Универзитет у Београду
Математички факултет



*- Мастер теза -*

# Примена истраживања података на одређивање карактеристика позиција епитопа у протеинима

Иван Грујичић

Ментор: др. Ненад Митић

Београд, октобар 2014

**Ментор:**

др. Ненад Митић
*Математички факултет*
*Универзитет у Београду*

**Чланови комисије:**

др. Саша Малков
*Математички факултет*
*Универзитет у Београду*

др. Мирјана Павловић
*Институт за општу и физичку хемију*
*Београд*

Датум одбране:

---

# Садржај

# 1 Увод

Термин "биоинформатика" подразумева примену рачунара у биологији и тесно је повезан са развојем структуралне биологије и настанком квантитативне биологије. Током напретка структуралне биологије рачунари се убрзано развијају и имају знатно већу процесорску моћ и капацитет складиштења података. Самим тим почиње већа примена рачунара у биологији за симулирање животних процеса, откривање правилности у молекулима живих бића и у разне друге сврхе на лакши, бржи и поузданији начин него раније.

Лускомбе мало формалније дефинише биоинформатику као концептуализацију биологије у погледу молекула (у смислу физичке хемије) и примене информационих техника (математике, рачунарских наука и статистике) ради разумевања и уређивања велике количине информација повезаних са тим молекулима [1]. Једна грана биоинформатике обухвата развој рачунарских алата за структуралну, секвенцијалну и функционалну анализу биолошких полимера (нуклеинских киселина или протеина), који се састоје од ниски (ланаца или секвенци) молекула мономера (нуклеотида, односно аминокиселина), као и образовање биолошких база података. Друга грана је примена наведених алата.

У овом раду је представљена примена једне методе истраживања података, методе правила придруживања, у откривању карактеристика позиција и дужина антигених региона у протеинима. Одређена експериментална истраживања су показала да постоји однос између структуре протеина и одређених људских болести као што је на пример канцер и болести повезане са процесима ћелијског сигнализирања и регулације [2], као и између структурних и антигених региона протеина [3], [4]. Овај рад би требало да прошири сазнања о расподели експериментално утврђених и предвиђених антигених детерминанти (епитопа) у уређеној и неуређеној структури протеина.

## 1.1 Биолошки полимери

Дезоксирибонуклеинске киселине (скр. ДНК), рибонуклеинске киселине (скр. РНК) и протеини су основне класе ланчаних биолошких полимера. Ген се може дефинисати као линеарни фрагмент ДНК секвенце или основна ДНК секвенца са специфичном функцијом, [5]. Целокупан скуп ДНК секвенци, односно скуп гена у ћелији организма се назива геном. Према функцији коју обављају гене можемо поделити на структурне и регулаторне. Структурни гени су гени који носе шифру за протеин, која се састоји од триплета нуклеотида ДНК који кодирају појединачне аминокиселине протеина. Ова шифра се преписује на информациону-РНК. У структурне гене спадају и они који се само преписују у РНК (транспортну-РНК и рибозомалну-РНК). Регулаторни гени се не преписују већ се за њих везују молекули који регулишу одређене процесе у организму. Протеини (или полипептиди) су полимери који се састоје од секвенце аминокиселина. Садрже комбинације 20 различитих аминокисели-

на повезаних пептидном везом. Аминокиселине су молекули који се састоје од амино и карбоксилне групе, везане за $\alpha$-$C$ атом и бочног радикала, и деле се на есенцијалне и неесенцијалне. Редослед аминокиселина у протеину одређује структуру/структуре протеина и од те структуре зависи његова функција у организму. Протеини међусобним интеракцијама или интеракцијама са другим типовима молекула могу да изграде различите сложене структуре, [6].

**Структура протеина.** Структура протеина се описује на четири нивоа што је последица специфичног везивања ланаца аминокиселина:

- Примарна структура представља линеарну секвенцу (редослед) аминокиселина у полипептидном ланцу протеина.
- Секундарна структура представља локалну просторну организацију атома полипептидне кичме, која је дефинисана водоничним везама између амино и карбоксилне групе суседних аминокиселина у секвенци аминокиселина у протеину (занемарују се везе између главног и бочног ланца, као и међусобне везе бочних ланаца). Секундарна структура је правилна (уређена) структура. Најчешће секундарне структуре су: алфа-завојнице, бета-плоче, окрети и петље.
- Терцијарна структура представља просторну (тродимензионалну, скр. 3Д) структуру протеина повезаног вишеструким везама.
- Кватернарна структура представља више повезаних полипептида (субјединица) у један већи протеин.

## 1.2 Уређени и неуређени протеини

У структурној биологији протеина је дуготрајно владала догма да је специфична функција једног протеина одређена његовом јединственом 3Д структуром. Иако је структура и функција полипептидног ланца одређена његовом аминокиселинском секвенцом, само известан број протеина је предодређен да образује јединствену структуру која је еволуирала тако да поседује јединствену биолошку функцију, [7]. Неки протеини се погрешно увијају, било спонтано или под утицајем различитих ћелијских и ванћелијских фактора, што се данас сматра за битан рани стадијум у настанку неких конформационих болести, као што су цистична фиброза и Алцхајмерова или Паркинсонова болест. Међутим, експериментални подаци добијени у *in vitro* физиолошким условима показују да велики број протеина има неуређену структуру (тј. нема дефинисану 3Д структуру). Ови протеини, познати као ИДП (енг. *intrinsically disordered proteins*) у потпуности или у појединим деловима немају дефинисану 3Д структуру, под горе наведеним експерименталним условима. У условима *in vivo* многи од ових протеина могу да имају вишеструке, важне биолошке функције. Нађено је да недостатак дефинисане (уређене) 3Д структуре може да представља

функционалну предност за ове протеине који им омогућава да интерреагују са различитим везивним партнерима (лигандима) као што су протеини, нуклеинске киселине, мали молекули или велике структуре попут мембрана.

Најзначајније експерименталне методе којима се испитује неуређеност протеина, односно његова 3Д структура су дифракциона кристалографија $X$-зрацима и нуклеарно магнетна резонантна спектроскопија. Мана ових метода је што захтевају много времена и имају велики број ограничења. Из тих разлога развијени су програми који омогућавају да се предвиде неуређени региони протеина, који се називају предиктори. Предиктори се могу поделити у две групе. Прву групу предиктора чине програми који раде предвиђање на основу физичко-хемијских особина аминокиселина у протеину (нпр. *iupred*). Другу групу предиктора чине програми који раде предвиђање на основу примена метода поравнања хомологних протеинских секвенци (на пример *disopred*).

Структуре које су до сада познате од уређене до најнеуређеније су: уређена структура (енг. *order*), топљива глобула (енг. *molten globule*), претопљива глобула (енг. *pre-molten globule*) и случајни навој (енг. *random coil*). Свака од ових структура може бити природно стање протеина и битна за неку биолошку функцију у организму. Протеини могу прелазити из једне структуре у другу и обратно након што дођу у интеракцију са другим макромолекулима или након промена у биохемијским процесима. Неуређени протеини имају већу површину од уређених протеина, конформациону флексибилност да се везују за више партнера, елементе молекуларног препознавања који прелазе у уврнуту структуру након везивања, позиције које се пост-транслаторно модификују и обично садрже кратке линеарне мотиве који су важни за интеракцију протеина са лигандима. Као што је горе поменуто, неуређени протеини могу да имају велики број функција, нпр. молекулско препознавање, везивање у мултипротеинске комплексе, модификација протеина, активирање ензима и животни век протеина. ИДП су често укључени у регулаторне/сигналне процесе у ћелији који захтевају високу специфичност и слаб афинитет везивања, што је условљено управо њиховом флексибилном структуром.

## 1.3 Имунолошки систем

Имунолошки систем чине организована ткива која препознају патогене и елиминишу их из организма. Имунолошки систем у случају инфекције домаћина реагује имунолошким одговором који може бити урођени или стечени. Урођени имунолошки одговор није специфичан, нема имунолошку меморију и чини први бедем одбране домаћина. Стечени имунолошки одговор је специфичан за одређени антиген и има имунолошку меморију. Антиген (енг. *antibody generator*) је молекул којег препознаје имунолошки систем домаћина. Док се фрагмент антигена који се везује за одговарајуће рецепторе на антиген везујућим ћелијама имунолошког система назива епитоп. Стечени имунитет се дели на хуморални и ћелијски.

Главне ћелије имунолошког система су лимфоцити, антиген презентујуће ћелије и ефекторне ћелије. Лимфоцити имају способност да препознају и реагују на антиген. Постоје различите врсте лимфоцита, то су Б-лимфоцити и Т-лимфоцити. Б-лимфоцити (Б-ћелије) су главни носиоци хуморалног стеченог имунитета, који се налази у телесним течностима антитела. Углавном су усмерени на нелинеарне епитопе који се налазе на површини молекула. За препознавање антигена код Б-лимфоцита задужени су Б-ћелијски рецептори (енг. *B cell receptor*, скр. *BCR*). Т-лимфоцити су главни носиоци ћелијског стеченог имунитета, који је усмерен углавном на линеарне епитопе антигена. За препознавање антигена код Т-лимфоцита задужени су Т-ћелијски рецептори (енг. *T cell receptor*, скр. *TCR*). И *BCR* и *TCR* се налазе на мембранама својих лимфоцита.

Т-лимфоцити се могу поделити на две гране. Прва коју чине *Th*-лимфоцити (енг. *T helper* или *CD4*) и *Tr*-лимфоцити (енг. *T regulatory*) су део ћелијског имунитета који има задатак да регулише адаптивни и урођени имунитет, и одлучује који тип имунолошког одговора ће тело произвести на одређено антитело. Другу грану чине *Tc*-лимфоцити (енг. *Cytotoxic T*, *Tc* или *CD8*). Ако вирус инфицира ћелију вирусни пептиди (епитопи) омогућавају *Tc*-лимфоцитима да препознају и убију инфицирану ћелију.

Имунолошки одговор се састоји из три фазе: препознавање антигена, активације лимфоцита и елиминације антигена. У зависности од препознатог антигена активира се одређени стечени или урођени имунолошки одговор. Стечени имунолошки одговор излаже епитопе протеина који потичу од антигена и сопствене протеине ћелијама имунолошког система. Епитопи из интактних протеина преко протеоличких механизама који се одвијају у специјализованим органелама антиген презентујућих ћелија, се даље преносе на површину ћелија заједно са протеинима главног хистокомпатибилног комплекса организма да би их препознале ћелије имунолошког система. Молекули главног хистокомпатибилног комплекса (енг. *Major Histocompatibility Complex*, скр. *MHC*) су фамилије гена, и састоје се од две подкласе *MHC I* и *MHC II*. Код човека ова фамилија гена се назива антигени људских леукоцита (енг. *humman leukocyte antigens*, скр. *HLA*), која има две подкласе *HLA I* и *HLA II*. Поред тога постоје и различити типови гена подкласа *HLA I* и *HLA II*. Типови гена класе *HLA I* су *HLA-A*, *HLA-B*, *HLA-C*, *HLA-E*, *HLA-G*, док типови гена класе *HLA II* су *HLA-DP*, *HLA-DQ*, *HLA-DR*. Ови молекули имају важну улогу у имунолошком систему.

## 1.4 Програми за предвиђање уређених и неуређених региона у протеину

Програми који су коришћени у овом раду за предвиђање уређених и неуређених региона у протеинима су: *anchor*, *disembl* прецизније његове верзије *disembl-hotloops* и *disembl-rem465*, *isunstruct*, *iupred* прецизније његове верзије *iupred-short* и *iupred-*

*long*, *ond*, *ronn* и *vsl2*. Следи кратак опис наведених програма.

Програм *anchor* предвиђа везујуће регионе протеина у неуређеним протеинима. Заснован је на приступу процене парова енергије (то је уједно основа и за *iupred*). Предвиђање *anchor*-а комбинује општу неуређену тенденцију са осетљивошћу структурне средине, предвиђајући тако неуређене везујуће регионе без било какве информације о партнерима протеина. Као улаз неопходна је једна аминокиселинска секвенца, али прихватљива је и листа мотива наведена као регуларан израз са или без њихових имена. Излаз овог програма је вероватноћа која указује на то колика је шанса да остатак буде део неуређеног везујућег региона дуж сваке позиције у секвенци.

Програм *disembl* има више верзија, између осталог *disembl-hotloops* и *disembl-rem465*. Верзије се разликују међусобно по критеријуму одређивања да ли је регион неуређен. Програм *disembl-hotloops* прави скуп података за разликовање уређених и неуређених врућих петљи (енг. *hot loops*) [8]. Програм *disembl-rem465* прави скуп података за предвиђање недостајућих координата. Цео скуп података се затим дели на скуп тест података и скуп тренинг података, а сваки од ова два скупа се дели на пет партиција применом технике крос-валидације. На тако припремљене податке примењује се техника методе класификације, вештачке неуронске мреже.

Програм *isunstruct* омогућава предвиђање неуређених региона у протеину, очекујући на улазу само задату ниску. Заснован је на моделу који апроксимира *Ising* модел [9] у коме је интеракција између суседа замењена са казном при промени стања. Сваки регион може бити означен као уређен или неуређен.

