УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Милица Којичић

РАЗВОЈ ВЕБ АПЛИКАЦИЈЕ ЗА ПРЕДИКЦИЈУ И МЕТАПРЕДИКЦИЈУ Т-ЋЕЛИЈСКИХ ЕПИТОПА

мастер рад

Београд, 2017.

Ментор:

доц. др Јована Ковачевић
Универзитет у Београду, Математички факултет

Чланови комисије:

проф. др Гордана Павловић-Лажетић
Универзитет у Београду, Математички факултет

Анђелка Зечевић
Универзитет у Београду, Математички факултет

Датум одбране: 29.9.2017.

Mojoj porodici

Наслов мастер рада: Развој веб апликације за предикцију и метапредикцију Т-ћелијских епитопа

Резиме: Имуноинформатика (енгл. *immunoinformatics*) је дисциплина биоинформатике која укључује примену рачунарских метода у решавању имунолошких проблема. Један од изазова имуноинформатике представља предикција Т-ћелијских епитопа. Т-ћелијски епитопи су врло значајне супстанце (по структури протеини), одговорне за имунитет организма. Наиме, епитоп је део антигена, а антиген је било која супстанца која може да изазове имуни одговор. Дакле, епитоп или антигентска детерминанта је структурна компонента антигена (низ аминокиселина) коју препознају ћелије имуног система, нарочито антитела и рецептори на Б-ћелијама и Т-ћелијама.

Идентификација епитопа, поготово оних који су карактеристични за туморске ћелије, од круцијалне је важности у дизајнирању вакцина и персонализованој туморској имунотерапији. Експерименталне методе за препознавање Т-ћелијских епитопа су веома скупе и временски захтевне, што је условило потребу за развојем рачунарских метода за њихову предикцију.

Главни циљ рада је развој веб апликације која комбинује неколико постојећих предикционих алата које корисник може да покреће са жељеним параметрима. Као улаз, веб апликација користи примарну секвенцу протеина, а као излаз, на основу одабраног предиктора, приказује који су делови протеина препознати као епитопи и са којим скором. У оквиру веб апликације је такође имплементиран и нови алат за метапредикцију епитопа.

Кључне речи: Т-ћелијски епитоп, предикциона метода, имуноинформатика

Садржај

Глава 1

Увод

Научна заједница је преплављена огромном количином биолошких података из различитих биолошких дисциплина. Како технике секвенцирања људског генома нове генерације (енгл. *next-generation sequencing*) све више напредују, тако се нагомилава количина доступних информација о људском геному. Те информације су од огромног значаја и захтевају напредне рачунарске алгоритме и алате који ће их обрађивати. Између осталог, и имуноинформатичке базе података се константно унапређују у циљу прилагођавања оваквој експанзији података [20].

Главни циљ имуноинформатике јесте да разуме и организује ту огромну количину података служећи се математичким и рачунарским методама у циљу добијања имунолошки смислених интерпретација тих података. Те рачунарске методе базирају се на статистици и машинском учењу и у стању су да моделују молекуларне интеракције и механизме имуног система.

У раду ће бити речи о једном од изазова имуноинформатике - предикцији Т-ћелијских епитопа. Епитоп је, наиме, део антигена изложен на површини ћелије у којој се тај антиген синтетише и који захваљујући својој позицији може бити препознат од стране ћелија имуног система. На пример, секвенцирањем генома човека који има тумор могу се добити информације о мутацијама специфичним за тог човека. На основу њих можемо добити информације о антигенима који су по структури протеини и који настају услед тих истих мутација које су узроковале тумор, а самим тим и о епитопима које имуни систем може да препозна, а затим и уништи такве „лоше ћелије". Захваљујући информацијама које добијамо секвенцирањем људског генома, егзома и РНК секвенцирањем, могуће је са великом прецизношћу добити информације о таквим епитопима.

Овакав развој доводи до нових концепата у дизајнирању вакцина састављених од изолованих епитопа који могу да стимулишу специфичне имуне одговоре. У односу на традиционални приступ где се при прављењу вакцина гаје одговарајући патогени у лабораторијским условима, овакав приступ има предност јер је бржи и јефтинији. Како би овакав приступ уопште био могућ, потребно је узети у обзир потенцијалне протеинске продукте и идентификовати оне који могу бити имуногени, односно оне који су способни да изазову имуни одговор [19]. У овом контексту, биоинформатика игра кључну улогу у анализи људских генома и препознавању одговарајућих епитопа.

Наиме, у раду ће бити представљени различити алати за предикцију Т-ћелијских епитопа. Алати су бесплатни за академске кориснике и могу се преузети са веб стране Центра за анализу биолошких секвенци Техничког Универзитета у Данској [3]. Централна тема рада је развој веб апликације која укључује 5 различитих предикционих алата и омогућава њихово покретање са различитим параметрима. Веб апликација такође садржи и нови, тзв. „мета" метод (метапредиктор) који комбинује резултате ових алата и врши коначну предикцију на основу различитих гласачких шема.

Након уводног поглавља, у глави 2 дат је кратак приказ основних имунолошких појмова и механизама. Затим су у глави 3 објашњене методе на којима алати за предикцију Т-ћелијских епитопа почивају, а након тога, у глави 4, описани су сами алати и начин на који раде. У глави 5 приказана је развијена апликација и њене функционалности. У глави 6 дат је осврт на постигнуте резултате и могућа унапређења.

Глава 2

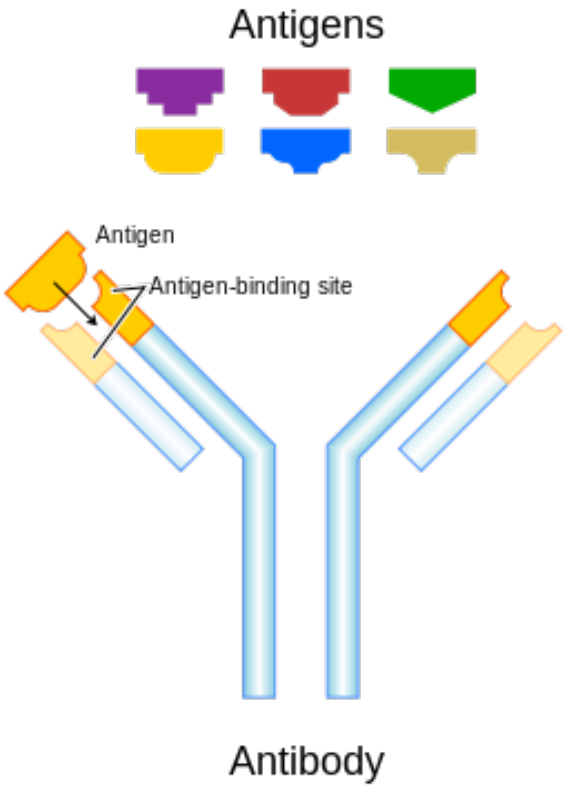
Основи имунологије

У овом поглављу дат је преглед основних имунолошких појмова неопходних за разумевање начина на који функционишу предикциони алати. На самом почетку биће речи о антигенима и антителима, затим о лимфоцитима као најважнијим ћелијама људског имуног система и на крају о протеинима главног комплекса хистокомпатибилности који играју кључну улогу у излагању делова антигена - епитопа на површину ћелије и без којих не би било могуће препознавање епитопа од стране ћелија имуног система.

2.1 Антигени и антитела

Антиген је сваки молекул који може да изазове имуни одговор, односно било која супстанца која може да изазове имуни систем да ствара антитела против ње. Они су по структури протеини, прецизније липопротеини, гликопротеини, нуклеопротеини или полисахариди велике молекулске масе [7]. Антигени су основа стеченог имунитета, јер се он развија тек након иницијалног уласка разних микроорганизама, токсина и других страних тела у организам. Механизам помоћу ког тело препознаје ову иницијалну инвазију заснива се управо на присуству антигена и њиховом препознавању. Као резултат јавља се имуни одговор [7].

Антитело (енгл. *antibody*) је гликопротеин који се производи као резултат стимуланса антигена. Свако антитело је направљено од стране имуног система тако да препознаје антиген након што су ћелије имуног система дошле у контакт с њим [6]. На слици 2.1 може се видети интеракција антигена и антитела.



Слика 2.1: Интеракција антигена и антитела

2.2 Лимфоцити

Да бисмо могли да представимо механизме имуног одговора, неопходно је навести основне карактеристике белих крвних зрнаца, односно лимфоцита, као једној од две основне врсте белих крвних зрнаца. Лимфоцити имају одбрамбену улогу тако што производе антитела и на тај начин учествују у имуном одговору. Постоје три основне групе лимфоцита које се разликују и морфолошки и структурално: Б-ћелије, Т-ћелије и тзв. ћелије природне убице (енгл. *natural killer cells*) [9].

Функција Б- и Т-ћелија је да препознају антигене у току процеса који се назива *излагање антигена*. Једном кад их препознају, ове ћелије генеришу специфичне одговоре који су организовани тако да елиминишу патогене или ћелије које су захваћене патогеном. Б-ћелије реагују на патогене тако што производе велику количину антитела која неутралишу стране објекте као што су бактерије и вируси. Као одговор на патогене, неке Т-ћелије производе цитокине - протеине који регулишу или помажу активни имуни одговор и те ћелије називамо

Т-помоћним ћелијама. С друге стране, постоје и *цитотоксичне Т-ћелије.* Оне испуштају токсична зрнца која садрже моћне ензиме. Ти ензими су способни да униште ћелије заражене неким патогеном или туморске ћелије. Што се ћелија природних убица тиче, о њима се најмање зна и сматра се да би могле бити одговорне за спречавање размножавања страних ћелија, посебно туморских [9].

