Aleli MHC klase II koji prepoznaju promiskuitetne epitope

U tabeli 24. su prikazani promiskuitetni epitopi MHC klase II koji se uvek javljaju zajedno. Prikazan je samo deo rezultata.

Visible rules:

ID	Rule	Support	▼ Confidence	Lift
7,509	[LKKSCTLYV] ==> [IIMSTSLRV]	88.3946%	100.0000%	1.03
7,508	[LILSLSLAL] ==> [IIMSTSLRV]	87.4275%	100.0000%	1.03
7,503	[INKKVSLLL] ==> [IIMSTSLRV]	86.8472%	100.0000%	1.03
7,498	[IMLNPSSRI] ==> [IIMSTSLRV]	85.1064%	100.0000%	1.03
7,494	[IRQSRNLRR]+[LKKSCTLYV] ==> [IIMSTSLRV]	84.5261%	100.0000%	1.03
7,486	[LLYSRGLLI] ==> [IIMSTSLRV]	84.1393%	100.0000%	1.03
7,483	[IRQSRNLRR]+[LILSLSLAL] ==> [IIMSTSLRV]	83.7524%	100.0000%	1.03
7,476	[LILSLSLAL]+[LKKSCTLYV] ==> [IIMSTSLRV]	83.5590%	100.0000%	1.03
7,479	[INKKVSLLL]+[LKKSCTLYV] ==> [IIMSTSLRV]	83.5590%	100.0000%	1.03
7,466	[LLYSRGLLI]+[LILSLSLAL] ==> [IIMSTSLRV]	83.1721%	100.0000%	1.03
7,457	[LSMLVSMLI] ==> [IIMSTSLRV]	82.5919%	100.0000%	1.03
7,452	[ILVNVSLGV] ==> [IIMSTSLRV]	82.2050%	100.0000%	1.03
7,451	[IMLNPSSRI]+[LKKSCTLYV] ==> [IIMSTSLRV]	82.2050%	100.0000%	1.03
7,442	[INKKVSLLL]+[IRQSRNLRR] ==> [IIMSTSLRV]	82.0116%	100.0000%	1.03
7,430	[IRQSRNLRR]+[LLYSRGLLI] ==> [IIMSTSLRV]	81.8182%	100.0000%	1.03
7,437	[LSMLVSMLI]+[IMLNPSSRI] ==> [IIMSTSLRV]	81.8182%	100.0000%	1.03
7,413	[LLYSRGLLI]+[LKKSCTLYV] ==> [IIMSTSLRV]	81.6248%	100.0000%	1.03
7,401	[ILVNVSLGV]+[LILSLSLAL] ==> [IIMSTSLRV]	81.4313%	100.0000%	1.03
7,395	[LKKSCTLYV]+[INLSRSAAR] ==> [IIMSTSLRV]	81.4313%	100.0000%	1.03
7,387	[INKKVSLLL]+[LILSLSLAL] ==> [IIMSTSLRV]	81.0445%	100.0000%	1.03
7,383	[IRQSRNLRR]+[LILSLSLAL]+[LKKSCTLYV] ==> [IIMSTSLRV]	81.0445%	100.0000%	1.03
7,364	[ILVNVSLGV]+[LKKSCTLYV] ==> [IIMSTSLRV]	80.8511%	100.0000%	1.03