Програм *iupred* се заснива на физичком објашњењу природе уређених/неуређених протеина [10]. Као улаз узима једну аминокиселину из секвенце и рачуна парове енергетског профила дуж секвенце. Након тога се вредности енергије пресликавају у вредности од 0 (комплетно уређен протеин) до 1 (комплетно неуређен протеин). Вредности преко 0.5 посматра као неуређене протеине. Пре извршавања програма могу се подесити параметри у зависности да ли се траже дугачки неуређени региони (*iupred-long*), кратки неуређени региони (*iupred-short*) или структурирани домени.

Програм *ond* се заснива на *CRF* (енг. *Conditional Random Fields*) методи која тачно предвиђа транзиције стрктура и неуређених региона протеина. Представља селективну надгледану методу машинског учења [11]. Ова метода тренинг скуп изводи из кристалне структуре високе резолуције. Профили се кодирају ниском аминокиселина дужине 9.

Програм *ronn* (енг. *regional order neural network*) за одређивање неуређених региона заснива се на претпоставци да ако два протеина имају сличну биолошку функцију онда примарне секвенце су такође сличне, што се лако може проверити техникама за поравнање секвенци које користе матрицу мутације да израчунају сличност секвенци. Секвенце се пореде са серијом секвенци познатог завијеног стања (уређеног, неуређеног

или њихове комбинације) и скор поравнања ових секвенци се користи да класификује сваку секвенцу као уређену или неуређену користећи довољно истриниране неуронске мреже [12].

Програм *vsl2* се заснива на физичко-хемијским особинама аминокиселина које улазе у састав протеина. Неки од улазних атрибута су хидрофобност, одређена комбинација аминокиселина, итд. Овај програм чине два подпрограма, *vsl2-m1* који проналази краће регионе (до 30 аминокиселина) и *vsl2-m2* који проналази дуже регионе (преко 30 аминокиселина). Као трећа компонента овог предиктора је мета-предиктор који има улогу да комбинује излазе из претходна два предиктора и даје коначан резултат.

## 1.5 Програми за предвиђање епитопа у протеину

Сада су кратко објашњени програми који су коришћени за предвиђање везивања епитопа са *HLA I* и *HLA II* класом. Програми који су коришћени за то су: *netmhc-3.0c*, *netmhcpan-2.8a*, *netmhcii-2.2* и *netmhciipan-2.0b*. Основна разлика између програма *netmhc* и *netmhcii* је та што први програм предвиђа епитопе за протеине *MHC I* класе, док други програм ради предвиђање за протеине *MHC II* класе. Исто важи и за програме *netmhcpan* и *netmhciipan*. Следи кратак опис наведених програма.

Програм *netmhc* генерише високу прецизност предвиђања *MHC* пептидне кичме. Заснива се на тренирању методе применом вештачких неуронских мрежа и матрицом скора специфичне позиције (енг. *position-specific scoring matrices*) над људским и не-људским алелима. На улазу програм захтева фаста датотеку протеина и затим предвиђа афинитет пептида дужине 8, 10 и 11. Ако би било потребно да се ради предвиђање за пептиде дужине 9 тада би требало додати неку аминокиселину ако је у питању пептид дужине 8, иначе потребно је обрисати једну или две аминокиселине зависно од тога да ли је у питању пептид дужине 10 или 11. Програм *netmhcii* је заснован на техници *NN*-поравнања (енг. *neural network-based alignment*) методе вештачких неуронских мрежа за истовремено проналажење језгра и афинитета везивања. *NN*-поравнање користи нови алгоритам који омогућава исправку пристрасности из тренинг података. Метод је оцењен на четири независна велика скупа података који покривају 14 људских алела *MHC II* класе.

Програм *netmhcpan* је заснован на методи вештачких неуронских мрежа. Трениран је на великом броју квантитативних *MHC* везујућих података. Генерише квантитативно предвиђање афинитета било којих интеракција између пептида и молекула *MHC I* класе. Предвиђање је могуће над људским и не-људским алелима. Као улаз програм захтева фаста датотеку протеина, а затим анализира све могуће пептиде дужине 9. За друге дужине пептида предвиђање се добија апроксимацијом. Програм *netmhciipan* је исто заснован на методи вештачких неуронских мрежа, али је могуће да ради предвиђање за пептиде дужине 9-15. Оно што је важно за оба програма јесте што имају високу тачност предвиђања епитопа за различите групе протеина.

## 1.6 *IEDB* - база података експерименталних епитопа

Подаци који су коришћени у овом раду су преузети из *IEDB* (енг. *Immune Epitope Database*, скр. *IEDB*) базе података експерименталних епитопа. Главни задатак *IEDB*-а је прикупљање и складиштење имунолошких информација попут Б и Т ћелијских заразних патогена, експерименталних и сопствених антигена организма домаћина. Поред тога *IEDB* има задатак да развија методе за предвиђање и моделовање имунолошког одговора, као и да помаже у откривању вакцина и дијагноза разних болести. У *IEDB*-у складиштен је велики број експерименталних података о Б и Т ћелијским епитопима. Представљено истраживање је усмерено ка Т-ћелијским епитопима.

*IEDB* је *MySql* база података, чија је схема и експорт свих података из ње јавно доступно на [13]. Верзија *IEDB*-а која је преузета за ово истраживање је 2.2. За све Т-ћелијске епитопе из ове базе могуће је утврдити из којих су протеина, њихове почетне и завршне позиције у протеинима, организме из којих потичу и друге важне информације. У ово истраживање укључени су они епитопи за које је утврђено да протеини којима припадају имају више од пет експериментално утврђених епитопа и за које су успешно добијени подаци предвиђања предиктора описаних у подпоглављу 1.4.

# 2 Опис задатка

Овај рад представља покушај откривања нових сазнања о карактеристикама позиција и дужина предвиђених и експерименталних епитопа у предвиђеним уређеним и неуређеним регионима протеина помоћу методе истраживања података. Истраживање је урађено на групи протеина који припадају Т-лимфоцитима и имају више од пет епитопа. За главну методу истраживања одабрана је метода правила придруживања.

Главни задатак рада се заправо може поделити на два подзадатка. Први подзадатак је формирање консензуса уређених/неуређених региона за све протеине на основу добијених резултата свих предиктора за предвиђање уређених/неуређених региона. Консензус се формира на основу резултата осам предиктора који су коришћени за предвиђање уређених/неуређених региона посматрајући појединачне позиције сваког протеина засебно. Ако су за неку позицију сви предиктори рекли да припада уређеном региону онда се она означава као таква. Исто важи и ако су за неку позицију сви предиктори предвидели да припада неуређеном региону. Док за позиције које су неки предиктори означили да припадају уређеном, а неки неуређеном региону, означава се да припада мешаном региону. Затим се из тако припремљених података извлаче неке занимљиве информације за које није потребна примена правила придруживања и представљене су сликама или табелама.

Други подзадатак је да се на основу претходно припремљених података (првенствено се мисли на консензус) пронађу сви епитопи који цели припадају уређеном, неуређеном или мешаном региону, као и сви епитопи који се налазе на прелазима региона. Да би се на такве епитопе применила метода правила придруживања и пронашла нека занимљива правила која укључују позиције и дужине предвиђених епитопа и њихове карактеристике ком региону припадају или какве су им карактеристике везивања.

# 3 Опис примењене методе истраживања података

Као први узрок развоја истраживања података јесте драстичан пад цена компоненти за складиштење података. Док је други узрок драстичан пораст количине података, али и слаба корист од њих јер из те гомиле података на први поглед није могло ништа значајно да се закључи. Други узрок је на неки начин последица првог узрока.

Истраживање података је кључни део процеса откривања знања из велике количине података. Фазе процеса откривања знања су следеће:

1. Постављање циљева анализе и разумевање података.
2. Издвајање битних података из целог скупа које желимо даље да обрађујемо.
3. Обрада и чишћење података, проналажење разних специфичности података (нпр. екстремних вредности) и сређивање недостајућих вредности.
4. Трансформација података у облик погодан за остваривање циљева анализе.
5. Одређивање методе и технике истраживања података, и утврђивање потребних параметара.
6. Примена алгоритама истраживања података на претходно сређене податке.
7. Визуелизација и тумачење добијених резултата.
8. Оцењивање и примена добијених резултата у пракси.

Откривање знања је мултидисциплинарна област која у средиште свог процеса ставља истраживање података.

## 3.1 Појам истраживања података

Истраживање података представља процес анализе прикупљених података да би се из њих извукле корисне информације. Под подацима се подразумевају јако велики скупови података који никако не би могли ручно да се обрађују. Проналажење скупова правила података може да помогне да се предвиди неко понашање или неки будући тренд. У контексту неког великог ланца супермаркета истраживање података може да помогне да се проучи понашање купаца. На пример које артикле купци најчешће купују заједно или да се уради подела купаца по неком критеријуму, итд. Поред тога методе истраживања података се користе и на многим другим местима које у средиште модела свог пословања стављају клијента да би своје пословање учинили ефикаснијим. Све ово иде у прилог томе да истраживање података има велики потенцијал да постаје један од најбитнијих фактора у информатичком сектору.

Задаци истраживања података се деле у две групе:

- **Предиктивни задаци** чији је циљ предвиђање вредности конкретног (циљног) атрибута на основу вредности осталих (независних) атрибута.
- **Дескриптивни задаци** чији је циљ пронаћи обрасце (корелације, трендове, кластере, трајекторије, аномалије) који описују односе између података.

Код предиктивног моделовања суштина је да се модел за циљани атрибут гради као функција независних атрибута. У овај начин моделовања спадају методе *класификације* и *регресије*. Разлика између њих је та што класификација користи дискретне циљне атрибуте, док регресија користи континуалне циљне атрибуте. Заједничко им је да се у оба случаја прави линеарни модел који минимизује грешку између вредности предикције и правих вредности циљног атрибута. Више о регресији и њеним техникама се може прочитати у [14], а о класификацији у [15].

У дескриптивно моделовање спадају методе правила придруживања, кластер анализе и откривање аномалија. Циљ методе правила придруживања је да пронађе обрасце који су у форми правила импликације из исказне логике, о чему ће више бити речи у 3.3. Један пример примене правила придруживања је проналажење групе гена који имају сличну функционалност, што је и тема овог рада. Кластер анализа подразумева проналажења група података на основу неке задате мере сличности између података на које се примењује. Кластер анализу можемо да применимо када желимо на пример да групишемо текстове, тако што за сваки текст одредимо кључне речи које се појављују у њему и на основу тих кључних речи израчунамо сличност текстова да би их поделили у групе. Док је задатак откривања аномалија проналажење података који су значајно различити (прилично одскачу) од осталих података. Откривање аномалије је јако важно када се на пример посматрају подаци неког корисника кредитне картице. Ако постоји неки податак који јако одскаче из профила понашања корисника кредитне картице, банка може да посумња да је дошло до злоупотребе и да контактира клијента.

## 3.2 Припрема података

Важан корак пре примене неке методе истраживања података је припрема података у погодан облик за даљу анализу. Овде је представљено на који начин су подаци припремљени и смештени у базу података.

Као што је речено у 1.6, подаци коришћени у овом истраживању су преузети из *IEDB* базе података која представља базу података експерименталних епитопа. Циљ припреме података је био да се на локалном рачунару направи идентична база података како би се олакшали наредни кораци припреме, као и само истраживање. Подаци који су преузети из *IEDB*-а су били припремљени за рад са *MySql* базом података, док је жеља била да радимо са *IBM DB2* базом података. Дакле, било је потребно да се све дефиниције табела и подаци трансформишу у облик за рад са

*IBM DB2* базом података. У ту сврху написан је програм *mySqlToDb2* за Виндовс (енг. Windows) и Линукс (енг. Linux) платформу у програмском језику Јава, који иде уз електронску верзију овог рада.

Програм *mySqlToDb2* на улазу очекује датотеку која се добије експортом *MySql* базе података. Као резултат свог рада даје одвојене *.sql* датотеке за прављење сваке табеле посебно, као и *.load* датотеке са подацима за сваку табелу. Програм се састоји из три дела. Први део програма дели улазну датотеку на више датотека тако да у свакој датотеци буде дефиниција и подаци само једне табеле. Други део програма дели датотеке добијене као резултат првог дела програма на две датотеке тако да се у једној датотеци налази дефиниција табеле, а у другој датотеци подаци те табеле. Трећи део програма од датотека које садрже податке табела који су добијени као резултат рада другог дела програма прави *.load* датотеке. Детаљније упутство у вези програма се може видети при преузимању програма.