2.3 Главни комплекс хистокомпатибилности

Како бисмо комплетирали основе имунологије, неопходно је представити и главни комплекс хистокомпатибилности (енгл. *major histocompatibility complex - MHC*). Наиме, ради се о скупу протеина на површини ћелије кодираних великом породицом гена који контролишу главни део имуног система свих кичмењака тако што препознају стране молекуле. Главна функција *MHC* молекула је да се вежу за антигене који се налазе у ћелији и да их прикажу на површини ћелије како би их одговарајуће Т-ћелије препознале. Сваки *MHC* ген има необично велики број *алела.* Алели су различити облици једног истог гена, тако да се из тог разлога каже да су *MHC* гени јако полиморфни - кодирају велики број различитих протеина. Протеини кодирани од стране *MHC* гена се код људи називају *људски леукоцитни антигени* (енгл. *human leucocyte antigen - HLA).*

За потребе овог истраживања од интереса су антигени који су по својој структури протеини, тако да ћемо у наставку, када говоримо о антигенима, заправо подразумевати да говоримо о антигенима који су по свом саставу протеини.

У ћелијама се протеини константно синтетишу и деградирају. Ти протеини могу бити синтетисани од стране самог организма домаћина, али и од стране других биолошких ентитета. Антигени могу бити ћелијски (енгл. *self*) или ванћелијски (енгл. *non self*). Ћелијски антигени су они који се нормално синтетишу у ћелији и нису штетни по ћелију, док су ванћелијски антигени они који припадају нпр. бактеријама или вирусима. Врло је важно да имуни систем може да их разликује, како би знао на које од њих треба да реагује и како не би уништавао здраве ћелије организма домаћина. На површину ћелије никад се не излажу цели антигени, већ само делови антигена, односно кратке пептидне секвенце величине најчешће од 8 до 11 аминокиселина. Сваки *MHC* молекул на ћелијској површини приказује молекуларну фракцију протеина, односно пептид, такозвани *enumon.* Епитоп је дакле, антигенска детерминанта,

односно део антигена за који могу да се вежу антитела, Б- или Т-ћелије.

MHC молекули се деле на 3 класе [20]:

1. *MHC молекули класе I* се могу наћи у свим ћелијама и излажу епитопе цитотоксичним Т-ћелијама. Цитотоксичне Т-ћелије поред својих цитотоксичних Т-ћелијских рецептора експримирају и тзв. *CD8 рецепторе*. Када се Т-ћелија својим CD8 рецептором веже за *MHC* молекул класе I и ако се притом Т-ћелијски цитотоксични рецептор веже за епитоп који *MHC* молекул излаже, у том случају ће цитотоксична Т-ћелија изазвати програмирану смрт ћелије на чијој је површини овај комплекс *MHC* молекул/пептид изложен. *MHC* класа I се код људи састоји од *HLA-A*, *HLA-B* и *HLA-C* гена. Захваљујући информацијама које добијамо секвенцирањем људског генома, егзома и РНК секвенцирањем могуће је са великом прецизношћу добити информације о *HLA* типовима пацијента. Сваки човек носи два алела за сваки од *HLA* типова класе I (*HLA-A*, *HLA-B* и *HLA-C*), тако да свака особа може да експримира шест различитих *MHC* I типова.

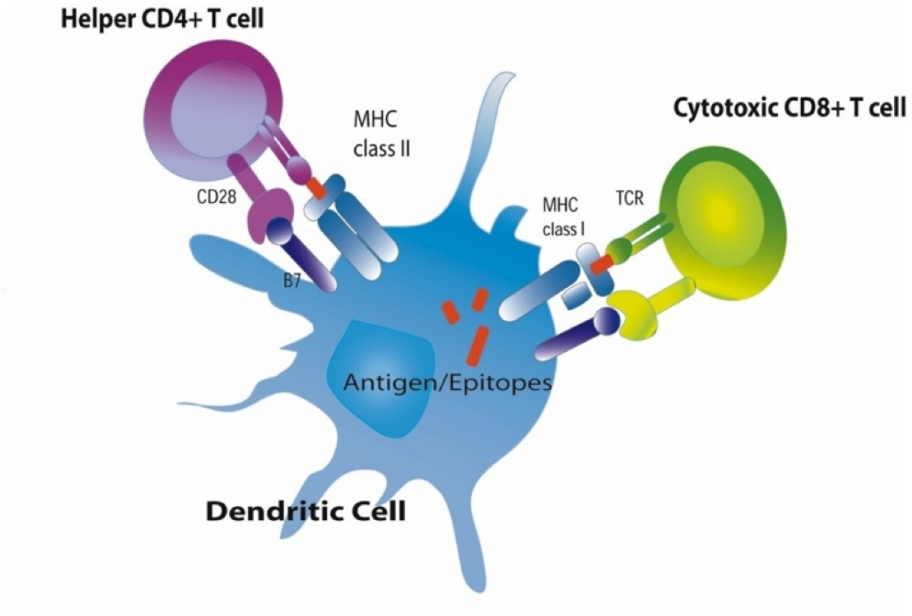
2. *MHC молекули класе II* су присутни само у ћелијама које излажу антигене, нпр. дендритским ћелијама или макрофагама које обавештавају Т-помоћне ћелије да је страно тело ушло у организам, а које затим изазивају активни имуни одговор од стране осталих ћелија имуног система. Они реагују са *CD4 рецепторима* који су присутни на Т-помоћним ћелијама.

3. *MHC молекули класе III* су разни протеини који немају везе са антигенским процесирањем и излагањем.

На слици 2.2 дат је приказ имуног одговора Т-ћелија.

2.4 Процес излагања антигена

Пошто су нам од интереса Т-ћелијски епитопи који се везују за *MHC* молекуле класе I, у наставку ће бити описан начин на који се они излажу на површину ћелије. Важно је напоменути да се овај процес разликује од процеса везивања пептида за *MHC* молекуле класе II.



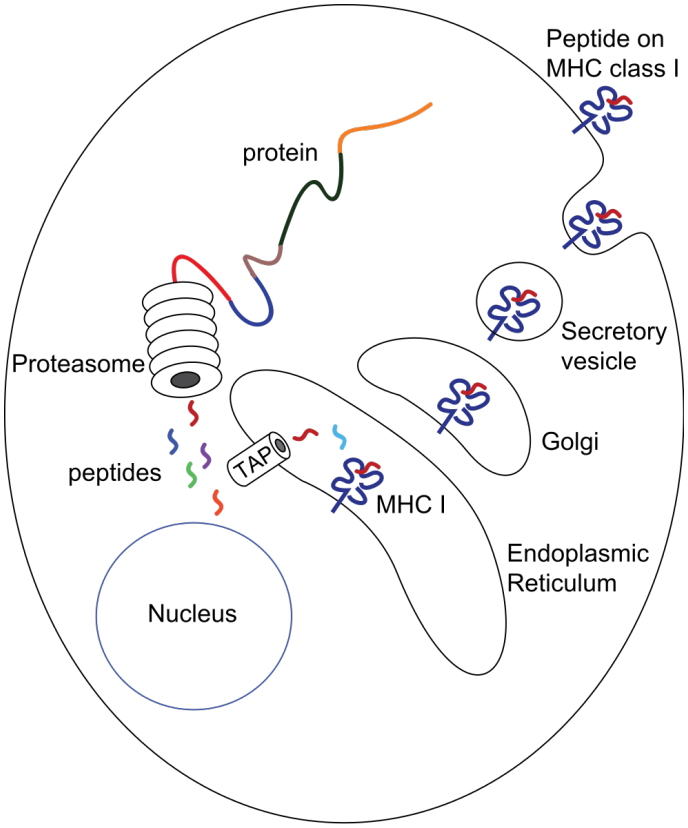
Слика 2.2: Имуни одговор Т-ћелија

Антигене, који су по саставу протеини, протеозоми у ћелијској цитоплазми разграђују на више мањих пептида. Након тога, долази до пребацивања пептида у ендоплазматични ретикулум од стране посебних протеина, тзв. *транспортера повезаних са антигенским процесирањем* (ТАП). У ендоплазматичном ретикулуму пептиди се повезују са *MHC* молекулима класе I који се ту синтетишу. Тај комплекс *MHC*-I/пептид улази у Голџијев апарат где се подвргава одређеним ензиматским процесима након чега га обухвата секреторна везикула која се стапа са ћелијском мембраном и излаже на ћелијској мембрани одакле може да интерагује са Т-ћелијама. На слици 2.3 дат је сликовит приказ излагања антигена посредством *MHC* молекула класе I [10].