Tabela 24. Epitopi MHC klase II koji se uvek zajedno javljaju

6 Zaključak

U radu je analizirana veza između neuređenih / uređenih delova i antigenih regioni u proteinu, učestalost pojavljivanja antigenih regiona u različitim strukturama proteina za sve, do danas, poznate ljudske alele. U obzir je uzeta i hidrofobna vrednost antigenih regiona – epitopa i ustanovljen je interval hidropatije u neuređenim regionima proteina. Pronalaženje navedenih regiona je izvedeno programima za predviđanje VSL2, NetMhcPan i NetMhciiPan. Nad dobijenim rezultatima je sprovedeno istraživanja podataka (klasterovanje i pravila pridruživanja). Relativna jednostavnost tih tehnika leži u činjenici da oni podatke nad kojima se primenjuju obrađuju u potpunosti sintaksno i bez potrebe za ekspertskom pomoći. Ove tehnike su primenjene u cilju klasterovanja antigenih regiona (epitopa). Veoma je važno grupisati epitope i alele koje ih prepoznaju prema strukturi i funkcionalnoj grupi proteina kojoj pripadaju, jer se tako može ustanoviti ponašanje imunološkog sistema. Ustanovljena pravila i rezultati su važni za pravljenje i poboljšanje vakcina zasnovanih na peptid koktelima za različite vrste bolesti: alergije, autoimune bolesti, kancerogena oboljenja, itd. Rezultati klasterovanja epitopa prema strukturi proteina kojoj pripadaju, hidrofobnoj vrednosti i alelima koje ga prepoznaju dobijeni tehnikama prikazanim u ovom radu su velike tačnosti. Za potrebe obrade podataka dobijenih navedenim programima je napisana aplikacija EPDIS. Aplikacija je napisana u programskom jeziku Java, verzija 6. EPDIS aplikacija je omogićila grafički prikaz rezultata navedenih programa, analizu rezultata za pojedinačne proteine i alele, skladištenje dobijenih epitopa i neuređenih regiona u relacionu bazu podataka DB2 sistema, i određivanje preklapanja intervala koji čine epitop sa uređenim / neuređenim delovima proteina i pridruživanje hidrofobne vrednosti svakom epitopu.

Dosadašnje pretpostavke su da je broj epitopa u uređenim regionima veći nego u neuređenim, ali su sve pretpostavke o raspodeli epitopa po uređenim / neuređenim regionima zasnovane na analizi pojedinačnog proteina za jedan ili mali podskup postojećih ljudskih alela.

U ovom radu su analizirana 654 proteina iz različitih funkcionalnih grupa i 1469 alela MHC klase I i 517 alela MHC klase II. Rezultati su sledeći:

1) Broj epitopa u uređenim regionima je 3.68 (MHC klasa I) i 3.54 (MHC klasa II) puta veći nego u neuređenim. Tvrđenje važi u približno jednakom odnosu za obe klase MHC I i II. Ovaj rezultat je potvrda predpostavke, iznesene u [1] da su neuređeni delovi proteina slabiji antigeni.
2) Broj prepoznatih epitopa za MHC klasu II je 3.36 puta veći nego za MHC klasu I i pored činjenice da MHC klasa II ima skoro 3 puta manje alela.
3) Aleli MHC klase I koji prepoznaju najviše epitopa, za proteine iz DisProt baze, PDB baze i kancer – testis antigene su jednako dominantni u uređenoj i neuređenoj strukturi proteina. Izdvojeni najdominantniji aleli se poklapaju sa najčešćim ljudskim alelima.

4) Aleli MHC klase I koji prepoznaju najviše epitopa za bakterijske proteine čine nepreklapajući skup alela sa navedenim u rezultatu 3), to aleli HLA_B grupe. Ne zavise od strukture proteina, (uređeni / neuređeni delovi u proteinu).
5) Aleli MHC klase II koji prepoznaju najviše epitopa su isti za sve funkcionalne grupe analiziranih proteina uključujući i bakterijske. Najbrojniji su u svim strukturama proteina i poklapaju se sa najčešćim ljudskim alelima.
6) Tehnikom pravila pridruživanja utvrđen je interval hidrofobnosti za epitope u neuređenim regionima, kao i aleli koji su slični, jer u preko 50% slučajeva prepoznaju iste epitope tzv. promiskuitetne epitope.

Uspostavljanjem korelacije između neuređenih / uređenih delova proteina i antigenih regiona (epitopa) koji se vezuju za molekule HLA-I i HLA-II klase je:

- utvrđen odnos slabih i jakih epitopa i njihova zastupljenost u neuređenim / uređenim delovima proteina.
- uspostavljena korelacije između neuređenih / uređenih delova proteina i antigenih regiona (epitopa) u hidrofilnim-hidrofobnim regionima.