Након написаног *mySqlToDb2* програма стекли су се услови да се направи *IEDB* база података на локалном рачунару. Наредни корак је био филтрирање података, односно издвајање података који су нам потребни. Као што је већ наведено у опису задатка потребно је издвојити Т-ћелијске епитопе *HLA I* и *HLA II* класе, за које у направљеној бази података постоји више од пет епитопа. Сви потребни подаци који су испуњавали постављене услове су издвојени упитом и смештени су у табелу *EPITOPE_PROTEIN*, чија дефиниција се може видети на слици 1.

| Schema | Name | Table | Type |
|---|---|---|---|
| IVAN | PROTEIN_ID | EPITOPE_PROTEIN | VARCHAR(20) |
| IVAN | PID | EPITOPE_PROTEIN | INTEGER |
| IVAN | EPITOPE_ID | EPITOPE_PROTEIN | INTEGER |
| IVAN | ORGANISAM_NAME | EPITOPE_PROTEIN | VARCHAR(150) |
| IVAN | SEQUENCE | EPITOPE_PROTEIN | VARCHAR(535) |
| IVAN | ALLELA | EPITOPE_PROTEIN | VARCHAR(85) |
| IVAN | IMM_TYPE | EPITOPE_PROTEIN | VARCHAR(50) |
| IVAN | START_POS | EPITOPE_PROTEIN | INTEGER |
| IVAN | END_POS | EPITOPE_PROTEIN | INTEGER |
| IVAN | LENGTH | EPITOPE_PROTEIN | INTEGER |
| IVAN | EPITOPE_COUNT | EPITOPE_PROTEIN | SMALLINT |
| IVAN | VALUE | EPITOPE_PROTEIN | VARCHAR(50) |
| IVAN | TAXONOMY_1 | EPITOPE_PROTEIN | VARCHAR(20) |
| IVAN | UREDJENOST | EPITOPE_PROTEIN | VARCHAR(1) |

Слика 1: Дефиниција табеле *EPITOPE_PROTEIN*

Колона *PROTEIN_ID* представља идентификатор протеина (антигена, односно имуногена) у *IEDB* бази података, колона *PID* представља *GI* протеина (*GI* протеина

представља низ цифара које су редом додељене свакој секвенци која је обрађена у *NCBI* (енг. *National Center for Biotechnology Information*) бази података, и нема никакве везе са идентификатором протеина), колона *EPITOPE_ID* представља идентификатор епитопа, колона *ORGANISAM_NAME* представља име организма из којег потиче антиген, колона *SEQUENCE* представља секвенцу аминокиселина, колона *ALLELA* представља да ли је у питању *HLA I* или *HLA II* алела, колона *IMM_TYPE* представља тип имуногена, колона *START_POS* представља почетну позицију епитопа у протеину, колона *END_POS* представља завршну позицију епитопа у протеину, колона *LENGTH* представља дужину епитопа, колона *EPITOPE_COUNT* представља број епитопа протеина, колона *VALUE* означава да ли је епитоп позитиван или негативан, колона *TAXONOMY_1* означава којој групи припада организам чији је епитоп (вирусима, еукаријама или бактеријама) и колона *UREDJENOST* означава ком региону припада епитоп (уређен, неуређен, мешани, прелазни) што је одређено на основу вредности из табеле *KONSENZUS*.

Дакле, у табели *EPITOPE_PROTEIN* су смештени сви експериментални подаци (протеини са припадајућим епитопима) који су важни за даље истраживање. Након тога за све протеине из табеле *EPITOPE_PROTEIN* потребно је преузети фаста датотеку из *NCBI*-а како би се урадило предвиђање уређених/неуређених региона и предвиђање епитопа. Програми који се користе за наведена предвиђања су описани у 1.4 и 1.5. Сада следи опис табела у које су смештени подаци који су добијени као резултат рада тих програма.

Табела *DISORDER* са слике 2 садржи податке о уређеним и неуређеним регионима протеина. Колоне *PID* и *PROTEIN_ID* представљају исто што и у табели *EPITOPE_PROTEIN*. Колона *START_POS* представља почетну позицију региона, колона *END_POS* представља завршну позицију региона, колона *ORDER_LEVEL* представља уређеност региона (уређен или неуређен) и колона *PREDICTOR* представља који програм је коришћен за предвиђање.

| Schema | Name | Table | Type |
|---|---|---|---|
| IVAN | PID | DISORDER | INTEGER |
| IVAN | PROTEIN_ID | DISORDER | VARCHAR(15) |
| IVAN | START_POS | DISORDER | SMALLINT |
| IVAN | END_POS | DISORDER | SMALLINT |
| IVAN | ORDER_LEVEL | DISORDER | CHAR(1) |
| IVAN | PREDICTOR | DISORDER | VARCHAR(19) |

Слика 2: Дефиниција табеле *DISORDER*

Табела *DISORDER_NUMERIC* са слике 3 садржи податке о уређеним и неуређеним регионима протеина за сваку позицију. Колоне *PID*, *PROTEIN_ID*, *ORDER_LEVEL* и *PREDICTOR* представљају исто што и у табели *DISORDER*. Колона *POSITION*

представља позицију у аминокиселини у протеину, колона *AA* представља која аминокиселина је у питању и колона *VALUE* представља нумеричку вредност (праг) на основу које предиктор одређује да ли је та позиција у уређеном или неуређеном региону.

| Schema | Name | Table | Type |
|---|---|---|---|
| IVAN | PID | DISORDER_NUMERIC | INTEGER |
| IVAN | PROTEIN_ID | DISORDER_NUMERIC | VARCHAR(15) |
| IVAN | POSITION | DISORDER_NUMERIC | SMALLINT |
| IVAN | AA | DISORDER_NUMERIC | CHAR(1) |
| IVAN | VALUE | DISORDER_NUMERIC | DECIMAL(6 , 5) |
| IVAN | ORDER_LEVEL | DISORDER_NUMERIC | CHAR(1) |
| IVAN | PREDICTOR | DISORDER_NUMERIC | VARCHAR(17) |

Слика 3: Дефиниција табеле *DISORDER_NUMERIC*

На основу вредности из табеле *DISORDER_NUMERIC* израчунат је консензус за сваку позицију сваког протеина и смештен у табелу *KONSENZUS*, али тако да су региони представљени у интервалима, а не по позицијама. Дефиниција табеле *KONSENZUS* дата је на слици 4.

| Schema | Name | Table | Type |
|---|---|---|---|
| IVAN | PID | KONSENZUS | INTEGER |
| IVAN | START_POS | KONSENZUS | SMALLINT |
| IVAN | LENGTH | KONSENZUS | SMALLINT |
| IVAN | UREDJENOST | KONSENZUS | VARCHAR(1) |

Слика 4: Дефиниција табеле *KONSENZUS*

Табела *EPITOPE* са слике 5 садржи податке о карактеристикама епитопа у протеинима. Колоне *PID*, PROTEIN_ID, *ALLELE*, *TAXONOMY_1* и *UREDJENOST* представљају исто што и у табели *EPITOPE_PROTEIN*. Колона *POSITION* представља почетну позицију епитопа, колона *EPITOPE* представља секвенцу епитопа, колона *OFFSET_CORE* представља дужину језгра епитопа, колона *CORE* представља секвенцу језгра епитопа, колона *AFF_LOG* представља логаритмовану вредност афинитета везивања епитопа, колона *AFF* представља афинитет везивања епитопа, колона *RANK* представља тачност афинитета везивања епитопа, колона *BONDING* представља јачину везивања епитопа (јако везујућ или слабо везујућ), колона *PROGRAM* представља који програм је коришћен за предвиђање епитопа, колона *EP_LENGTH* представља дужину епитопа и колона *TAXONOMY_2* представља подгрупу организма којој епитоп припада.

| Schema | Name | Table | Type |
|---|---|---|---|
| IVAN | PID | EPITOPE | INTEGER |
| IVAN | PROTEIN_ID | EPITOPE | VARCHAR(15) |
| IVAN | POSITION | EPITOPE | SMALLINT |
| IVAN | EPITOPE | EPITOPE | VARCHAR(20) |
| IVAN | OFFSET_CORE | EPITOPE | SMALLINT |
| IVAN | CORE | EPITOPE | VARCHAR(20) |
| IVAN | AFF_LOG | EPITOPE | DECIMAL(8 , 3) |
| IVAN | AFF | EPITOPE | DECIMAL(8 , 2) |
| IVAN | RANK | EPITOPE | DECIMAL(8 , 2) |
| IVAN | BONDING | EPITOPE | CHAR(2) |
| IVAN | PROGRAM | EPITOPE | VARCHAR(20) |
| IVAN | ALLELE | EPITOPE | VARCHAR(30) |
| IVAN | EP_LENGTH | EPITOPE | SMALLINT |
| IVAN | TAXONOMY_1 | EPITOPE | VARCHAR(20) |
| IVAN | TAXONOMY_2 | EPITOPE | VARCHAR(35) |
| IVAN | UREDJENOST | EPITOPE | VARCHAR(1) |

Слика 5: Дефиниција табеле *EPITOPE*

Након свих наведених корака подаци су припремљени и спремни за даље истраживање. Као што је речено у опису задатка прво ће дијаграмима бити представљене значајне информације које се виде из њих, а након тога ће над овим подацима бити примењена метода правила придруживања. Како су подаци припремљени у *IBM DB2* бази података, алат коришћен за примену методе правила придруживања је *IBM Infosphere Warehouse*.

## 3.3 Метода правила придруживања

Како је као главна метода овог истраживања коришћена метода правила придруживања она ће бити мало детаљније описана у овом подпоглављу. Метода правила придруживања је корисна за откривање значајних скривених односа у великом скупу података. Откривени односи се могу представити у форми правила придруживања или скупом фреквентних ставки. На пример ако су извршене следеће трансакције у неком супермаркету:

1. {хлеб,млеко}
2. {хлеб,пелене,пиво,јаја}
3. {млеко,пелене,пиво,сок}
4. {хлеб,млеко,пелене,пиво}
5. {хлеб,млеко,пелене,сок},

може се извући правило {пелене} → {пиво}. Ово правило говори да постоји јака веза између продаје пелена и пива, јер већина купаца која купује пелене купује и пиво. На овај начин продавац може да прати трендове продаје и да побољшава могућности своје продаје.

### 3.3.1 Терминологија

Овде су објашњени основни термини који се најчешће јављају када је у питању метода правила придруживања. За почетак се претпоставља да је задата бинарна репрезентација података транскације, 0 ако ставка није присутна у трансакцији и 1 ако јесте. Како је присуство ставке у трансакцији много важније, онда ће подаци бити представљени асиметричном бинарном репрезентацијом. Та репрезентација је једноставнија јер не узима у обзир неке друге аспекте као на пример цену, количину и слично.

Нека је скуп свих ставки $S = \{s_1, s_2, ..., s_D\}$ и скуп свих трансакција $T = \{t_1, t_2, ..., t_N\}$, тако да свака трансакција $t_i \in T$ садржи подскуп ставки из $S$. Скуп од нула или више ставки се назива скуп ставки, ако он има $k$ ставки можемо га назвати и $k$-скуп ставки. Важно својство скупа ставки $X$ је бројач подршке који представља број трансакција које садрже тај скуп ставки и означава се са $\sigma(X)$. На пример, нека трансакција $t_j$ садржи скуп ставки $X$, то значи да је $X \subset t_j$, а тада је $\sigma(X) = |\{t_j | X \subseteq t_j, t_j \in T\}|$, где $|\cdot|$ означава кардиналност скупа.

Правило придруживања је правило облика $X \to Y$, где важи $X \cap Y = \emptyset$, где су $X$,$Y$ скупови ставки. Део $X$ се назива тело (енг. *body*) правила, а део $Y$ се назива глава (енг. *head*) правила. Процена правила придруживања се ради помоћу мера подршке и поузданости. Подршка представља заправо колико често је неко правило применљиво на задати скуп ставки и формално се може дефинисати као $S(X \to Y) = \frac{\sigma(X \cup Y)}{N}$, где је $N$ укупан број ставки. Подршка је важна мера јер правила са малом подршком нису занимљива за даље разматрање. Док је поузданост мера која говори да ако неко правило $X \to Y$ има већу поузданост, већа је шанса да се $Y$ појави у истој трансакцији када и $X$. Дакле, поузданост представља фреквенцију појављивања ставки из $Y$ у трансакцији која садржи $X$, а она се дефинише на следећи начин $C(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$.

На основу претходног може се дефинисати правило придруживања као задатак у коме је за задати скуп трансакција потребно пронаћи таква правила која задовољавају неке минималне унапред задате вредности за подршку ($minsup$) и поузданост ($minconf$). Односно потребно је пронаћи скуп правила $P = \{p \mid S(p) \geq minsup \wedge C(p) \geq minconf\}$.

Прва идеја како пронаћи таква правила је примена грубе силе, да се за свако могуће правило рачуна подршка и поузданост, али то је јако скупо. Ако је задат скуп са $d$ ставки требало би да се рачуна подршка и поузданост за $3^d - 2^{d+1} + 1$

правила, тако на пример ако је $d = 6$ тада имамо чак 602 правила за које је потребно рачунати подршку и поузданост. Некако је потребно да се смањи тако велики број правила, па се за почетак раздвоји рачунање подршке и поузданости. Након тога се да приметити да подршка правила зависи само од одговарајућег скупа ставки, јер јасно је да на пример:

- $\{a, b\} \to \{c\}$
- $\{a, c\} \to \{b\}$
- $\{b, c\} \to \{a\}$
- $\{a\} \to \{b, c\}$
- $\{b\} \to \{a, c\}$
- $\{c\} \to \{a, b\}$

сва претходна правила имају исту подршку јер представљају исти скуп ставки $\{a, b, c\}$. Самим тим ако скуп ставки $\{a, b, c\}$ није чест може се одбацити рачунање подршке за сва наведена правила. Па се алгоритам за проналажење одговарајућег скупа правила може поделити на два дела:

- генерисање честих скупова ставки, који проналази све скупове ставки који задовољавају $minsup$
- генерисање правила, који проналази скупове ставки који задовољавају $minconf$ из скупова ставки добијених у претходном кораку.