Јачина везе између пептида и *MHC* молекула - тзв. афинитет (енгл. *affinity*) се мери помоћу константе дисоцијације (Kd). У хемији, биохемији и фармакологији, константа дисоцијације је специфични тип константе равнотеже који мери склоност објекта да се реверзибилно раздвоји у мање компоненте [8]. Примери дисоцијације су раздвајање хемијског комплекса у молекуларне компоненте. Нека су $[P]$, $[M]$ и $[PM]$ редом моларне концентрације пептида, *MHC* молекула и њиховог комплекса. Тада се одговарајућа константа дисоцијације израчунава као:

$$Kd = \frac{[P][M]}{[PM]}$$

Константа дисоцијације се изражава у моларима (mol/L) и одговара концентрацији лиганда (у овом случају пептида) при којој је концентрација протеина (*MHC* молекула) са везаним пепитидом једнака концентрацији *MHC* молекула без везаног пептида. Што је константа дисоцијације мања, јачи је афинитет између пептида и *MHC* молекула, тј. већа је концентрација комплекса молекула $[PM]$ у односу на концентрације појединачних супстанци $[P]$ и $[M]$. На пример, пептид са наномоларном (nM) константом дисоцијације се јаче везује за одређени *MHC* молекул него пептид са микромоларном (μM) константом дисоцијације [8]. Ова мера се користи и у алатима за предикцију епитопа.



Слика 2.3: Излагање антигена посредством *MHC* молекула класе I

Неоантигени су антигени кодирани од стране *туморских гена*. Када говоримо о неоантигенима, мислимо на протеине који су услед неких мутација променили своју структуру, односно код којих је дошло до промене у једној

или више аминокиселина. Пошто се ради о протеинима који су се раније (пре мутација) нормално синтетисали у ћелијама и као такви нису били предмет препознавања ћелија имуног система, зовемо их новим (*нео*) антигенима. Овакви протеини изложени су од стране молекула *MHC* класе I и *MHC* класе II на површини *туморских* ћелија, тако да цитотоксичне T-ћелије онда могу да препознају овакве антигене и да униште туморске ћелије.

Студије у последњих пар година показују кључну улогу неоантигена у туморској имунотерапији. Најновије технологије секвенцирања људског генома напредују у циљу што бржег идентификовања мутација ткива захваћених тумором, тако да се данас захваљујући постојећим рачунарским алатима може открити много о пептидним секвенцама, односно епитопима који потичу од тако мутиране ДНК. Сазнање о таквим епитопима нам омогућава обећавајућу циљану туморску имунотерапију. У наставку ће бити дат преглед рачунарских метода и техника које детектују ове епитопе.

Глава 3

Предикционе методе

Предикција Т-ћелијских епитопа је нарочит изазов због високог степена полиморфизма *MHC* региона и огромне количине података која је последица различитости и комплексности самих корака при генерисању и излагању Т-ћелијских епитопа о којима је већ било речи у поглављу 2.4. Број познатих *HLA* алела је порастао са 1000 у 1998. години до преко 13 000 у 2015. години [12]. Од огромног броја различитих пептида који могу да буду генерисани од стране једног патогена, јако је мали проценат оних који заиста могу да изазову имуни одговор. Процењује се да је тај број 1 у 2000 или 1 у 5600 [2]. Овако мали број могућих имуногених пептида је последица три кључна корака: сечења протеина и транспорта добијених пептида у ендоплазматични ретикулум, затим везивања за *MHC* молекул и њиховог препознавања од стране цитотоксичних Т-ћелија.

Када је у питању предвиђање епитопа, главни циљ је истражити склоности антигенског пептида ка везивању за *MHC* молекуле (енгл. *binding affinity*), јер је то најрестриктивнији корак од претходна три поменута. Само 1 пептид од 40-200 пептида веже се за специфични *MHC* молекул са довољним афинитетом који може да изазове имуни одговор [2]. Врло је важно напоменути да се сви Т-ћелијски епитопи добро вежу за *MHC* молекуле, али није сваки пептид који се добро веже за *MHC* молекул у исто време и Т-ћелијски епитоп [5].

Експерименталне технике за препознавање Т-ћелијских епитопа су се показале као скупе, временски захтевне и неефикасне и због тога су развијене бројне рачунарске (енгл. *in silico*) методе које могу да моделују имунолошке процесе. Један од начина да се унапреди читав процес откривања потенцијалних епитопа јесте да се прво изврше предикциони алати који ће дати листу највероватнијих

пептида који представљају епитопе, а онда да се експериментално посматра тај мањи скуп потенцијалних епитопа [2].

Развијени су многи алати за предикцију Т-ћелијских епитопа, многи од њих се и даље унапређују, а многи од њих су доступни на различитим веб серверима. Ти алати почивају на више различитих принципа и деле се на *директне* и *индиректне*. Директни алати раде тако што узимају у обзир тродимензионалну структуру протеина или нуклеотидне секвенце протеина. Они посматрају амфипатичност пептида, затим обрасце или мотиве (енгл. *motifs*) - специфичне секвенце које су карактеристичне за места везивања пептида и *MHC* молекула. Проблем са оваквим алатима је тај што нису довољно прецизни и дају велики проценат лажних позитива. С друге стране, индиректне методе се базирају на статистици, неуронским мрежама, методама потпорних вектора, итд. За алгоритме који су засновани на неуронским мрежама показало се да предвиђају епитопе са највећом тачношћу [20].

3.1 Директне методе

У даљем тексту биће приказани најчешћи директни приступи у предикцији Т-ћелијских епитопа. Они се, дакле, ослањају на анализу секвенци аминокиселина из којих се пептиди састоје, као и на анализу структуре пептида за које се врше предикције. Биће речи о амфипатичности, приступу заснованом на мотивима и квантитативним матрицама.

Амфипатичност

Амфипатичност се односи на својство молекула да у исто време испољавају хидрофилна (воле воду, растварају се у води) и хидрофобна (не воле воду, одбијају се од воде) својства. Амфипатични региони могу бити присутни код протеина, као и код пептида. Т-ћелијски епитопи могу формирати амфипатичне структуре код којих се хидрофобични региони периодично понављају, те се та особина може користити при предикцији потенцијалних епитопа. Важно је имати у виду да није свака таква структура потенцијални епитоп [20].

Приступ заснован на мотивима

Овај приступ је најстарији, али и најшире коришћен метод предвиђања епитопа. Базира се на налажењу одређених региона на нуклеотидним секвенцама који садрже позната места везивања *MHC* молекула и пептида. Наиме, пептиди који се везују за *MHC* молекуле садрже одређене аминокиселине на одређеним позицијама и таква места називамо мотивима. Мотив присутан у пептиду који се веже за један *MHC* молекул се може разликовати од мотива на пептиду који се везује за неки други *MHC* молекул. Алати који раде по овом принципу функционишу тако што траже познате мотиве у датој протеинској секвенци, а затим генеришу листу мотива за дати протеин. Такви мотиви се онда кластерују и врши се предикција епитопа у зависности од релативне заступљености мотива у датој протеинској секвенци. Неки алати узимају у обзир и амфипатичност и мотиве. При оваквим предикцијама, могу се узети у обзир и аминокиселине које су непожељне на неким позицијама у протеину. Наиме, постоје алгоритми који раде по принципу бодовања. Свакој аминокиселини на одређеној позицији у оквиру пептида се додељује одређена вредност у зависности од њене фреквенције у експериментално потврђеним епитопима или везујућим пептидима (који не морају нужно да представљају епитопе). Та вредност може да варира од високе позитивне вредности - нпр. 15, која представља идеалну вредност, малих позитивних вредности - нпр. 1, до негативних вредности које означавају да је нека аминокиселина непожељна на тој позицији у пептиду. Вредности на свим позицијама се сумирају и дају крајњу вредност пептида на основу које се даље врши предикција [20].

Квантитативне матрице

Квантитативне матрице (енгл. *quantitative matrix*) приказују квантитативну вредност квалитативних информација/података. Те матрице су заправо тежинске матрице које за сваку аминокиселину на одређеној позицији у пептиду садрже њен допринос у коначној предикционој вредности везивања тог пептида за *MHC* молекул. Такве матрице постоје за различите *MHC* молекуле. Квантитативна матрица садржи вредност која означава утицај (пожељан, неутралан или непожељан) одређене аминокиселине на некој позицији на везивање тог пептида за одређени *MHC* молекул. Алгоритми који користе квантитативне матрице деле протеин на преклапајуће фрагменте (пептиде) одређене дужине.

Ако се, на пример, одабере дужина 10, онда ће свакој аминокиселини у пептиду дужине 10 бити додељен одређени коефицијент у зависности од вредности те аминокиселине на одређеној позицији у квантитативној матрици. Крајња вредност за пептид се добија сабирањем или множењем коефицијената свих аминокиселина у пептиду и за пептиде који имају вредност већу од неког задатог параметра се сматра да ће да се вежу за *MHC* молекул [20]. На слици 3.1 приказана је једна таква матрица.