Strategija razvijanja vakcina, zasnovanih na epitopima proteina, a u cilju pokrivanja što veće ljudske populacije je identifikovanje HLA alela koji su prisutni u što većem broju individua iz svih mogućih etničkih grupa i koji bi se vezivali bar za jedan peptid u vakcini. Iz tog razloga sledi da su peptidi koji se vezuju za više od jednog alela (tzv. promiskuitetni epitopi) od prevashodnog interesa. Aleli koji prepoznaju najviše epitopa MHC klase I za bakterijske proteine čine izdvojeni skup koji obuhvata isključivo alele HLA klase B. Iako je obrađeni uzorak bakterijskih proteina bio relativno mali, ovaj podatak bi mogao da bude u korelaciji sa saznanjima o HLA-B restrikciji nekih bakterijskih i virusnih antigena. Takođe je nađeno da bakterijski proteini koji ulaze u sastav vakcina imaju manje predviđenih MHC-vezujućih peptida nego kontrolni proteini iz skoro svih subcelularnih bakterijskih lokacija. Ovaj efekat je previše izražen da bi mogao biti posledica nepreciznosti softvera ili aminokiselinskog sastava epitopa, kao što je to ranije predloženo (npr. veći broj hidrofobnih amino kiselina u MHC-II epitopima). Predpostavlja se da su patogeni organizmi evoluirali pod uticajem imunološkog sistema domaćina tako da su u njihovim površinskim proteinima eliminisane mnoge sekvence koje mogu da izazovu odgovor adaptivnog imunološkog sistema domaćina [23].

6.1 **Dalji rad**

Postoji više podela proteina iz DisProt baze: prema strukturalno-funkcionalnim osobinama, prema strukturi, prema funkciji i prema organizmima iz kojih vode poreklo. Planirano je ispitivanje:

- da li važe utvrđena pravila i pojedinačno za sve definisane grupe.
- da li HLA_B restrikcija utvrđena za bakterijske proteine važi za sve bakterijkse proteine.

Za 654 analizirana proteina su pronađeni svi neuređeni regioni, izračunata je njihova dužina, broj i vrsta epitopa koje sadrže. U planu je:

- Klasifikovanje neuređenih regiona prema epitopima koje sadrže (jake, slabe), broju epitopa i dužini neuređenog regiona.
- Analiza položaja epitopa u neuređenim sekvencama (pri krajevima ili u sredini).
- Analiziranje prisustva amino kiselina koje ulaze u sastav epitopa na osnovu karakteristika (polarnost, šarža, hidropatija) u neuređenim i uređenim regionima proteina.

7 Literatura

[1] Carl, P.L., Temple, B.R.S., Cohen, P.L., *Most nuclear systemic autoantigens are extremely disordered proteins*: implications for the etiology of systemic autoimmunity, Arthritis Research & Therapy, 2005, 7, R1360-R1373

[2] Dis Prot, release 5.0, 2010, http://www.disprot.org.

[3] Pengfei Han, Xiuzhen Zhang, Raymond S. Norton and Zhi-Ping Feng: "*Large-scale prediction of long disordered regions in proteins using random forests", BMC Bioinformatics 2009 , 10, 8, 1-8*

[4] Sickmeier M., Hamilton J.A., LeGall T., Vacic V., Cortese M.S., Tantos A., Szabo B., Tompa P., Chen J., Uversky V.N., Obradović Z., Dunker A.K. "*DisProt: the Database of Disordered Proteins*". Nucleic Acids Res., 35 (database issue): D786-793.

[5] Ian H. Witten, Eibe Frank; „*Data Mining, Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, 2005 ,Secend Edition.