Сложеност приступа грубе силе при рачунању подршке за сваки скуп ставки је $O(N \cdot M \cdot \omega)$, где је $N$ број трансакција, $M$ број кандидата и $\omega$ максимална дужина трансакције, што је у пракси много. Због тога се тежи да се смањи комплексност рачунања. Један начин је да се смањи број кандидата $M$ коришћењем Априори алгоритма или да се смањи број поређења $N \cdot M$ користећи неке ефикасне структуре података.

### 3.3.2 Априори алгоритам за генерисање честих скупова ставки

Један од начина да се при генерисању честих скупова ставки смањи комплексност рачунања као што је већ споменуто јесте Априори алгоритам за генерисање честих скупова ставки. Априори алгоритам за генерисање честих скупова ставки је један од најпопуларнијих алгоритама у области Истраживања података. Заснива се на особини метрике којом се мери подршка, да подршка скупа правила никада није већа од подршке његових подскупова. Односно, за $\forall X, Y$ важи $X \subseteq Y \Rightarrow S(X) \geq S(Y)$, где су $X, Y$ скупови ставки, ова особина се назива **анти-монотоност подршке**. Из претходног се може закључити да ако је скуп ставки чест, онда је сваки његов подскуп чест, односно ако скуп ставки није чест онда ни његови надскупови нису

чести. Овим принципом смањује се простор претраге.

Скица Априори алгоритам за генерисање честих скупова ставки:

$F_1 =$ `cesti skupovi stavki kardinalnosti 1;`
`for (`$k = 1; F_k \neq \emptyset; k++$`) do`
    $C_{k+1} = apriori\_gen(F_k)$
    `for (svaku transakciju` $t \in Y$ `) do`
        $C'_t = podskup(C_{k+1}, t)$
        `for (svaki kandidat` $c \in C'_t$ `) do`
            $c.count++$
        `end`
        $F_{k+1} = \{C \in C_{k+1} | c.count \geq minsup\}$
    `end`
`end`
`vrati` $\bigcup_k F_k$

Скуп $C_k$ представља скуп кандидата, док скуп $F_k$ представља скуп честих ставки величине $k$. Метода $apriori\_gen(F_k)$ која се позива заправо генерише нове потенцијалне кандидате. Више детаља о алгоритму се може прочитати у [16].

Приказани алгоритам је итеративан, са максималним бројем итерација $k_{max} + 1$, где је $k_{max}$ максимална величина честих скупова ставки. Да би се ефикасно генерисали кандидати у претходном алгоритму потребно је да се избегне генерисање непотребних кандидата. Да би се избегло генерисање непотребних кандидата мора да скуп кандидата буде комплетан и да се не генеришу исти кандидати више пута. Из тог разлога постоји више техника за генерисање кандидата које се користе у овом алгоритму:

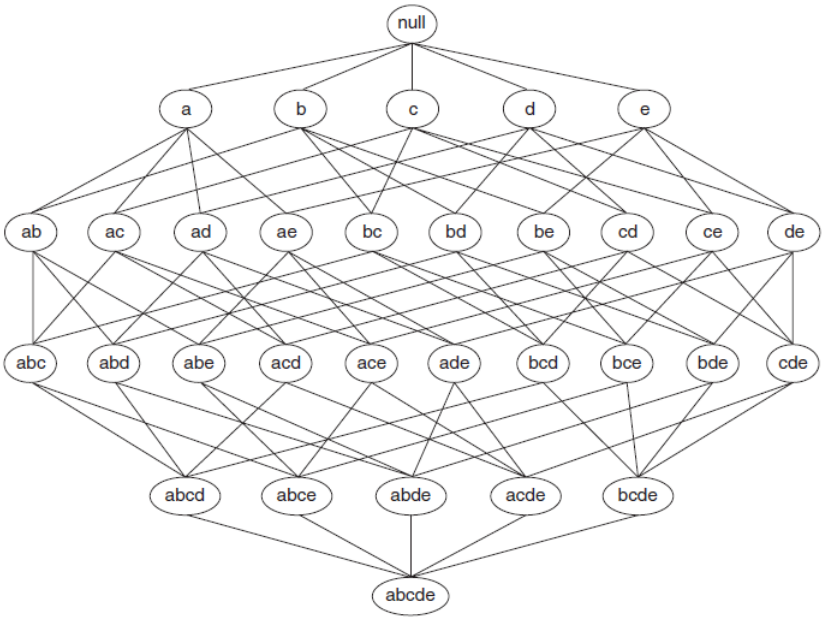
1. Примена грубе силе која разматра сваки подржани скуп дужине $k$ као потенцијалног кандидата и затим примењује брисање непотребних кандидата.

2. $F_{k-1} \times F_1$ техника која генерише кандидате тако што сваки чест скуп ставки величине $k-1$ проширује са осталим честим скуповима ставки. Потенцијалан проблем јесте појава дупликата што се рашава лексикографским сортирањем.

3. $F_{k-1} \times F_{k-1}$ техника која генерише кандидате спајањем парова $A, B$ честих скупова ставки за које важи $A = \{a_1, ..., a_{k-1}\}$ и $B = \{b_1, ..., b_{k-1}\}$ тако да је $a_i = b_i$ за $\forall i = \overline{1, k-2}$ и $a_{k-1} = b_{k-1}$.

Други начин да се повећа ефикасност рачунања честих скупова ставки јесте да се смањи број поређења. Број поређења се смањује тако што ће се уместо да се сваки кандидат тражи у свакој трансакцији користити напредне структуре података. То је могуће пребројавањем и коришћењем *префиксне структуре* [17] или *хеш стабла* [18].

**Алтернативне технике генерисања честих скупова ставки.** Да би се представила листа свих могућих ставки користи се структура решетке, слика 6. Како скуп са $k$ ставки може генерисати $2^k - 1$ честих скупова ставки и како $k$ може бити јако велико, можемо имати велики простор претраге. Једна група алтернативних техника за генерисање честих скупова ставки се заснива на алгоритмима који одређују начин обиласка решетке. Другу групу чине алгоритам ФП-раста (енг. FP-growth) [19] и алгоритам ФП-максимума (енг. FP-max) [20] који омогућавају ефикасно проналажење честих скупова ставки из своје структуре. Алгоритам ФП-раста користи структуру ФП-дрвета (енг. FP-tree), док алгоритам ФП-максимума комбинује структуре ФП-дрвета и ФП-низа (енг. FP-array). Обе структуре представљају компресивну репрезентацију података.




Слика 6: Пример структуре решетке

### 3.3.3 Априори алгоритам за генерисање правила

Након што је решен први део проблема који се бави генерисањем честих скупова ставки, остало је још да се реши други део проблема чији је задатак генерисање правила придруживања.

Сваки чест скуп ставки $Y$ од $k$ елемената може да произведе $2^k - 2$ правила придруживања (игноришемо правила облика $Y \to \emptyset$ и $\emptyset \to Y$). Правила се могу извести дељењем скупа ставки $Y$ на два непразна подскупа $X$ и $Y - X$, тако да правило $X \to Y - X$ задовољава $minconf$. Оно што се може приметити јесте да сва та правила задовољавају $minsup$ јер су генерисана из скупа честих ставки. За разлику од подршке, поузданост нема својство анти-монотоности. Што значи да за правило

$X \to Y$, поузданост тог правила може бити већа, мања или једнака поузданости неког другог правила $X' \to Y'$ за које важи $X' \subseteq X$ и $Y' \subseteq Y$. То је важно имати у виду због наредног закључка који омогућава одсецања у алгоритму. Ако правило $X \to Y - X$ има поузданост мању од $minconf$, тада и правило $X' \to Y - X'$ где је $X' \subseteq X$ има поузданост мању од $minconf$. Најпознатији алгоритам за генерисање правила придруживања јесте Априори алгоритам за генерисање правила.

Скица Априори алгоритма за генерисање правила:

```
for (svaki cesti skup stavki kardinalnosti k, f_k, k ≥ 2) do
    H_1 = {i|i ∈ f_k}
    ap_genrules(f_k, H_1)
end

funkcija ap_genrules(f_k, H_1)
k = |f_k|
m = |H_m|
if (k > m + 1) then
   H_{m+1} = apriori_gen(H_m)
   for (svaki h_{m+1} ∈ H_{m+1} ) do
       conf = σ(f_k) / σ(f_k − h_{m+1})
       if (conf ≥ minconf ) then
          dodaj pravilo (f_k − h_{m+1} → h_{m+1})
       else
          odbaci h_{m+1} ∈ H_{m+1}
       end
   end
   ap_genrules(f_k, H_{m+1})
end
```

Променљива $k$ представља величину скупа честих ставки, док је $m$ величина последичног правила. Скуп $f_k$ је скуп честих ставки величине $k$, док је $H_m$ скуп последичног правила. Позивом процедуре *ap_genrules* са одговарајућим вредностима за $f_k$ и $H_m$ се генеришу одговарајућа правила. Више детаља о алгоритму се може прочитати у [21].

### 3.3.4 Алгоритам *SIDE* за генерисање правила

У 3.2 је наведено да је за истраживање коришћен алат *IBM Infosphere Warehouse*. Тај алат кориснику при коришћењу методе правила придруживања поред добро познатог Априори алгоритма за генерисање правила који је описан у 3.3.3 нуди још један алгоритам за генерисање правила придруживања. У питању је алгоритам *SIDE* (енг. *SImultaneous Depth-first Expansion*). Корисник дакле има могућност да бира који од ова два алгоритма жели да користи. У овом раду је одлучено да се користи алгоритам *SIDE*, па следи кратак опис разлика у имплементацији алгоритма *SIDE*

и алгоритма Априори.

Априори алгоритам који је имплементиран у *IBM Infosphere Warehouse* генерише правила претрагом у ширину. Испоставља се да је Априори алгоритам још увек присутан у *IBM Infosphere Warehouse* из историјских разлога, јер је алгоритам *SIDE* много бржи. *SIDE* алгоритам такође итеративно претражује простор кандидата за правила придруживања, и креће од једноставног правила које даље развија све док може понављајући све кораке док има нових кандидата. Али за разлику од Априори алгоритма *SIDE* покушава да комбинује претрагу у ширину и претрагу у дубину. Разлог комбиновања ова два начина претраге је што ни једна претрага посебно није довољно добра. Ако би се применила само претрага у дубину може се догодити да понестане меморије у раду са великом количином података јер је меморијски захтевна процедура, подаци се тада могу сместити на диск али би то значајно успорило алгоритам. Док се применом само претраге у ширину враћамо на Априори алгоритам.

Алгоритам *SIDE* има четири дела:

1. Пролази се кроз све податке и прикупљају статистички подаци о честим ставкама и паровима честих ставки.

2. Трансформишу се трансакције које садрже парове честих ставки у бинарну репрезентацију. То се ради јер у већини случајева то омогућава да се сви подаци задрже у меморији.

3. Главни корак је итеративно генерисање листе правила који су кандидати, оцењивање правила према подацима и одбацивање правила која нису задовољила постављене критеријуме. Нова дужа правила се генеришу на основу краћих која су задовољила постављене критеријуме у претходној итерацији.

4. На крају се рачуна лифт мера за скупове ставки и правила придруживања, и ради се филтрирање по задатом критеријуму лифт мере.

### 3.3.5 Рачунање квалитета правила придруживања

Применом техника методе правила придруживања може се добити велики број правила која задовољавају праг подршке и поузданости. Ипак, циљ је издвојити само најзанимљивија правила. Објективна мера за рачунање квалитета правила придруживања се заснива на подацима. Добија се из табеле контигената која у општем облику за правило $X \to Y$ изгледа као табела 1.

У табели 1 $f_{11}$ је подршка за $X$ и $Y$, $f_{10}$ је подршка за $X$ и $\overline{Y}$, $f_{01}$ је подршка за $\overline{X}$ и $Y$, $f_{00}$ је подршка за $\overline{X}$ и $\overline{Y}$, $f_{+1} = f_{11} + f_{01}$, $f_{1+} = f_{11} + f_{10}$, $f_{+0} = f_{10} + f_{00}$, $f_{0+} = f_{01} + f_{00}$ и $N$ је величина групе.

| | Y | $\overline{Y}$ | |
|---|---|---|---|
| X | $f_{11}$ | $f_{10}$ | $f_{1+}$ |
| $\overline{X}$ | $f_{01}$ | $f_{00}$ | $f_{0+}$ |
| | $f_{+1}$ | $f_{+0}$ | $N$ |

Табела 1: Табела контигената

Иако се све до сада ослањало на меру подршке и поузданости, испоставља се да оне имају одређене недостатке. Да би се боље сагледали недостаци поузданости потребно је размотрити следећи пример. Нека је потребно истражити односе између људи који играју фудбал и који играју кошарку, и нека су задати подаци из табеле 2.

| | фудбал | $\overline{\text{фудбал}}$ | |
|---|---|---|---|
| кошарка | 15 | 5 | 20 |
| $\overline{\text{кошарка}}$ | 75 | 5 | 80 |
| | 90 | 10 | 100 |

Табела 2: Табела са информацијама

На основу информација из табеле могуће је проценити правило {кошарка} $\rightarrow$ {фудбал}. Очигледно је да је подршка овог правила 15%, а поузданост 75%. Ова чињеница би била прихватљива да не важи да део људи који играју фудбал без обзира да ли играју кошарку је 80% и да део људи који играју и кошарку и фудбал је 75%. Што значи да се особи која игра кошарку смањује шанса да игра и фудбал са 80% на 75%. Због тога је претходно правило погрешно упркос високој поузданости. То се може десити ако се игнорише мера подршке.