Amino acid/position	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	0.52	−0.67	−0.25	−0.29	−0.35	−0.55	−0.1	−0.34	−0.05
C	0	−2	−0.4	0.29	1	1.67	1.33	0.67	1
D	−1.6	−2	0.08	0.34	−0.75	−0.86	−0.82	−0.4	−1.69
E	−1.41	−1.64	−1.48	−0.05	−0.43	−0.92	−1.08	−0.04	−2
F	0	−1.08	1.05	−0.4	1.28	0.27	1.39	−0.53	−2
G	0.91	−1.82	−0.47	1.18	0.3	−0.4	−0.11	0.13	−1.82
H	0.22	−2	0.22	0.22	−0.29	−0.5	0.93	−0.22	−2
I	−0.27	0.89	−0.62	−1.09	−0.62	0	−0.27	−0.07	0
K	0.25	−1.47	−1.14	−0.75	−0.77	−1.56	−1.2	−0.63	−1.43
L	0.51	1.62	1.24	−0.29	0.19	0.44	0.38	0.22	1.31
M	−0.67	1.47	0.29	1.43	1.33	1.67	0	0.4	1
N	−0.22	−2	0.29	−1	−1.11	−0.82	−0.22	−0.44	−2
P	−0.5	−2	−0.5	0.59	0.62	0.88	0.17	0.11	−2
Q	−0.75	−1.14	−1.64	0.26	−0.82	−0.35	−0.22	0.33	−1.33
R	0.17	−0.86	−0.29	0.32	−0.11	−1.11	−0.8	−0.15	−1.2
S	0.76	−2	0.4	0.5	0	0.11	−0.53	0.1	−1.08
T	−0.88	−0.75	−0.81	−0.92	−0.5	−0.67	−0.24	0.92	−0.71
V	−0.81	−0.88	0.22	−0.83	0	1.23	0.44	−0.5	1.38
W	−1.38	−1.6	−0.1	−1.64	−0.11	−1.47	−0.86	−1	−2
X	2	2	0	0	0	2	2	0	0
Y	−0.12	−2	0.09	−2	0.43	−0.12	−0.25	0	−1.43

Слика 3.1: Пример квантитативне матрице

3.2 Индиректне методе

Индиректне методе се базирају на машинском учењу и статистици. У даљем тексту ће бити речи о неуронским мрежама и методи потпорних вектора.

Неуронске мреже

Неуронске мреже (енгл. *artificial neural networks*) представљају метод машинског учења способан да уочи нелинеарности међу подацима. Оне могу да буду трениране тако да практично науче карактеристике одговарајућих образаца (тренинг скуп података) и да се након тога користе да препознају сличне обрасце у новим подацима. Неуронска мрежа се састоји од одређеног броја повезаних чворова (процесних елемената) које називамо вештачким неуронима. Чворови могу бити повезани у више слојева. Обично постоји један улазни слој, један или више средишњих (скривених) слојева и један излазни слој. Улазни слој је једини који прима податке из спољашње средине. Следећи (скривени) слој(еви) прослеђују релевантне податке до излазног слоја где добијамо коначан резултат. Средишњи слојеви се најчешће састоје из неурона који могу бити повезани на такав начин да што боље осликавају обрасце у тренинг подацима. Док неуронска мрежа учи обрасце који су присутни у тренинг подацима, она формира одговарајуће везе међу чворовима и додељује им одговарајуће тежине. Неуронска мрежа може успешно да препозна одговарајуће обрасце у тест подацима, али исто тако може и да направи грешке. Те грешке се могу редуковати ажурирањем тежина између чворова мреже. Неуронске мреже се користе у разне сврхе, као што су предикције генских експресија, секундарних структура протеина, Б- и Т-ћелијских епитопа итд. Неуронске мреже које се тренирају у циљу предикције Т-ћелијских епитопа обично као улазне чворове имају секвенцу аминокиселина, а као излазни слој обично имају један чвор који за улазну секвенцу аминокиселина каже да *јесте епитоп* или *није епитоп* [20].

Метод потпорних вектора

Метод потпорних вектора (енгл. *support vector machines*) је још један од метода надгледаног машинског учења са јаком статистичком основом. Овај метод такође учи и препознаје обрасце који су присутни у подацима. За дати скуп тренинг података, од којих сваки податак има обележје категорије којој припада, метод прави модел који нове податке сврстава у једну од категорија. Подаци се представљају као вектори и ако се налазе у дводимензионалном простору, онда се раздвајају правом у две категорије, а ако су пак представљени као тачке у вишедимензионалном простору, онда је потребна хиперраван која ће их раздвојити у категорије. Нови подаци се мапирају у простор истих димензија као и

тренинг подаци и додељује им се категорија у зависности са које стране праве или хиперравни се налазе. Метод потпорних вектора има широку примену у модерној биологији. Користи се, на пример, у анализи генских експресија у нормалним и туморским ћелијама да издвоји гене који се експримирају само у туморским ћелијама. Такође, може се користити и у идентификовању Т-ћелијских епитопа међу многим пептидима који нису епитопи [20].

Глава 4

Предикциони алати

Предиктори за *MHC* молекуле класе I су врло ефикасни, покривају широк скуп *HLA* алела и процењује се да достижу тачност (енгл. *accuracy*) од 90-95%. Пошто је везивање пептида за *MHC* молекул главни фактор који утиче на то да пептид буде имуноген, већина алата се фокусира на ову фазу процесирања пептида. Информације о епитопима који су добијени експерименталним путем користе се за тренирање алгоритма све док алгоритам не достигне максималну ефикасност у предвиђању нових *MHC*-I/пептид структура, односно епитопа. Као што је већ речено, сами алгоритми почивају на различитим принципима, па су самим тим и различите комплексности и тачности. Обично се алгоритми деле на две групе: једноалелски (енгл. *allele-specific*) и вишеалелски (енгл. *pan-specific*). Први се односе на оне код којих је модел трениран независно за сваки *HLA* алел. Овакви приступи су непрактични због сталног откривања нових *HLA* алела. Вишеалелски алгоритми, с друге стране, односе се на оне код којих је модел трениран над скупом података који покрива велики број *HLA* алела. Алгоритми који припадају другој групи су се показали као веома моћни, јер омогућавају предикције за све познате *MHC* молекуле, укључујући и оне за које постоји ограничен број података о везивању са пептидима или их уопште нема. У наставку ће бити дат преглед неколико алата за предикцију Т-ћелијских епитопа [20].

4.1 NetMHC

NetMHC је алат који предвиђа везивање пептида за *MHC* молекуле користећи неуронске мреже. Оне су тренирane за 81 различит људски *HLA* алел,

укључујући *HLA-A*, *HLA-B*, *HLA-C* и *HLA-E* алеле. Такође су могуће предикције за 41 животињски *MHC* алел. Предикције могу да се врше за пептиде дужине од 8-14 аминокиселина, али треба узети у обзир да се *MHC* молекули углавном добро везују за пептиде дужине 9, док предикције за пептиде који су дужи од 11 аминокиселина треба узети са резервом. Метод користи експерименталне податке из *IEDB* базе (познати епитопи и њихове склоности ка везивању са познатим *HLA* типовима) и *SYFPEITHI* базе (колекција *MHC* лиганда и пептидних мотива како за људску, тако и за животињску врсту). Тренинг метод који овај алат користи показао се као најбољи тренутно доступан. Алат је коришћен за предвиђање везивања *MHC* молекула и различитих вирусних пептида укључујући САРС, инфлуенцу и ХИВ и резултат је у просеку био 75-80% експериментално потврђених резултата. Његове перформансе су даље валидиране на новим скуповима података који нису редундантни са подацима над којима је метод треniран. Други *MHC* предиктори су тренирани над пептидима исте дужине као што је она коју предвиђају, но с обзиром да су много ређи подаци за пептиде чија је дужина различита од 9, самим тим је и могућност предикције за пептиде дужине различите од 9 ограничена.

Оно што *NetMHC* разликује од других алата јесте то што он, иако је треniран над скупом пептида који су дужине 9, може да врши предикцију за пептиде дужине од 8 до 14 аминокиселина, што му даје предност у односу на остале *MHC* предикторе. Ипак, *NetMHC* може да врши предикције само за оне *MHC* алеле над којима је треniран [16].

На слици 4.1 приказан је излаз из *NetMHC-a*. За сваки пептид израчунат је његов афинитет ка задатом *HLA* типу изражен у наномоларима (nM). Колона *affinity* се заправо односи на константу дисоцијације дефинисану у поглављу 2.4. Дакле, што је мања константа дисоцијације, пептид се јаче веже за *MHC* молекул. Подразумевана горња граница за вредност *affinity* колоне за пептиде који се добро вежу за *MHC* је 50, а за пептиде који се слабо вежу та вредност је 500. Јаки и слаби епитопи се одређују у односу на дистрибуцију афинитета израчунатих за 400 000 случајно одабраних пептида. Позиција епитопа у растуће сортираном низу ових вредности се назива ранг и он се користи као мера јачине епитопа. Епитоп се дефинише као јак уколико се вредност његовог афинитета налази у првих 0,5% свих израчунатих вредности, док се епитоп сматра слабим ако је његов афинитет између 0,5 и 2% свих израчунатих вредности. Колона $1 - \log_{50000}(aff)$ је само логаритамски трансформисана вредност афинитета и