[6] Depil S, Moralès O, Castelli FA, Delhem N, François V, Georges B, Dufossé F, Morschhauser F, Hammer J, Maillère B, Auriault C, Pancré *„Determination of a HLA II promiscuous peptide cocktail as potential vaccine against EBV latency II malignancies*." V. Journalof immunotherapy, 2007, 30, 215 - 226

[7] Lin H. H., Zhang G. L., Tongchusak S., Reinherz E. L., Brusić V. "*Evaluation of MHC II peptide binding prediction servers: applications for vaccine research*"; BMC Bioinformatics, 2008, 9 (Suppl. 12) , 1-10

[8] 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.

[9] Tong J. C., Tan T. W., Ranganathan S.: "*Methods and protocols for prediction of immunogenic epitopes*"; Bioinformatics, 2006, 8, (2) 96 – 108.

[10] Brusić V., Flower D. R.: "*Bioinformatics tools for identifying T-cell epitopes*"; Drug Discovery Today: BIOSILICO Volume 2, 1, 1, 2004, 18-23.

[11] Hand D.J., Manila H., Smyth P. *"Principles of data mining"*. Cambridge (MA): The MIT Press, 2001

[12] Dunham M.H., *"Data mining introductory and advanced topics"*, Prentice Hall, 2003

[13] Bishop, Christopher, „*Pattern Recognition and Machine Learning"*, Springer, 2006

[14] Cheng Y., LeGall T., Oldfield C. J, Dunker A. K., Uversky V. N.: „*Abundance of Intrinsic Disorder in Protein Associated with Cardiovascular Disease*". Biochemistry. 2006, 45(35):10448-60.

[15] Uversky V. N.: „*TheMysterious Unfoldome: Structureless, Underappreciated, Yet Vital Part of Any Given Proteome*"; Journal of Biomedicine and Biotechnology Volume 2010 (2010), Article ID 568068.

[16] Campen A., Williams R.M., Brown C.J., Meng J.W., Uversky V.N., Dunker A.K.: "*TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder"***.** *Protein and Pept Lett* 2008, **15(9)**:956–963.

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

[18] Purcell, A. W., & Gorman, J.J., „*Immunoproteomics, Mollecular and cellular proteomics"*, 2004, 3, 193-208

[19] Yang, X. & Yu, X., „*An introduction to epitope prediction methods and software", Reviews in medical virology*, 2009, 19, 77-96

[20] Rothbard, J.B. & Taylor, W.R., „*A sequence pattern common to T cell epitopes"; The EMBO Journal, 1988, 7, 93-100*

[21] Prato, S., Fleming, J., Schmidt, G.,Corradin, G., Lopey, J.A., *„Cross-presentation of a human malaria CTL epitope is conformation dependent", Mollecilar immunology*, 2006, 43, 2031-2036

[22] Carmicle, S., Steede, N.K., Landry, S.J. : „*Antigen three-dimensional structure guides the processing and presentation of helper T-cell epitopes", Molecular immunology*, 2007, 44, 1159-1168

[23] Halling-Brown, M., Shaban, R., Framton, D., Sansom, C.E., et al, Proteins accessible to immune surveillance show significant T-cell epitope depletion: Implication for vaccine design, Mollecular immunology, 2009, 46, 2699-2705.

[24] Niketić, V., „*Principi strukture i aktivnosti proteina"*, Izdavač: Hemijski fakultet, Univerzitet u Beogradu, Beograd, 1995.

[25] Ohkuri, T. Wakita, D., Chamoto, K., Togashi, Y, Kitamura, H., Nishimura, T., „*Identification of novel helper epitopes of MAGE-A4 tumour antigen: useful tool for the propagation of Th1 cells*", British Journal of Cancer, 2009, 100, 1135-1143

[26] Abbas, A. K. & Lichtman, A.H., „*Cellular and molecular immunology"*, Fifth ed., Pub. Saunders, 2003.

[27] Powel, A.M. & Black, M.M., „*Epitope spreading: protection from pathogens, but propagation of autoimmunity?", Clinical and experimental dermatology*, 2001, 26, 427-433

[28] Old, L.J., „*Cancer/Testis (CT) antigens – a new link between gametogenesis and cancer", Cancer immunity*, 2001, 1, 1-7