Недостатак подршке огледа се у томе да некада могу да буду одбачена занимљива правила. То се може догодити у случајевима када је подршка тих ставки мања од *minsup*, али њихова фреквенција појављивања у трансакцијама јако велика. На пример, дати су подаци груписани у три групе $G_1$, $G_2$ и $G_3$, и нека су информације о тим групама представљене у табели 3.

| Група | $G_1$ | $G_2$ | $G_3$ |
|---|---|---|---|
| Подршка | <2% | 2-95% | >95% |
| Фреквенција ставки | 2000 | 400 | 20 |

Табела 3: Табела са информацијама о групама

Ако се над подацима чије су информације представљене у табели 3 постави превелика подршка (нпр. 20%) постоји велика шанса да ће бити одбачен велики број занимљивих правила. Слично, ако се подршка постави на јако малу вредност можемо укључити у испитивање велики број правила која нам нису интересантна. Тако да је потребно водити рачуна које вредности се узимају за подршку. Због тога постоје одређене

метрике које могу да помогну у томе да се процени квалитет правила придруживања, међутим свака од њих има своје недостатке.

Прва метрика је **лифт мера**, која се дефинише на следећи начин за промењљиве које нису бинарне

$$Lift = \frac{c(X \to Y)}{s(Y)}.$$

За бинарне промењљиве лифт мера се назива **интересни фактор** и дефинише се као

$$I(X,Y) = \frac{s(X,Y)}{s(A) \times s(B)} = \frac{N \cdot f_{11}}{f_{1+} \cdot f_{01}}.$$

Недостатак ове метрике је тај да када има вредност 1, тада су подаци независни и овом мером се не може добити ништа што би било од значаја.

Друга метрика је **корелација** која представља статистички засновану технику за анализу односа између пара промењљивих. За непрекидне промењљиве корелација се дефинише на следећи начин:

$$corr(X,Y) = \frac{covarijansa(X,Y)}{standardnaDevijacija(X) \cdot standardnaDevijacija(Y)}, \text{ где је}$$

$$covarijansa(X,Y) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x}) \cdot (y_k - \overline{y}),$$

$$standardnaDevijacija(X) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2},$$

$$standardnaDevijacija(Y) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2},$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k,$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k.$$

За бинарне промењљиве корелација се рачуна помоћу $\phi$-коефицијента, који се дефинише као

$$\phi = \frac{f_{11} \cdot f_{00} + f_{01} \cdot f_{10}}{\sqrt{f_{1+} \cdot f_{+1} \cdot f_{0+} \cdot f_{+0}}}.$$

Вредности корелације су у интервалу $[-1, 1]$ и ако је вредност корелације 0 то значи да су промењљиве статистички независне. Мана ове метрике је што се може догодити да пар промењљивих које се заједно јако често појављују и пар промењљивих које се јако ретко појављују имају исту вредност корелације у случају када се користи $\phi$-коефицијент. До тога долази јер $\phi$-коефицијент даје једнаку важност ставкама које су присутне и ставкама које нису присутне у трансакцији.

Трећа метрика је алтернативна метрика за рад са асиметричним бинарним подацима, назива се **IS-мера**. Дефинише се као

$$IS(X,Y) = \sqrt{I(X,Y) \times s(X,Y)} = \frac{s(X,Y)}{\sqrt{s(X) \cdot s(Y)}}.$$

Потенцијални недостатак ове мере је сличан као недостатак поузданости. Може се десити да има велике вредности за слабо корелисане променљиве, односно за правила која нису значајна.

### 3.3.6 Рад са различитим врстама атрибута

До сада је све посматрано кроз асиметричну бинарну репрезентацију података, сада је потребно проширити репрезентацију. Прво ћемо представити како се примењује метода правила придруживања над категоричким атрибутима у које спадају симетрични бинарни атрибути и именски атрибути. Да би могло да се ради са оваквим атрибутима потребно је да се трансформишу у погодан облик, односно у асиметричне бинарне атрибуте. Онда над таквим подацима се могу примењивати алгоритми. При овој трансформацији прави се нова ставка за сваки појединачни атрибут. Први проблем који се може јавити јесте да неки атрибут има велики број могућих вредности (на пример имена држава), тада је могуће агрегирати вредности таквих атрибута. Други проблем који се може јавити јесте ако расподела вредности јако одскаче на једну страну (на пример 95% вредности неког атрибута има исту вредност), у том случају проблем се може решити одабиром ставке са високом учесталошћу. Дакле, рад са категоричким атрибутима се своди на асиметричну бинарну репрезентацију.

Када је у питању рад са *непрекидним атрибутима* прича је мало компликованија. Постоје три методе које се користе у раду са непрекидним атрибутима. Метода заснована на дискретизацији је најопштија метода за рад са непрекидним атрибутима. Овим приступом групишу се вредности непрекидних атрибута у коначан број интервала, да би се тако добијени интервали даље мапирали у асиметричне бинарне атрибуте. Кључан параметар ове методе је број интервала. Ако је број интервала јако мали подршка може да буде недовољна, а ако је број интервала јако велики поузданост може да буде недовољна. Овај проблем се најчешће решава тако што се проба са више различитих вредности параметра који представља број интервала на колико се дели, који имају смисла за конкретан проблем и изабере најбољи. Друга метода је статистички заснована метода која прво издваја циљне атрибуте од остатка података и примењује Априори алгоритам или алгоритам ФП-раста за генерисање честих скупова ставки над тим остатком података, након што смо их представили бинарном репрезентацијом. За сваку честу ставку потребно је израчунати одабрану статистику (нпр. медијану) за одговарајућу циљну променљиву. Чест скуп ставки постаје правило за укључивање циљног атрибута као последичног правила. Тада се примењују статистички тестови да би се одредила интересантност правила. Правило је интересантно ако је статистика популације покривене правилом различита од статистике популације која није покривена правилом. Трећа метода, заснована на недискретизацији користи се када је занимљивије пронаћи везу између непрекидних атрибута него између њихових дискретних интервала, као на пример код анализе текстова.

# 4 Резултати

Прво су представљени неки од резултата који су добијени извршавањем одговарајућих *sql* упита. Ти резултати су представљени одговарајућим табелама или графицима.

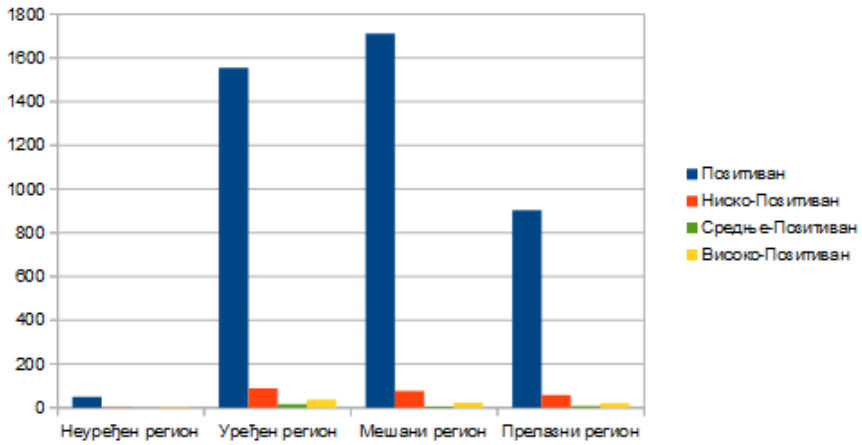
У табели 4 приказана је расподела протеина по групи организама којима припадају. Као што се да приметити највећи број протеина припада групи вируса, а након провере испоставило се да су сви ти вируси еукариотски.

| | Укупно |
|---|---|
| вируси | 478 |
| бактерије | 60 |
| еукарије | 158 |

Табела 4: Расподела протеина по врсти организама

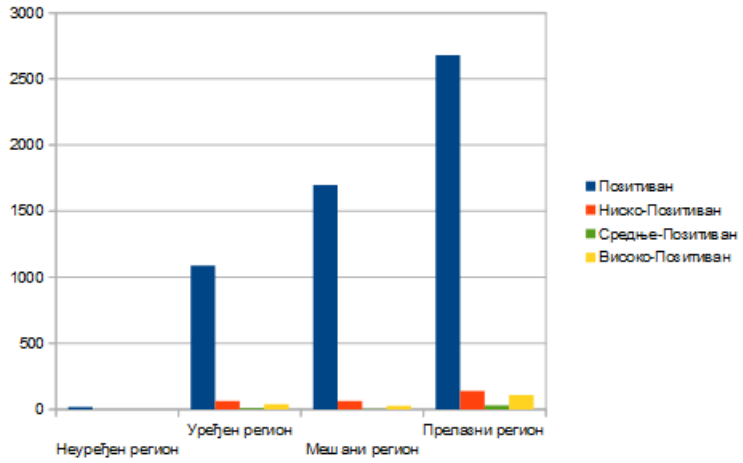
На слици 7 је представљена расподела позитивних експерименталних епитопа по групама (позитиван, ниско-позитиван, средње-позитиван, јако-позитиван) и по регионима којима епитоп може да припада (неуређен, уређен, мешани, прелазни) за *HLA-I* класу. Са слике се да видети да највећи број ових епитопа припада мешаном региону, док има јако мало епитопа у неуређеном региону. Такође, очигледно је да највише има оних епитопа који припадају групи чисто позитивних епитопа.

У обзир су узети само епитопи који се цели налазе у неком од наведених региона. То важи и за све наредне статистике. Исто је урађено и за предвиђене епитопе, рачунати су само предвиђени епитопи који су цели упали у неки од региона.



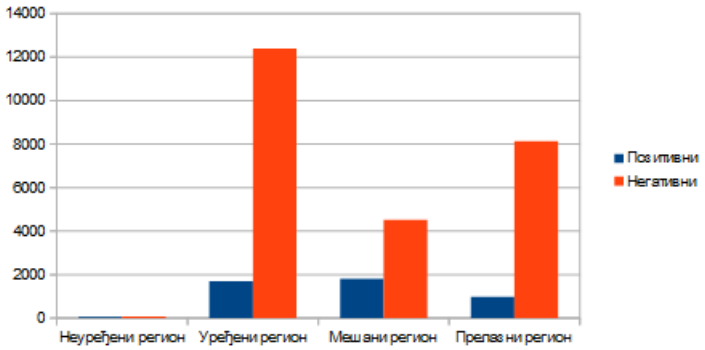

Слика 7: Расподела позитивних експерименталних епитопа по регионима код *HLA-I* класе

На слици 8 је представљена расподела позитивних експерименталних епитопа по групама и по регионима за *HLA-II* класу. Код ове класе највећи број епитопа се налази у прелазном региону, али и овде највише епитопа припада групи чисто позитивних. Дакле, расподела епитопа је мало другачија него код *HLA-I* класе.



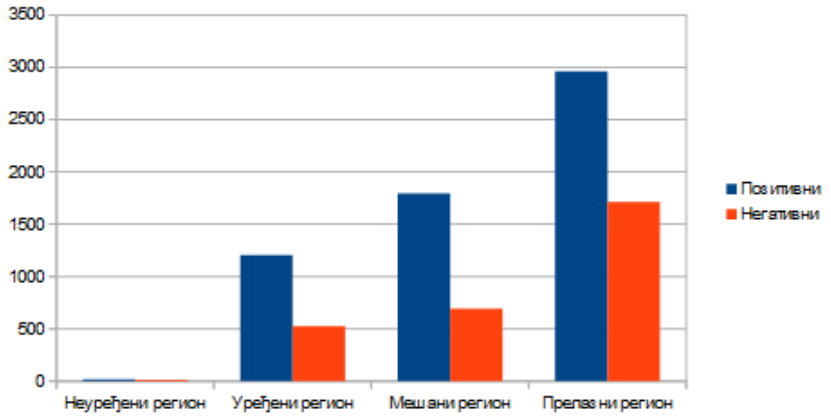

Слика 8: Расподела позитивних експерименталних епитопа по регионима код *HLA-II* класе

На слици 9 је приказан однос свих позитивних и свих негативних експерименталних епитопа *HLA-I* класе. На основу тога може се закључити да негативних епитопа има значајно више од позитивних, и да већина негативних епитопа припада уређеном региону.




Слика 9: Расподела позитивних и негативних експерименталних епитопа по регионима код *HLA-I* класе

На слици 10 је приказан однос свих позитивних и свих негативних експерименталних епитопа *HLA-II* класе. На њој се да видети да код *HLA-II* класе има више позитивних епитопа и да највећи број њих припада прелазном региону. Види се да је ова расподела доста другачија од расподеле коју имамо код *HLA-I* класе.