представља ништа друго него скалиран афинитет на интервалу [0,1]. Колона *Bindlevel* показује на јаке/слабе епитопе.

```
--------------------------------------------------------------------------------
Milicas-MacBook-Pro:netMHC-4.0 milicakojicic$ ./netMHC test/test.fsa -l 9 -a HLA-B2705 -s
# /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHC-4.0/Darwin_x86_64/bin/netMHC test/test.fsa -l 9 -a HLA-B2705 -s
# Tue Sep 12 22:32:34 2017
# User: milicakojicic
# PWD : /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHC-4.0
# -l       9                   Peptide length (multiple lengths separated by comma e.g. 8,9,10)
# -a       HLA-B2705           HLA allele name
# -s       1                   Sort output on decreasing affinity
# Command line parameters set to:
#       [-a line]            HLA-B2705           HLA allele name
#       [-f filename]                            Input file (by default in FASTA format)
#       [-p]                 0                   Switch on if input is a list of peptides (Peptide format)
#       [-l string]          9                   Peptide length (multiple lengths separated by comma e.g. 8,9,10)
#       [-s]                 1                   Sort output on decreasing affinity
#       [-rth float]         0.500000            Threshold for high binding peptides (%Rank)
#       [-rlt float]         2.000000            Threshold for low binding peptides (%Rank)
#       [-listMHC]           0                   Print list of alleles included in netMHC
#       [-xls]               0                   Save output to xls file
#       [-xlsfile filename]  NetMHC_out.xls      File name for xls output
#       [-t float]           -99.900002          Threshold for output
#       [-thrfmt filename]   /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHC-4.0/Darwin_x86_64/data/threshold/%s.thr Format for threshold filenames
#       [-hlalist filename]  /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHC-4.0/Darwin_x86_64/data/allelelist File with covered HLA names
#       [-rdir filename]     /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHC-4.0/Darwin_x86_64 Home directory for NetMHC
#       [-tdir filename]     /var/folders/ks/kkpvbs_x7yl73v797cpfl4br0000gn/T/ Temporary directory (Default $$)
#       [-syn filename]      /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHC-4.0/Darwin_x86_64/data/synlists/%s.synlist Format of synlist file
#       [-v]                 0                   Verbose mode
#       [-dirty]             0                   Dirty mode, leave tmp dir+files
#       [-inptype int]       0                   Input type [0] FASTA [1] Peptide
#       [-version filename]  /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHC-4.0/Darwin_x86_64/data/version File with version information
#       [-w]                 0                   w option for webface

# NetMHC version 4.0

# Read 132 elements on pairlist /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHC-4.0/Darwin_x86_64/data/allelelist
# Input is in FSA format

# Peptide length 9
# Rank Threshold for Strong binding peptides   0.500
# Rank Threshold for Weak binding peptides   2.000
-----------------------------------------------------------------------------------
  pos          HLA      peptide         Core Offset  I_pos  I_len  D_pos  D_len        iCore        Identity 1-log50k(aff) Affinity(nM)    %Rank  BindLevel
-----------------------------------------------------------------------------------
  166    HLA-B2705    IRLGLALNF    IRLGLALNF      0      0      0      0      0    IRLGLALNF 143B_BOVIN_P293         0.666        37.18     0.15 <= SB
   55    HLA-B2705    RRSSWRVIS    RRSSWRVIS      0      0      0      0      0    RRSSWRVIS 143B_BOVIN_P293         0.511       197.96     0.80 <= WB
  126    HLA-B2705    FRYLSEVAS    FRYLSEVAS      0      0      0      0      0    FRYLSEVAS 143B_BOVIN_P293         0.362       998.75     2.50
   54    HLA-B2705    ARRSSWRVI    ARRSSWRVI      0      0      0      0      0    ARRSSWRVI 143B_BOVIN_P293         0.345      1193.05     2.50
   40    HLA-B2705    ERNLLSVAY    ERNLLSVAY      0      0      0      0      0    ERNLLSVAY 143B_BOVIN_P293         0.318      1597.37     3.00
  221    HLA-B2705    LRDNLTLWT    LRDNLTLWT      0      0      0      0      0    LRDNLTLWT 143B_BOVIN_P293         0.285      2293.84     3.50
   98    HLA-B2705    LQLLDKYLI    LQLLDKYLI      0      0      0      0      0    LQLLDKYLI 143B_BOVIN_P293         0.282      2361.09     3.50
   59    HLA-B2705    WRVISSIEQ    WRVISSIEQ      0      0      0      0      0    WRVISSIEQ 143B_BOVIN_P293         0.243      3614.25     4.50
   82    HLA-B2705    YREKIEAEL    YREKIEAEL      0      0      0      0      0    YREKIEAEL 143B_BOVIN_P293         0.232      4078.94     5.00
```

Слика 4.1: Приказ командне линије и излаза из *NetMHC-a*

Верзија *NetMHC-a* која је коришћена у веб апликацији је *4.0*. Као улаз, алат може да прими секвенце протеина у *FASTA* формату или секвенце пептида које морају да буду исте дужине. *FASTA* формат је текстуални формат за представљање нуклеотидних или пептидних секвенци где су нуклеотиди или аминокиселине представљене једним словом. Формат дозвољава коментаре који претходе секвенцама и обично представљају имена секвенци. Пептидни формат је у овом случају само низ пептида исте дужине записаних један испод другог. Пример *FASTA* и пептидног улаза се може видети на фрагментима кода 4.1 и 4.2.

```
>sekvenca 1
ASTPGHTIIYEAVCLHNDRTTIP
>sekvenca 2
ASQKRPSQRHGSKYLATASTMDHARHGFLPRHRDTGILDSIGRFFGGDRGAPK
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKD
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQ
```

Код 4.1: *FASTA* формат

```
ILYQVPFSV
VVMGTLVAL
ILDEAYVMA
KILSVFFLA
```

Код 4.2: Пример пептидног формата

При покретању алата могуће је одабрати пептидну дужину (која може бити 8-14), као и *HLA* тип за који се врши предикција. Ако се не задају ови параметри, подразумевани параметри су 9 за пептидну дужину и HLA-A0201 за *HLA* тип. Такође, могуће је задати и параметре који се односе на горњу границу за пептиде који се добро вежу за дати *MHC* молекул, као и горњу границу за пептиде који се слабије вежу и њих корисник може подешавати по потреби. Оба параметра се односе на колону *%Rank*. Као што је претходно речено, подразумевана горња граница за пептиде који се добро вежу је 0,5, док је та граница за пептиде који се слабије вежу 2. И коначно, могуће је сортирати и филтрирати пептиде по афинитету.

4.2 NetMHCpan

За највећи број *HLA* алела и даље не постоје експериментални подаци о везивању. *NetMHCpan* је алат који омогућава предикције везивања пептида за *било који MHC* молекул чија је секвенца позната користећи неуронске мреже (дакле, не само за оне над којима је трениран). Метод је трениран на преко 180 000 квантитативних података који покривају 172 *HLA* алела код човека (*HLA-A*, *HLA-B*, *HLA-C* и *HLA-E*), као и на одређеном броју животињских *MHC* алела. Омогућено је предвиђање за пептиде дужине 8-14 аминокиселина [1].

На слици 4.2 приказан је излаз из *NetMHCpan-a*. Параметри за покретање *NetMHCpan-a* су исти као и код *NetMHC-a*, осим што *NetMHCpan* може да ради над ширим скупом *HLA* типова. Излази се подударају са излазима из *NetMHC-a*.

4.3 NetCTL

NetCTL је алат за предвиђање епитопа који се везују за *MHC* молекуле класе I и интегрише три кључна корака у излагању антигена на површину ћелије, а о којима је било речи у поглављу 2.4: деградацију протеина на мање

```
Milicas-MacBook-Pro:netMHCpan-3.0 milicakojicic$ ./netMHCpan -p test/test.pep -l 11 -a HLA-A02:01 -s
# /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHCpan-3.0/Darwin_x86_64/bin/netMHCpan -p test/test.pep -l 11 -a HLA-A02:01 -s
# Tue Sep 12 23:11:40 2017
# User: milicakojicic
# PWD : /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHCpan-3.0
# Host: Darwin Milicas-MacBook-Pro.local 16.7.0 x86_64
# -p       1                    Use peptide input
# -l       11                   Peptide length [8-11] (multiple length with ,)
# -a       HLA-A02:01           HLA allele
# -s       1                    Sort output on descending affinity
# Command line parameters set to:
#       [-rdir filename]     /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHCpan-3.0/Darwin_x86_64 Home directory for NetMHpan
#       [-syn filename]      /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHCpan-3.0/Darwin_x86_64/data/synlist Synaps file
#       [-v]                 0                    Verbose mode
#       [-dirty]             0                    Dirty mode, leave tmp dir+files
#       [-tdir filename]     /var/folders/ks/kkpvbs_x7yl73v797cpfl4br0000gn/T//netMHCpanXXXXXX Temporary directory (made with mkdtemp)
#       [-hlapseudo filename] /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHCpan-3.0/Darwin_x86_64/data/MHC_pseudo.dat File with HLA pseudo sequences
#       [-hlaseq filename]                        File with full length HLA sequences
#       [-a line]            HLA-A02:01           HLA allele
#       [-f filename]                             File name with input
#       [-w]                 0                    w option for webface
#       [-s]                 1                    Sort output on descending affinity
#       [-p]                 1                    Use peptide input
#       [-rth float]         0.500000             Rank Threshold for high binding peptides
#       [-rlt float]         2.000000             Rank Threshold for low binding peptides
#       [-l string]          11                   Peptide length [8-11] (multiple length with ,)
#       [-xls]               0                    Save output to xls file
#       [-xlsfile filename]  NetMHCpan_out.xls    Filename for xls dump
#       [-t float]           -99.900002           Threshold for output
#       [-thrfmt filename]   /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHCpan-3.0/Darwin_x86_64/data/threshold/%s.thr Format for threshold filenames
#       [-expfix]            0                    Exclude prefix from synlist
#       [-version filename]  /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHCpan-3.0/Darwin_x86_64/data/version File with version information
#       [-inptype int]       0                    Input type [0] FASTA [1] Peptide
#       [-listMHC]           0                    Print list of alleles included in netMHCpan
#       [-allname filename]  /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netMHCpan-3.0/Darwin_x86_64/data/allelenames File with print names for alleles

# NetMHCpan version 3.0

# Tmpdir made /var/folders/ks/kkpvbs_x7yl73v797cpfl4br0000gn/T//netMHCpanganR5y
# Input is in PEPTIDE format

HLA-A02:01 : Distance to training data  0.000 (using nearest neighbor HLA-A02:01)

# Rank Threshold for Strong binding peptides   0.500
# Rank Threshold for Weak binding peptides   2.000
-----------------------------------------------------------------------------------
  Pos          HLA         Peptide       Core Of Gp Gl Ip Il        Icore        Identity   Score Aff(nM)   %Rank  BindLevel
-----------------------------------------------------------------------------------
    1  HLA-A*02:01       AAAWYLWEV  AAAWYLWEV  0  0  0  0  0    AAAWYLWEV         PEPLIST 0.82824     6.4    0.06 <= SB
    1  HLA-A*02:01       AEFGPWQTV  AEFGPWQTV  0  0  0  0  0    AEFGPWQTV         PEPLIST 0.14872 10003.0   13.00
    1  HLA-A*02:01       AASKQQMLM  AASKQQMLM  0  0  0  0  0    AASKQQMLM         PEPLIST 0.06531 24664.2   29.00
```