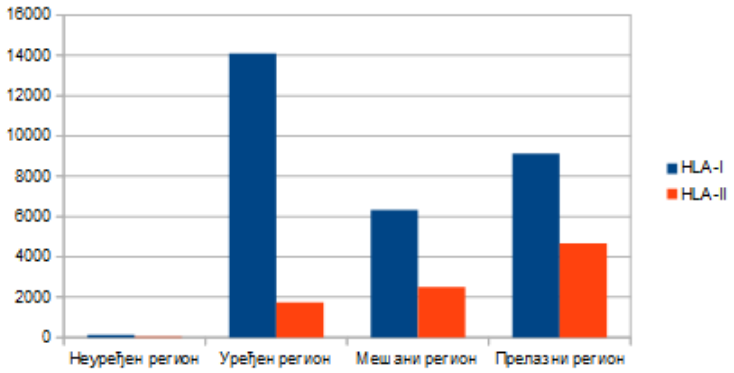

Слика 10: Расподела позитивних и негативних експерименталних епитопа по регионима код *HLA-II* класе

У табели 5 је приказана расподела експерименталних епитопа по регионима за сваку *HLA* класу. Из табеле се може видети да за обе класе важи да имају јако мало епитопа у неуређеном региону. Док се за *HLA-I* класу највећи број епитопа налази у уређеном региону, а за *HLA-II* класу у прелазном региону.

| | *HLA-I (%)* | *HLA-II (%)* |
|---|---|---|
| неуређени | 0.43 | 0.35 |
| уређени | 47.48 | 19.4 |
| мешани | 21.34 | 27.87 |
| прелазни | 30.75 | 52.38 |

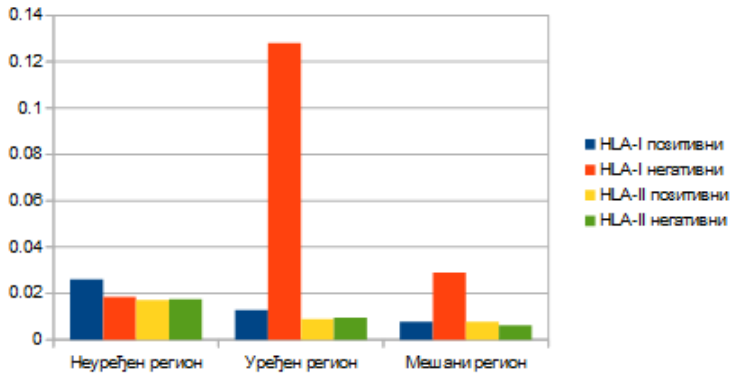
Табела 5: Расподела експерименталних епитопа по регионима у процентима

На слици 11 је графички приказано оно што стоји у табели 5 да би се после лакше уочила разлика у расподели експерименталних и предвиђених епитопа.



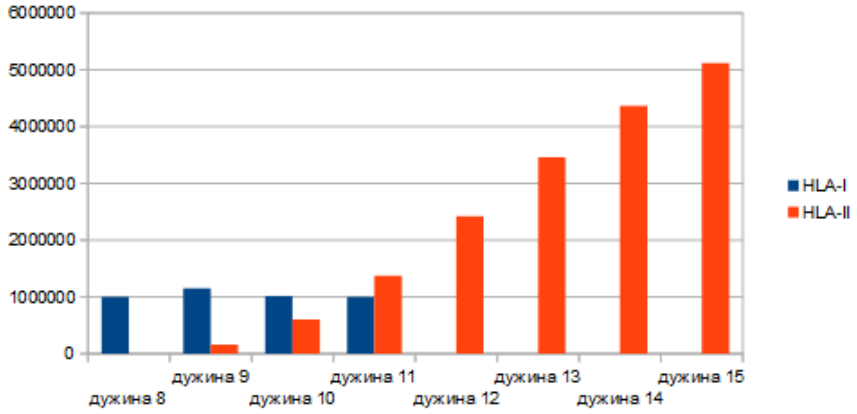

Слика 11: Расподела експерименталних епитопа по регионима

На слици 12 је приказана фреквенција појављивања експерименталних епитопа по регионима груписаних на позитивне и негативне. Са те слике види се да се негативни експериментални епитопи *HLA-I* класе којих има највише најчешће појављују у уређеном региону.




Слика 12: Фреквенција појављивања експерименталних епитопа по регионима

Следе статистике предвиђених података. На слици 13 приказана је расподела предвиђених епитопа по њиховим дужинама, раздвојених по *HLA* класи. Највише епитопа *HLA-II* класе је дугачко 15 аминокиселина, док за *HLA-I* класу епитопи су скоро равномерно распоређени по дужинама 8, 9, 10 и 11.



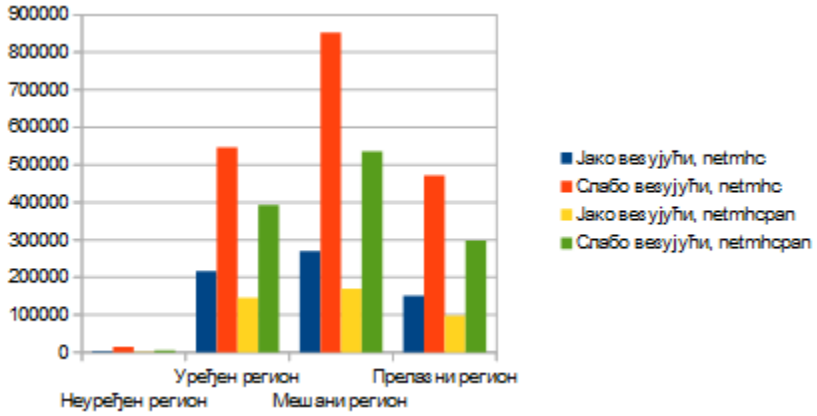

Слика 13: Расподела предвиђених епитопа по дужинама, груписаних по *HLA* класи

У табели 6 представљена је расподела предвиђених епитопа по регионима груписаних по начину везивања (јако-везујући, слабо-везујући) у процентима за *HLA-I* класу. На основу табеле се види да јако мало предвиђених епитопа (и јако-везујућих и слабо-везујућих) припада неуређеном региону, док се највећи проценат налази у мешаном региону.

| | *netmhc* | | *netmhcpan* | |
|---|---|---|---|---|
| | *Слабо везујући (%)* | *Јако везујући (%)* | *Слабо везујући (%)* | *Јако везујући (%)* |
| неуређени | 0.49 | 0.76 | 0.24 | 0.44 |
| уређени | 33.84 | 29.01 | 35.23 | 31.88 |
| мешани | 42.17 | 45.20 | 41.08 | 43.35 |
| прелазни | 23.5 | 25.03 | 23.45 | 24.42 |

Табела 6: Расподела предвиђених епитопа по регионима груписаних по јачини везивања у процентима за *HLA-I* класу

На слици 14 је графички представљена расподела предвиђених епитопа по регионима за *HLA-I* класу. Она заправо представља графичку интерпретацију табеле 6 да се лакше упореди еквивалентни график за *HLA-II* класу.



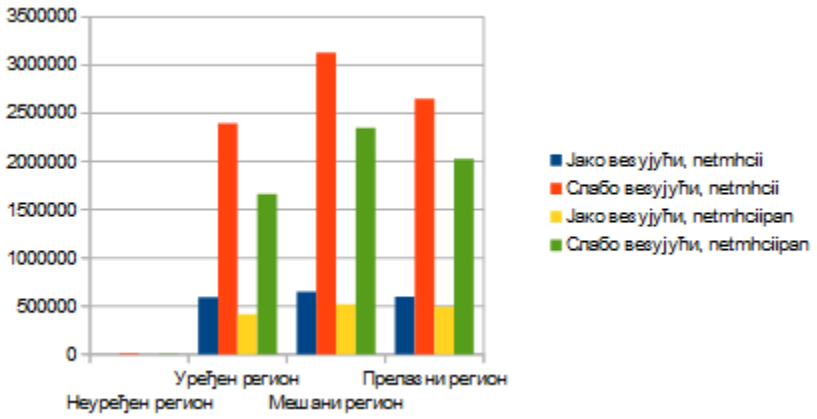

Слика 14: Расподела предвиђених епитопа по регионима груписаних по јачини везивања за *HLA-I* класу

У табели 7 је дата расподела предвиђених епитопа по регионима груписаних по јачини везивања у процентима за *HLA-II* класу, која је графички представљена на слици 15.

| | *netmhcii* | | *netmhciipan* | |
|---|---|---|---|---|
| | *Слабо везујући (%)* | *Јако везујући (%)* | *Слабо везујући (%)* | *Јако везујући (%)* |
| неуређени | 0.01 | 0.02 | 0.01 | 0.08 |
| уређени | 32.24 | 29.24 | 29.16 | 27.51 |
| мешани | 35.3 | 38.21 | 36.37 | 38.88 |
| прелазни | 32.45 | 32.43 | 34.46 | 33.53 |

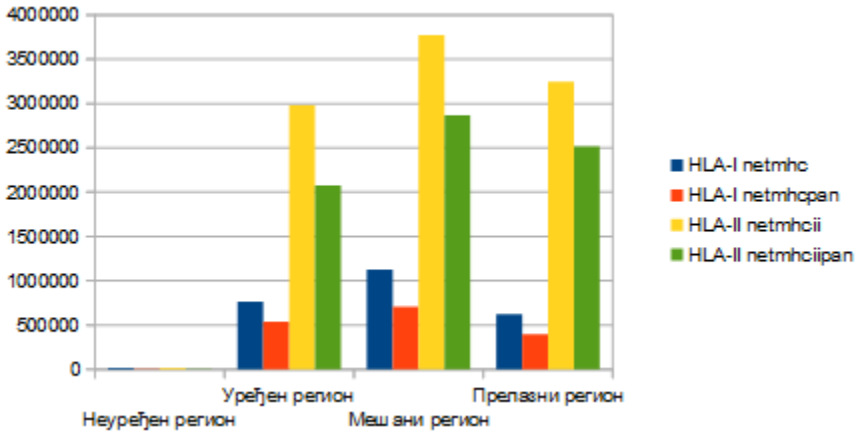
Табела 7: Расподела предвиђених епитопа по регионима груписаних по јачини везивања у процентима за *HLA-II* класу

Ако упоредимо графике са слика 14 и 15 види се да највећи проценат предвиђених епитопа за сваку групу у оба случаја припада мешаном региону, али да је та предност већа код епитопа *HLA-I* класе него код епитопа *HLA-II* класе.




Слика 15: Расподела предвиђених епитопа по регионима груписаних по јачини везивања за *HLA-II* класу

На слици 16 је приказана расподела предвиђених епитопа по регионима груписаних по *HLA* класи.

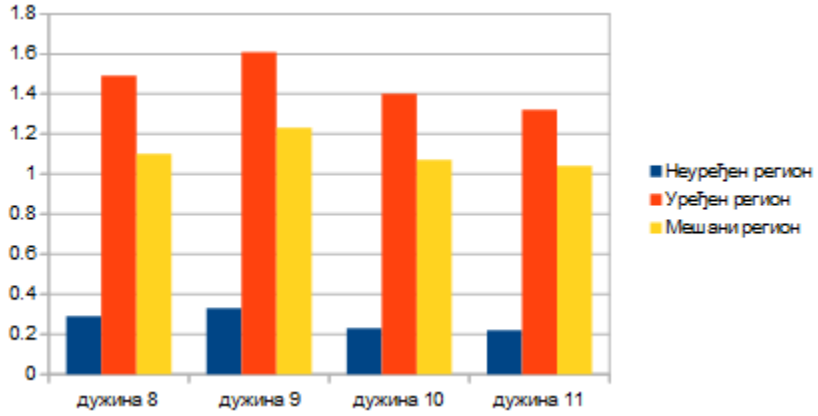



Слика 16: Расподела предвиђених епитопа по регионима груписаних по *HLA* класи

Са слике 16 да се приметити да је највећи број епитопа обе *HLA* класе у мешаном региону. Посматрајући тај график са њему еквивалентним графиком за експерименталне епитопе, видимо да су код предвиђених доминантнији епитопи *HLA-II* класе, док су код експерименталних епитопа доминантнији епитопи *HLA-I* класе. Поред тога код предвиђених епитопа пик за *HLA-I* епитопе је у мешаном региону, док је код експерименталних епитопа пик за исту класу епитопа у уређеном региону. За *HLA-II* класу експериментални епитопи имају мањи пик у прелазном региону, а предвиђени епитопи имају пик у мешаном региону.