Слика 4.2: Приказ командне линије и излаза из *NetMHCpan-a*

фрагменте - пептиде, ТАП транспорт пептида и везивање *MHC* молекула и пептида. Предикција везивања пептида за *MHC* молекул класе I заснована је на претходно описаном *NetMHC* алату, док се предикција ефикасности ТАП транспортног канала врши коришћењем квантитативних матрица. Предикција деградације протеина у протеозомама се врши алатом *NetChop*. То је алат који је такође базиран на неуронским мрежама и који је трениран да препознаје места сечења протеина у протеозомама. Коначна предикциона вредност *NetCTL* алата добија се као тзв. тежинска сума појединачних предикционих вредности у сва три главна корака. Овде је могућа предикција везивања пептида за неки од 12 *HLA* супертипова и то: A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58 и B62 и подржана је предикција само за пептиде дужине 9. Метод је трениран на 886 познатих лиганда *MHC* класе I. У предвиђању епитопа могу да се поставе различите границе за сваку од вредности ових појединачних корака. Рађено је детаљно упоређивање алата са другим алатима из научне заједнице (*WAPP*, *EpiJen*, *MHC-pathway*) и показано је да *NetCTL* има већу предикциону моћ од поменутих алата [17].

4.4 NetCTLpan

NetCTLpan је унапређена и проширена верзија *NetCTL-a*. Из тог разлога је само он укључен у веб апликацију. То је алат који предвиђа епитопе који се везују за *MHC* молекуле класе I и такође интегрише три кључна корака у излагању антигена на површину ћелије: деградацију протеина на мање фрагменте - пептиде, ТАП транспорт пептида и везивање *MHC* молекула и пептида. Коначна предикциона вредност *NetCTLpan* алата се добија као тзв. тежинска сума појединачних предикционих вредности у сва три главна корака. За сваки од ова три корака користе се исти алати као и у *NetCTL* алату, осим што се за предикцију везивања пептида за *MHC* молекул класе I користи *NetMHCpan* алат. *NetCTLpan* може да врши предикцију за пептиде дужине 8, 9, 10 и 11 аминокиселина.

Како би се што више смањио број лажних позитива, метод је оптимизован тако да постиже што већу специфичност (мера пропорције негативних исхода који су коректно идентификовани као такви, нпр. проценат неепитопа који су коректно идентификовани као неепитопи). Такав приступ, с друге стране, води до потенцијалног губитка у сензитивности (пропорција позитивних исхода који су коректно идентификивани као такви, нпр. проценат епитопа који су коректно идентификовани као епитопи).

Метод је трениран и валидиран над великим скуповима података и експериментално идентификованим пептидима који се везују за *MHC* молекуле класе I и Т-ћелијским епитопима. Предиктивна моћ *NetCTLpan* алата показала се бољом у односу на све постојеће алате за предикцију Т-ћелијских епитопа. У односу на *NetCTL* и *NetMHCpan* овај алат може да смањи експериментални труд за проналажење епитопа за 15% и 40%, респективно [18]. Све претходне методе су ограничене чињеницом да врше предикцију везивања пептида за врло ограничен скуп различитих *MHC* молекула класе I, док *NetCTLpan* врши предикцију за све *MHC* молекуле класе I чије су протеинске секвенце познате. Скуп *MHC* молекула класе I са којима *NetCTLpan* ради константно се повећава како се базе које их складиште допуњавају (*IMGT/HLA* и *IPD-MHC*).

На слици 4.3 приказан је излаз из *NetCTLpan-a*. Колоне *Cle*, *TAP* и *MHC* се односе редом на скорове сваког од поменутих корака (деградацију протеина на пептиде, ТАП транспорт пептида и везивање *MHC* молекула и пептида). Колона *Comb* представља комбиновани предикциони скор који комбинује вредно-

сти претходне три колоне. Што се улазних параметара тиче, слични су улазима претходних алата. *NetCTLpan* нуди додатну могућност да му се задају тежине које се дају сваком од корака у комбинованој предикционој вредности. Такође, могуће је сортирати пептиде по свакој од вредности појединачних корака, као и по комбинованој предикционој вредности, по којој је такође омогућено и филтрирање пептида.

```
Milicas-MacBook-Pro:netCTLpan-1.1 milicakojicic$ ./netCTLpan test/test.fsa -s 0 -a HLA-A02:01
# /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netCTLpan-1.1/Darwin_x86_64/bin/netCTLpan test/test.fsa -s 0 -a HLA-A02:01
# Tue Sep 12 23:16:45 2017
# User: milicakojicic
# PWD : /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netCTLpan-1.1
# -s       0                   Sort output on score: 0 [comb], 1 [MHC], 2 [Cle], 3 [TAP] <0 No sort
# -a       HLA-A02:01          HLA allele
# Command line parameters set to:
#       [-rdir filename]     /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netCTLpan-1.1/Darwin_x86_64 Home directory for NetMHpan
#       [-c filename]        /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netCTLpan-1.1/Darwin_x86_64/bin/clepred Cleavage prediction code
#       [-t filename]        /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netCTLpan-1.1/Darwin_x86_64/bin/tapmat_pred_fsa Tap prediciton code
#       [-m filename]        /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netCTLpan-1.1/netMHCpan-2.3//netMHCpan MHC binding prediction code
#       [-v]                 0                    Verbose mode
#       [-dirty]             0                    Dirty mode, leave tmp dir+files
#       [-tdir filename]     /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netCTLpan-1.1/tmp Temporary directory (Default $$)
#       [-hlaseq filename]                        File with full length HLA sequences
#       [-a line]            HLA-A02:01           HLA allele
#       [-f filename]                             File name with input
#       [-s int]             0                    Sort output on score: 0 [comb], 1 [MHC], 2 [Cle], 3 [TAP] <0 No sort
#       [-l int]             9                    Peptide length [8-11]
#       [-xls]               0                    Save output to xls file
#       [-xlsfile filename]  NetCTLpan_out.xls    Filename for xls dump
#       [-thr float]         -99.900002           Threshold for output
#       [-listMHC]           0                    Print list of alleles included in netMHCpan
#       [-thrfmt filename]   /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netCTLpan-1.1/Darwin_x86_64/data/threshold/%s.thr Format for threshold filenam
es
#       [-wt float]          0.025000             Weight of tap
#       [-wc float]          0.225000             Weight of Clevage
#       [-ethr float]        1.000000             Threshold for epitopes
#       [-version filename]  /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/netCTLpan-1.1/Darwin_x86_64/data/version File with version information

# NetCTLpan version 1.1


# Peptide length 9
# NetCTLpan predictions for HLA-A*02:01 allele.

#  N   Sequence Name       Allele      Peptide      MHC      TAP      Cle     Comb  %Rank
  219 143B_BOVIN_(P29  HLA-A*02:01    QLLRDNLTL  0.42500  1.04100  0.97391  0.67015   3.00
  195 143B_BOVIN_(P29  HLA-A*02:01    AFDEAIAEL  0.40200  1.10000  0.97574  0.64904   3.00
   43 143B_BOVIN_(P29  HLA-A*02:01    LLSVAYKNV  0.47200  0.47500  0.48171  0.59226   4.00
  171 143B_BOVIN_(P29  HLA-A*02:01    ALNFSVFYY  0.23200  3.04200  0.97083  0.52649   5.00
   57 143B_BOVIN_(P29  HLA-A*02:01    SSWRVISSI  0.24900  0.93200  0.96898  0.49032   6.00
  160 143B_BOVIN_(P29  HLA-A*02:01    MQPTHPIRL  0.23400  0.89500  0.95489  0.47122   6.00
   28 143B_BOVIN_(P29  HLA-A*02:01    AVTEQGHEL  0.20300  1.30200  0.97058  0.45393   6.00
   97 143B_BOVIN_(P29  HLA-A*02:01    VLQLLDKYL  0.33500  0.86500  0.35484  0.43646   7.00
   46 143B_BOVIN_(P29  HLA-A*02:01    VAYKNVVGA  0.26000 -0.18000  0.79787  0.43502   7.00
  169 143B_BOVIN_(P29  HLA-A*02:01    GLALNFSVF  0.16700  2.44000  0.88531  0.42720   7.00
```

Слика 4.3: Приказ командне линије и излаза из *NetCTLpan-a*

4.5 PickPocket

С обзиром да постоје на хиљаде *MHC* алела, тешко је експериментално потврдити вероватноће везивања пептида за сваки *MHC* молекул. *PickPocket* је метод који може да на основу познатих података врши предвиђање везивања лиганда за оне *MHC* молекуле за које не постоје експериментални подаци.