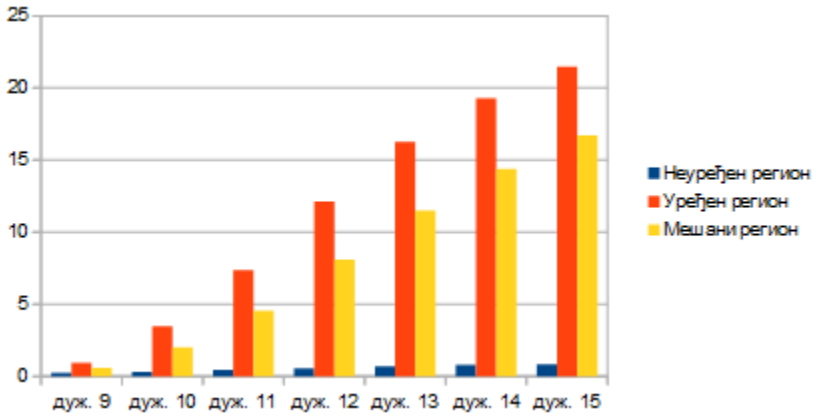
На слици 17 је приказана фреквенција појављивања предвиђених епитопа у регионима на 100 аминокиселина за сваку могућу дужину *HLA-I* класе. Са ове слике се може закључити да се предвиђени епитопи *HLA-I* класе било које дужине најчешће

јављају у уређеном региону.




Слика 17: Фреквенција појављивања предвиђених епитопа по регионима за *HLA-I* класу

На слици 18 је приказана фреквенција појављивања предвиђених епитопа у регионима на 100 аминокиселина за сваку могућу дужину *HLA-II* класе. Са ове слике се види да и за *HLA-II* класу важи да се предвиђени епитопи ове класе било које дужине најчешће јављају у уређеном региону.

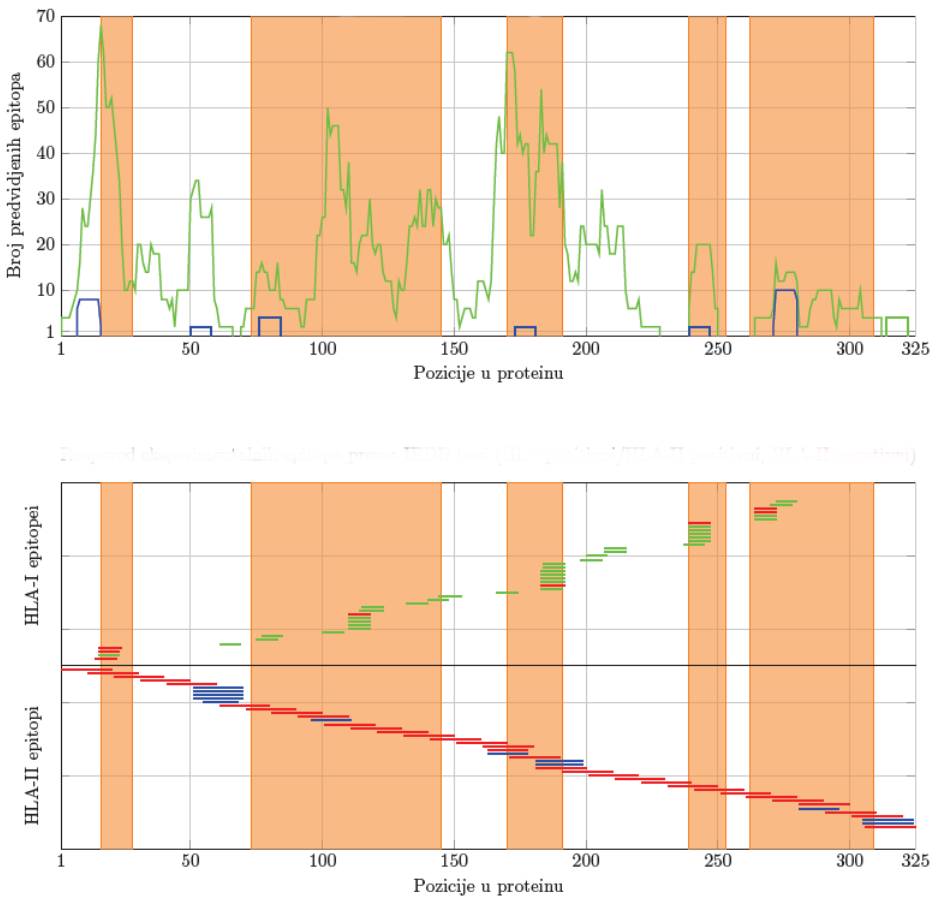



Слика 18: Фреквенција појављивања предвиђених епитопа по регионима за *HLA-II* класу

Наредне две слике, 19 и 20, представљају консензус дужине 9 уређених/неуређених региона за две различите бактерије. Горње слике представљају распоред предвиђених епитопа, док доње слике представљају распоред експерименталних епитопа. Зелена боја на горњим сликама представља епитопе *HLA-I* класе, а плава боја епитопе *HLA-II* класе. Зелена боја на доњој слици представља позитивне епитопе класе *HLA-I*, плава боја представља позитивне епитопе класе *HLA-II*, а црвена боја означава негативне епитопе обе класе. Бела боја на графицима означава мешани регион,

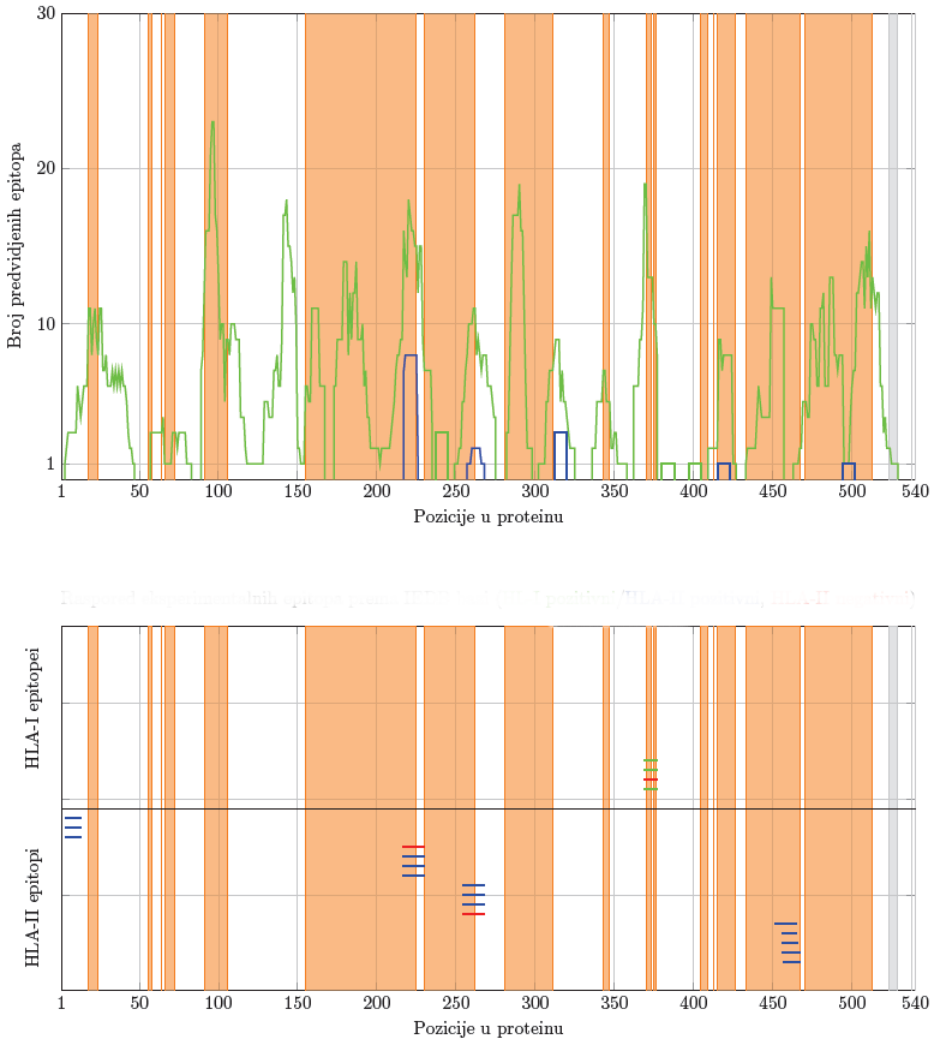
наранџаста боја означава уређен регион, а сива боја означава неуређен регион.

На слици 19 за предвиђене епитопе (горњи график) види се да епитопа *HLA-I* класе има много више него епитопа *HLA-II* класе, али и да се пикови обе класе углавном поклапају. Код експерименталних епитопа *HLA-I* класе види се да су доминантнији позитивни епитопи, док су код *HLA-II* класе доминантнији негативни епитопи. Оно што се још да видети јесте да се пикови код предвиђених епитопа поклапају са појавом негативних експерименталних епитопа код *HLA-I* класе, односно позитивним експерименталним епитопима код *HLA-II* класе.




Слика 19: Распоред епитопа у консензусу региона протеина 398979

На слици 20 за предвиђене епитопе се види да епитопа *HLA-I* класе има значајно више у односу на епитопе *HLA-II* класе, и да им се пикови углавном поклапају. За експерименталне епитопе обе *HLA* класе види се да су јако ретки. Још се да приметити да се појава експерименталних епитопа поклапа са пиковима код предвиђених епитопа.




Слика 20: Распоред епитопа у консензусу региона протеина 116244

## 4.1 Правила придруживања

Следи приказ неких издвојених, занимљивих правила придруживања која су добијена истраживањем. На слици 21 су приказана издвојена правила за експерименталне податке, док су на слици 22 приказана правила за предвиђене податке.

У току истраживања је добијен велики број правила са високом подршком и поузданошћу, али је већина тих правила била очигледна и бесмислена када је у истраживање била укључена позиција протеина. Због тога је из истраживања искључена позиција протеина, након чега је добијен велики број правила из којих су се могла изабрати нека која су занимљива. Да би се пронашла занимљива правила било је потребно додатно анализирати добијена правила.

| Rule | Support | Confidence | ▼ Lift |
|---|---|---|---|
| [ALLELA=HLA-DR]+[TAXONOMY_1=viruses] ==> [LENGTH >= 12 AND < 16 {=3/4}] | 0,5364% | 81,8182% | 10,80 |
| [LENGTH >= 8 AND < 12 {=2/4}]+[VALUE2=Negative]+[TAXONOMY_1=eukaryota] ==> [ALLELA=HLA-A*02:01] | 1,2981% | 78,5266% | 6,80 |
| [UREDJENOST=O]+[LENGTH >= 8 AND < 12 {=2/4}]+[VALUE2=Negative]+[TAXONOMY_1=eukaryota] ==> [ALLELA=HLA-A*02:01] | 0,5286% | 77,5665% | 6,72 |
| [ALLELA=HLA-DR1]+[VALUE2=Negative]+[TAXONOMY_1=viruses] ==> [LENGTH >= 16 {=4/4}] | 0,7773% | 98,0392% | 5,88 |
| [ALLELA=HLA-DR1]+[VALUE2=Negative]+[UREDJENOST=P] ==> [LENGTH >= 16 {=4/4}] | 0,5415% | 96,3134% | 5,77 |
| [VALUE2=Negative]+[TAXONOMY_1=bacteria]+[UREDJENOST=P] ==> [LENGTH >= 16 {=4/4}] | 1,5754% | 94,1176% | 5,64 |
| [ALLELA=HLA-DR1]+[TAXONOMY_1=viruses]+[UREDJENOST=P] ==> [LENGTH >= 16 {=4/4}] | 0,8784% | 92,6230% | 5,55 |
| [ALLELA=HLA-DR1]+[TAXONOMY_1=viruses]+[VALUE2=Positive] ==> [LENGTH >= 16 {=4/4}] | 0,5882% | 84,0741% | 5,04 |
| [TAXONOMY_1=bacteria]+[VALUE2=Positive]+[UREDJENOST=P] ==> [LENGTH >= 16 {=4/4}] | 1,3189% | 80,1575% | 4,80 |
| [ALLELA=HLA-DR1]+[VALUE2=Positive]+[UREDJENOST=P] ==> [LENGTH >= 16 {=4/4}] | 0,5389% | 79,0875% | 4,74 |
| [ALLELA=HLA-DRB1*04:01]+[VALUE2=Positive]+[UREDJENOST=P] ==> [LENGTH >= 16 {=4/4}] | 0,5208% | 70,2797% | 4,21 |
| [ALLELA=HLA-DRB1*04:01]+[TAXONOMY_1=viruses]+[LENGTH >= 16 {=4/4}] ==> [VALUE2=Positive] | 0,6037% | 95,4918% | 3,50 |
| [ALLELA=HLA-DR]+[LENGTH >= 16 {=4/4}] ==> [VALUE2=Positive] | 0,8525% | 89,4022% | 3,28 |
| [ALLELA=HLA-DR]+[UREDJENOST=N] ==> [VALUE2=Positive] | 0,5674% | 83,2700% | 3,05 |
| [ALLELA=HLA-DR]+[UREDJENOST=P] ==> [VALUE2=Positive] | 0,8110% | 82,5858% | 3,03 |
| [ALLELA=HLA-DR]+[TAXONOMY_1=eukaryota] ==> [VALUE2=Positive] | 0,9432% | 82,5397% | 3,02 |
| [UREDJENOST=N]+[TAXONOMY_1=viruses]+[LENGTH >= 16 {=4/4}] ==> [VALUE2=Positive] | 1,8008% | 79,2474% | 2,90 |
| [ALLELA=HLA-DR]+[LENGTH >= 12 AND < 16 {=3/4}] ==> [VALUE2=Positive] | 0,8887% | 78,4897% | 2,88 |
| [UREDJENOST=O]+[LENGTH >= 12 AND < 16 {=3/4}]+[TAXONOMY_1=viruses] ==> [VALUE2=Positive] | 0,7074% | 78,2235% | 2,87 |
| [UREDJENOST=O]+[LENGTH >= 12 AND < 16 {=3/4}] ==> [VALUE2=Positive] | 1,3525% | 78,1437% | 2,86 |
| [UREDJENOST=O]+[TAXONOMY_1=eukaryota]+[LENGTH >= 16 {=4/4}] ==> [VALUE2=Positive] | 0,7307% | 77,4725% | 2,84 |
| [LENGTH >= 12 AND < 16 {=3/4}]+[TAXONOMY_1=eukaryota]+[UREDJENOST=P] ==> [VALUE2=Positive] | 1,1401% | 77,0578% | 2,82 |
| [UREDJENOST=O]+[LENGTH >= 12 AND < 16 {=3/4}]+[TAXONOMY_1=eukaryota] ==> [VALUE2=Positive] | 0,5156% | 76,8340% | 2,82 |
| [LENGTH >= 12 AND < 16 {=3/4}]+[TAXONOMY_1=eukaryota] ==> [VALUE2=Positive] | 2,2853% | 73,1343% | 2,68 |
| [LENGTH >= 12 AND < 16 {=3/4}]+[UREDJENOST=P] ==> [VALUE2=Positive] | 2,4486% | 72,9730% | 2,67 |
| [TAXONOMY_1=eukaryota]+[UREDJENOST=P]+[LENGTH >= 16 {=4/4}] ==> [VALUE2=Positive] | 2,3216% | 71,6800% | 2,63 |
| [UREDJENOST=N]+[TAXONOMY_1=bacteria]+[LENGTH >= 16 {=4/4}] ==> [VALUE2=Positive] | 0,5415% | 71,5753% | 2,62 |
| [LENGTH >= 12 AND < 16 {=3/4}]+[TAXONOMY_1=viruses] ==> [VALUE2=Positive] | 2,7880% | 70,7429% | 2,59 |
| [LENGTH >= 12 AND < 16 {=3/4}]+[UREDJENOST=N]+[TAXONOMY_1=viruses] ==> [VALUE2=Positive] | 0,9121% | 70,4000% | 2,58 |
| [VALUE2=Negative]+[TAXONOMY_1=bacteria]+[LENGTH >= 16 {=4/4}] ==> [UREDJENOST=P] | 1,5754% | 72,6404% | 2,03 |
| [ALLELA=HLA-B*49:01]+[LENGTH >= 8 AND < 12 {=2/4}]+[TAXONOMY_1=viruses] ==> [VALUE2=Negative] | 2,5082% | 100,0000% | 1,38 |
| [ALLELA=HLA-A*33:03]+[UREDJENOST=P] ==> [VALUE2=Negative] | 1,2593% | 100,0000% | 1,38 |
| [LENGTH >= 8 AND < 12 {=2/4}]+[ALLELA=HLA-A*33:03]+[UREDJENOST=P] ==> [VALUE2=Negative] | 1,2593% | 100,0000% | 1,38 |

Слика 21: Правила придруживања за експерименталне податке

| Rule | Support | Confidence | ▼ Lift |
|---|---|---|---|
| [ALLELE=HLA-B27:09]+[BONDING=WB] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,2037% | 100,0000% | 3,44 |
| [ALLELE=B3501]+[TAXONOMY_1=viruses] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1908% | 100,0000% | 3,44 |
| [UREDJENOST=N]+[ALLELE=HLA-B15:03]+[BONDING=WB] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1853% | 100,0000% | 3,44 |
| [ALLELE=HLA-B35:01]+[TAXONOMY_1=viruses] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1676% | 100,0000% | 3,44 |
| [ALLELE=HLA-B15:03]+[BONDING=SB] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1667% | 100,0000% | 3,44 |
| [ALLELE=HLA-B27:09]+[BONDING=WB]+[TAXONOMY_1=viruses] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1634% | 100,0000% | 3,44 |
| [ALLELE=HLA-B35:01]+[BONDING=WB] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1614% | 100,0000% | 3,44 |
| [TAXONOMY_1=viruses]+[ALLELE=HLA-B15:01] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1602% | 100,0000% | 3,44 |
| [UREDJENOST=N]+[ALLELE=HLA-B15:03]+[BONDING=WB]+[TAXONOMY_1=viruses] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1541% | 100,0000% | 3,44 |
| [ALLELE=B3501]+[BONDING=WB]+[TAXONOMY_1=viruses] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1527% | 100,0000% | 3,44 |
| [UREDJENOST=O]+[ALLELE=HLA-B15:03]+[BONDING=WB] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1504% | 100,0000% | 3,44 |
| [ALLELE=HLA-B15:03]+[UREDJENOST=P] ==> [EP_LENGTH < 11.9 {=1/4}] | 0,1501% | 100,0000% | 3,44 |
| [EP_LENGTH < 11.9 {=1/4}]+[ALLELE=B1501] ==> [BONDING=WB] | 0,1613% | 99,8628% | 1,25 |
| [UREDJENOST=N]+[ALLELE=HLA-DQA10401-DQB10402] ==> [BONDING=WB] | 0,1740% | 98,4462% | 1,23 |
| [EP_LENGTH >= 11.9 AND < 13.4 {=2/4}]+[ALLELE=DRB1_0802] ==> [BONDING=WB] | 0,2472% | 98,3548% | 1,23 |
| [ALLELE=HLA-DQA10301-DQB10302]+[TAXONOMY_1=viruses] ==> [BONDING=WB] | 0,2393% | 98,2402% | 1,23 |
| [EP_LENGTH >= 11.9 AND < 13.4 {=2/4}]+[ALLELE=DRB1_0802]+[TAXONOMY_1=viruses] ==> [BONDING=WB] | 0,2004% | 98,1955% | 1,23 |
| [ALLELE=HLA-DQA10301-DQB10302] ==> [BONDING=WB] | 0,3352% | 98,1368% | 1,23 |
| [ALLELE=DRB1_0802]+[UREDJENOST=P] ==> [BONDING=WB] | 0,2552% | 97,9074% | 1,22 |
| [TAXONOMY_1=viruses]+[ALLELE=HLA-DQA10401-DQB10402] ==> [BONDING=WB] | 0,2908% | 97,8905% | 1,22 |
| [ALLELE=HLA-DQA10401-DQB10402] ==> [BONDING=WB] | 0,4037% | 97,7223% | 1,22 |
| [ALLELE=HLA-DRB10802]+[TAXONOMY_1=viruses]+[UREDJENOST=P] ==> [BONDING=WB] | 0,1689% | 97,7073% | 1,22 |
| [ALLELE=DRB1_0802]+[TAXONOMY_1=viruses]+[UREDJENOST=P] ==> [BONDING=WB] | 0,1890% | 97,6441% | 1,22 |
| [ALLELE=DRB1_0802]+[EP_LENGTH >= 13.4 AND < 14.9 {=3/4}] ==> [BONDING=WB] | 0,1958% | 97,6416% | 1,22 |
| [ALLELE=DRB1_0802]+[TAXONOMY_1=viruses] ==> [BONDING=WB] | 0,5910% | 97,6259% | 1,22 |
| [ALLELE=DRB4_0101]+[EP_LENGTH >= 11.9 AND < 13.4 {=2/4}]+[TAXONOMY_1=viruses] ==> [BONDING=WB] | 0,2204% | 97,6088% | 1,22 |
| [UREDJENOST=N]+[ALLELE=DRB1_0802] ==> [BONDING=WB] | 0,3095% | 97,5705% | 1,22 |
| [ALLELE=HLA-DRB10802]+[UREDJENOST=P] ==> [BONDING=WB] | 0,2384% | 97,4444% | 1,22 |
| [ALLELE=DRB1_0802]+[TAXONOMY_1=viruses]+[EP_LENGTH >= 13.4 AND < 14.9 {=3/4}] ==> [BONDING=WB] | 0,1580% | 97,4088% | 1,22 |
| [UREDJENOST=N]+[ALLELE=DRB1_0802]+[TAXONOMY_1=viruses] ==> [BONDING=WB] | 0,2757% | 97,3833% | 1,22 |
| [ALLELE=DRB4_0101]+[EP_LENGTH >= 11.9 AND < 13.4 {=2/4}] ==> [BONDING=WB] | 0,2957% | 97,2217% | 1,21 |
| [EP_LENGTH >= 14.9 {=4/4}]+[ALLELE=DRB1_0802] ==> [BONDING=WB] | 0,2301% | 97,1659% | 1,21 |
| [EP_LENGTH >= 11.9 AND < 13.4 {=2/4}]+[ALLELE=DRB1_0405] ==> [BONDING=WB] | 0,3588% | 97,1525% | 1,21 |

Слика 22: Правила придруживања за предвиђене податке

# 5 Закључак

На основу резултата представљених у претходном поглављу може се закључити више ствари. Позитивни експериментални епитопи се различито групишу у зависности којој *HLA* класи припадају. Поред тога којој *HLA* класи припадају експериментални епитопи зависи однос позитивних и негативних епитопа у уређеним и неуређеним регионима протеина. Такође можемо закључити да за прикупљене протеине предвиђени епитопи великом већином припадају *HLA-II* класи, а да опет већина њих припада мешаном региону (није у чистом уређеном региону, а није ни у чистом неуређеном региону), док је јако мали број епитопа за обе *HLA* класе у чистом неуређеном региону.

## 5.1 Даљи рад

У даљем раду планира се анализа сваке статистике посебно за сваки предиктор који је коришћен, анализа њихових међусобних резултата и анализа њихових резултата у односу на експерименталне податке. Поред тога планира се да се у истраживање више укључи предвиђање *anchor*-а, које је овде коришћено само за прецизније процене позиција епитопа.

# 6 Литература

[1] N. M. Luscombe, D. Greenbaum, M. Gerstein, *What is Bioinformatics? A Proposed Definition and Overview of the Field*, Department of Molecular Biophysics and Biochemistry, Yale University, New Haven (USA), 2001.

[2] LM. Iakoucheva, CJ. Brown , JD. Lawson, Z. Obradovic , *Dunker AK: Intrinsic disorder in cell-signaling and cancer-associated proteins*, Journal of Molecular Biology, 323(3):573-584, 2002.

[3] N. S. Mitić, M. D. Pavlović, D. R. Jandrlić, *Epitope distribution in ordered and disordered protein regions— Part A. T-cell epitope frequency, affinity and hydropathy*, Journal of Immunological Methods, 406, 83-103., 2014.

[4] M. D. Pavlović, D. R. Jandrlić, N. S. Mitić, *Epitope distribution in ordered and disordered protein regions. Part B — Ordered regions and disordered binding sites are targets of T- and B-cell immunity,*Journal of Immunological Methods, 407, 90-107., 2014.

[5] E. Feedwell, A. Narayanan, *Intalligent bioinformatics*, Wiley, 2005

[6] Никетић В., *Принципи структуре и активности протеина*, Издавач: Хемијски факултет, Универзитет у Београду, Београд, 1995.

[7] B. Uverski, *Article ID 568068*, Journal of Biomedicine and Biotechnology, 2010.

[8] C. A. F. Andersen, A. G. Palmer, S. Brunak, and B. Rost, *Continuum secondary structure captures protein flexibility*, 10:175-184., 2002.

[9] M. Y. Lobanov, O. V. Galzitskaya, *The Ising model for prediction of disordered residues from protein sequence alone*, Physical Biology, volume 8, number 3, 2010.

[10] Z. Dosztanyi, V. Csizmok, P. Tompa and I. Simon, *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content*, Bioinformatics, volume 21, issue 16, pages 3433-3434, 2005.

[11] J. Lafferty, A. McCallum, F. Perreira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, 18th International Conference on Machine Learning (ICML), page:282–289., 2001.

[12] Z. R. Yang, R. Thomson, P. McNeil, R. M. Esnouf, *RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins*, Bioinfomatics, volume 21, number 16, pages 3369–3376, 2005.

[13] http://www.iedb.org/database_export.php, последња посета октобар 2014.

[14] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, *Introduction to Linear Regression Analysis*, Wiley, 2013.

[15] P-N. Tan, M. Steinbach, V. Kumar, *Introduction to data mining*, Addison Wesley, 2006.

[16] V. Kumar, X. Wu, *Top 10 algorithms in data mining*, CRC Press, 2009.

[17] C. Vajiramedhin, J. Werapun, *EBPA: An Efficient Data Structure for Frequent Closed Itemset Mining*, Applied Mathematical Sciences, volume 7, 2013.

[18] E-H. Han, G. Karypis, V. Kumar, *Scalable parallel data mining for association rules*, IEEE Transactions on knowledge and data engineering, 2000.

[19] J. Han, J. Pei, Y. Yin, R. Mao, *Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach*, Data Mining and Knowledge Discovery, 8, 53–87, 2004.

[20] G. Grahne, J. ZhuFast, *Algorithms for Frequent Itemset Mining Using FP-Trees*, IEEE Transactions on knowledge and data engineering, volume 17, number 10, 2005.

[21] R. Agrawal, R. Srikant, *Fast algorithms for mining association rules*, IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120