За сваки део пептидног лиганда, тзв. резидуе, користимо информације о сличним пептидним деловима за које знамо како се вежу за који *MHC* молекул. Из скупа *MHC* молекула са познатим лигандима, прави се библиотека која се састоји од матрица везивања. Свака од матрица из библиотеке садржи

вероватноће везивања *MHC* молекула за сваку аминокиселину у пептидном лиганду. Како би се конструисала матрица везивања за било који *MHC* молекул (за ког, потенцијално, не постоје експериментални подаци), упоређује се сваки део тог *MHC* молекула са деловима *MHC* молекула из овако направљене библиотеке. Најједноставнија имплементација претпоставља да ће део датог *MHC* молекула за сваку резидуу пептидног лиганда имати исту вредност везивања као њему најсличнији део познатих *MHC* молекула из библиотеке. У напреднијим имплементацијама, вредност везивања се рачуна као тежински просек базиран на сличностима ка свим деловима *MHC* молекула из библиотеке. У оба случаја, предиктоване вредности за сваки део *MHC* молекула се комбинују у матрицу вредности и ова матрица може да се користи за предвиђање нових лиганда за тај *MHC* молекул.

На слици 4.4 приказан је излаз из *PickPocket-a*.

```
Milicas-MacBook-Pro:pickpocket-1.1 milicakojicic$ ./PickPocket test/test.fsa -s
# /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/pickpocket-1.1/Darwin_x86_64/bin/pickpocket_pmbec test/test.fsa -s -d /Users/milicakojicic/Documents/ALAT
I_ZA_PREDIKCIJU/pickpocket-1.1/Darwin_x86_64/data/PMBEC.lib
# Tue Sep 12 23:18:52 2017
# User: milicakojicic
# PWD : /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/pickpocket-1.1
# -s       1                   Sort output on descending affinity
# -d       /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/pickpocket-1.1/Darwin_x86_64/data/PMBEC.lib Matrix definitions
# Command line parameters set to:
#       [-rdir filename]     /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/pickpocket-1.1/Darwin_x86_64 Home directory for NetMHpan
#       [-v]                 0                   Verbose mode
#       [-dirty]             0                   Dirty mode, leave tmp dir+files
#       [-tdir filename]     /var/folders/ks/kkpvbs_x7yl73v797cpfl4br0000gn/T/ Temporary directory (Default $$)
#       [-hlapseudo filename] /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/pickpocket-1.1/Darwin_x86_64/data/MHC_pseudo.dat File with HLA pseudo sequenc
es
#       [-hlaseq filename]                       File with full length HLA sequences
#       [-a line]            HLA-A02:01          HLA allele
#       [-f filename]                            File name with input
#       [-w]                 0                   w option for webface
#       [-s]                 1                   Sort output on descending affinity
#       [-p]                 0                   Use peptide input
#       [-l int]             9                   Peptide length [8-11]
#       [-inptype int]       0                   Input type [0] FASTA [1] Peptide
#       [-listMHC]           0                   Print list of alleles included in PickPocket
#       [-c filename]        /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/pickpocket-1.1/Darwin_x86_64/data/contact.txt Contacts
#       [-d filename]        /Users/milicakojicic/Documents/ALATI_ZA_PREDIKCIJU/pickpocket-1.1/Darwin_x86_64/data/PMBEC.lib Matrix definitions
#       [-pow float]         10.000000           Power for scoring function
# Input is in FSA format

# Peptide length 9
-----------------------------------------------------------------------------
  pos          HLA     peptide        Identity 1-log50k(aff)
-----------------------------------------------------------------------------
    0  HLA-A*02:01  QLLRDNLTL 143B_BOVIN_P293        0.560
    0  HLA-A*02:01  LLSVAYKNV 143B_BOVIN_P293        0.545
    0  HLA-A*02:01  YLIPNATQP 143B_BOVIN_P293        0.543
    0  HLA-A*02:01  ELQDICNDV 143B_BOVIN_P293        0.534
    0  HLA-A*02:01  LGLALNFSV 143B_BOVIN_P293        0.510
    0  HLA-A*02:01  DMAAAMKAV 143B_BOVIN_P293        0.508
    0  HLA-A*02:01  LLDKYLIPN 143B_BOVIN_P293        0.498
    0  HLA-A*02:01  VLQLLDKYL 143B_BOVIN_P293        0.492
    0  HLA-A*02:01  IMQLLRDNL 143B_BOVIN_P293        0.466
    0  HLA-A*02:01  LQLLDKYLI 143B_BOVIN_P293        0.465
    0  HLA-A*02:01  AFDEAIAEL 143B_BOVIN_P293        0.455
    0  HLA-A*02:01  NLLSVAYKN 143B_BOVIN_P293        0.454
    0  HLA-A*02:01  YLSEVASGD 143B_BOVIN_P293        0.436
```

Слика 4.4: Приказ командне линије и излаза из *PickPocket-a*

PickPocket се показао као врло прецизан метод за предикцију везивања пептида над широким скупом *MHC* алела, како за људску, тако и за животињске врсте. Такође се показао као врло робустан метод у случајевима када је сличност неког новог *MHC* молекула са *MHC* молекулима за које су познате вредности о везивању за пептиде мала, за разлику од метода који почивају на не-

уронским мрежама. Консензус метод који комбинује *PickPocket* и *NetMHCpan* је показао супериорне предикционе перформансе [14].

4.6 NetMHCcons

Тачност свих предикционих метода највише зависи од тога колико је експерименталних података о специфичностима везивања неког *MHC* молекула доступно. Показано је да консензус метод као комбинација два или више алата најбоље ради у пракси [13]. С обзиром на постојање великог броја предикционих алата, неком крајњем кориснику тих алата који није експерт је врло тешко да одабере најбољи алат за одређени *MHC* молекул.

На слици 4.5 приказан је излаз из *NetMHCcons-a*.

```
Milicas-MacBook-Pro:netMHCcons-1.1 milicakojicic$ ./netMHCcons -f test/test.fsa -s -a HLA-A02:01
# Method: NetMHCcons

# Input is in FASTA format

# Peptide length 9

# Threshold for Strong binding peptides (IC50)  50.000 nM
# Threshold for Weak binding peptides (IC50)    500.000 nM

# Threshold for Strong binding peptides (%Rank) 0.5%
# Threshold for Weak binding peptides (%Rank)   2%

# Allele: HLA-A02:01

# Distance to the nearest neighbour ( HLA-A02:01 ) in the training set: 0.000

# NetMHCcons = NetMHC+NetMHCpan

-----------------------------------------------------------------------------------
   pos       Allele         peptide         Identity 1-log50k(aff) Affinity(nM)  %Rank   BindingLevel
-----------------------------------------------------------------------------------
   219   HLA-A02:01       QLLRDNLTL  143B_BOVIN_P293         0.465       328.32   4.00  <=WB
    43   HLA-A02:01       LLSVAYKNV  143B_BOVIN_P293         0.407       608.33   5.00
    98   HLA-A02:01       LQLLDKYLI  143B_BOVIN_P293         0.343      1222.43   6.00
    97   HLA-A02:01       VLQLLDKYL  143B_BOVIN_P293         0.343      1222.43   6.00
   168   HLA-A02:01       LGLALNFSV  143B_BOVIN_P293         0.318      1610.82   7.00
   217   HLA-A02:01       IMQLLRDNL  143B_BOVIN_P293         0.283      2327.08   8.00
   195   HLA-A02:01       AFDEAIAEL  143B_BOVIN_P293         0.282      2352.40   8.00
   160   HLA-A02:01       MQPTHPIRL  143B_BOVIN_P293         0.270      2678.54   8.00
    57   HLA-A02:01       SSWRVISSI  143B_BOVIN_P293         0.266      2812.19   9.00
    28   HLA-A02:01       AVTEQGHEL  143B_BOVIN_P293         0.238      3786.74  10.00
   100   HLA-A02:01       LLDKYLIPN  143B_BOVIN_P293         0.234      3954.23  10.00
   174   HLA-A02:01       FSVFYYEIL  143B_BOVIN_P293         0.229      4174.04  10.00
    89   HLA-A02:01       ELQDICNDV  143B_BOVIN_P293         0.228      4219.45  10.00
   171   HLA-A02:01       ALNFSVFYY  143B_BOVIN_P293         0.215      4883.05  15.00
   104   HLA-A02:01       YLIPNATQP  143B_BOVIN_P293         0.213      4962.95  15.00
    46   HLA-A02:01       VAYKNVVGA  143B_BOVIN_P293         0.213      4989.87  15.00
    21   HLA-A02:01       DMAAAMKAV  143B_BOVIN_P293         0.190      6399.79  15.00
    90   HLA-A02:01       LQDICNDVL  143B_BOVIN_P293         0.186      6646.79  15.00
   128   HLA-A02:01       YLSEVASGD  143B_BOVIN_P293         0.185      6755.55  15.00
   167   HLA-A02:01       RLGLALNFS  143B_BOVIN_P293         0.180      7092.61  15.00
   147   HLA-A02:01       QAYQEAFEI  143B_BOVIN_P293         0.178      7287.08  15.00
   173   HLA-A02:01       NFSVFYYEI  143B_BOVIN_P293         0.172      7733.88  15.00
   158   HLA-A02:01       KEMQPTHPI  143B_BOVIN_P293         0.169      8032.38  15.00
    93   HLA-A02:01       ICNDVLQLL  143B_BOVIN_P293         0.165      8342.39  15.00
```

Слика 4.5: Приказ командне линије и излаза из *NetMHCcons-a*

NetMHCcons управо омогућава осврт на овај проблем и имплементира метод који за сваки *MHC* молекул аутоматски даје оптималну комбинацију предикционих метода, што омогућава крајњем кориснику прецизну предикцију. Ради се о консензус методи која укључује следеће алате: *NetMHC*, *NetMHCpan* и *PickPocket*. Показано је да комбинација *NetMHC* и *NetMHCpan* алата даје нај-

боље перформансе када се ради о предикцији везивања за алеле који постоје у тренинг скупу и за које постоје подаци о везивању са бар 50 лиганда, од којих се бар 10 са великом вероватноћом заиста и вежу за тај алел [13]. У супротном, *NetMHCpan* је најбољи алат. Када нема података за неки *MHC* алел, перформансе алата зависе од растојања тог алела од тренинг података. *NetMHCpan* је показао најбоље перформансе када се блиски суседи датом алелу налазе у тренинг скупу, док комбинација *NetMHCpan* и *PickPocket* алата превазилази моћ оба појединачна алата када су суседи тог алела удаљенији [13]. Очигледно је да се не могу сва три алата користити за произвољан *MHC* алел. На пример, *NetMHC* алат је могуће користити само ако је тај алел део тренинг скупа над којим је овај метод трениран, док је друга два алата могуће користити за било који *MHC* алел чија је протеинска секвенца позната. У овом консензус методу комбинација два или више алата биће укључена у крајњи резултат само ако је показано да је боља од сваког појединачног алата при истим условима. Овај алат илуструје како интеграција потпуно различитих алгоритамских приступа може да доведе до унапређене предикције.

Глава 5

Опис веб апликације

Централна тема овог рада јесте развој апликације која би корисницима омогућила обједињено коришћење претходно описаних алата кроз удобни веб интерфејс. Као улаз, апликација користи секвенце протеина, а као излаз, на основу одабраног предиктора, приказује који су делови протеина препознати као епитопи и са којим скором. У оквиру апликације је такође имплементиран и нови алат за метапредикцију епитопа. Метапредиктор користи предикције више постојећих алата и доноси коначну одлуку на основу различитих гласачких шема (консензуса).

Како би овај циљ био остварен, било је неопходно реализовати неколико задатака. Први задатак је подразумевао детаљно упознавање са начином на који описани алати раде, а затим и осмишљавање начина на који би нови метапредиктор функционисао. Даље, било је потребно одабрати погодни радни оквир (енгл. *framework*). И на крају, не мање важно, неопходно је било пронаћи и прилагодити одговарајуће експерименталне податке над којима би појединачни алати и метапредиктор могли да оперишу.

Пројекат је отвореног кода и може се преузети на адреси [21]. С друге стране, алати који су укључени у апликацију су бесплатни за коришћење само академским корисницима и не могу се користити у комерцијалне сврхе. Из тог разлога они се не налазе на наведеној адреси заједно са кодом пројекта. На адреси [21] дато је такође и упутство на који начин академски корисници могу да преузму алате, а затим и како да их прилагоде да би апликација могла да ради.

При развоју апликације коришћен је *Django web framework* [4] чији ће приказ бити дат у наставку.

5.1 Django

Django је веб радни оквир отвореног кода (енгл. *open source*) и доступан је под БСД лиценцом. Сам *Django* не би био могућ без других пројеката отвореног кода као што су *Apache*, *Python* и *PostgreSQL*. Написан је у *Python* програмском језику и прати *MVC* архитектуру (енгл. *Model-View-Controller*). Омогућава развој софтвера на једноставан и брз начин. *Django* ставља акценат на поновно коришћење и „прикључност" компоненти, брз развој и принцип „без понављања". *Django* програмерима омогућава да избегну уобичајене и честе безбедносне грешке, као што су уметање *SQL* упита (енгл. *SQL Injection*), уметање скриптова (енгл. *cross-site scripting*) и преваре унакрсним захтевима (енгл. *cross-site request forgery*). Изузетно је скалабилан, што значи да је способан да поднесе велики пораст обима послова, без угрожавања функционалности и поузданости система, па га због тога користе познати сајтови као што су *Instagram*, *Mozilla*, *National Geographic*, *Pinterest*, *Bitbucket* и многи други [4].

Основна идеја при развоју апликације била је омогућити корисницима да покрећу све наведене алате на знатно једноставнији начин у поређењу са покретањем из командне линије. *Django* је одабран због своје једноставности и робусности. Као што је већ речено, *Django* се сматра *MVC* радним оквиром, али заправо не имплементира *MVC* архитектуру на стандардан начин. У даљем тексту следи кратак опис *MVC* архитектуре.

5.2 MVC архитектура

Једна од кључних ствари при развоју корисничких апликација јесте раздвајање садржаја од презентације. У овом контексту акценат је на подели одговорности између различитих слојева апликације. Апликација је подељена на три главне компоненте и свака од њих обавља различите задатке:

1. Модел (енгл. *model*) је интерна репрезентација података и пословне логике. Садржи главне програмске податке као што су информације о објектима из базе података. Сви подаци се добијају од модела, али се он не може директно позвати, већ посредством контролора. Контролор је тај који од модела захтева податке, модел затим обрађује захтеве од контролора и враћа податке контролору. Модел не зна ништа о корисничком интерфејсу.

2. Поглед (енгл. *view*) је компонента која обезбеђује кориснику интерфејс преко кога корисник уноси податке и позива одговарајуће операције које треба да се изврше над моделом. Поглед приказује кориснику стање модела.

3. Контролор (енгл. *controller*) садржи главну контролу програма и одговоран је за његов ток. То је први слој у веб апликацијама који се позива када корисник позове неки УРЛ. Контролор је такође задужен за управљање корисничким захтевима (нпр. када корисник кликне на неки ГУИ елемент) и иницира активности на нивоу модела и промене на погледу [22].

Слика 5.1: Приказ *MVC* архитектуре

У наставку ће бити приказана архитектура имплементираног система, са освртом на битне компоненте *Django* радног оквира.

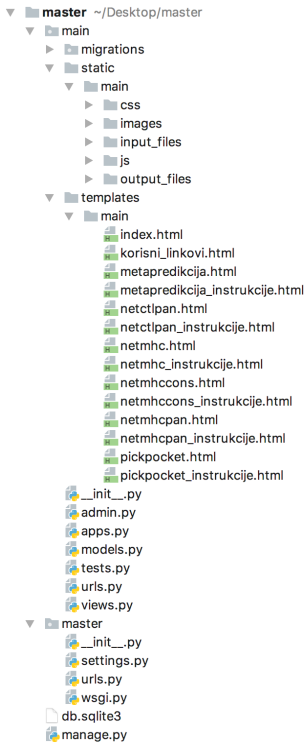
5.3 Структура пројекта

У *Django* интерпретацији *MVC* архитектуре поглед описује податке који се приказују кориснику и то не нужно са фокусом на то како подаци изгледају, већ са акцентом на томе *коју* се подаци приказују кориснику. Поглед, дакле, описује податке које корисник види, не како их види. У овом случају, поглед је *Python callback* функција за одређени УРЛ и та функција описује податке који се приказују. Пошто је кључно раздвојити садржај од презентације, неопходно је увести тзв. шаблоне (енгл. *template*), који су задужени за приказ података. Намеће се питање какву улогу контролор обавља. У овом случају, контролор би заправо био сам радни оквир: машинерија која шаље захтев одговарајућем погледу на основу постојеће *Django* УРЛ конфигурације о којој ће бити речи у наставку. Закључак је да *Django* архитектуру заправо можемо звати *MTV* архитектуром (енгл. *Model-Template-View*). У наставку ће бити речи о овим компонентама у току проласка кроз структуру пројекта.

Важно је на почетку објаснити разлику између пројекта и апликације у *Django* радном оквиру. Наиме, апликација може бити било шта што има неку одређену функцију, нпр. апликација за гласање, док је пројекат колекција различитих конфигурација и апликација за одређени веб сајт. Дакле, пројекат може да се састоји из више апликација и једна апликација може да се налази у више различитих пројеката. Овај пројекат се састоји из једне *main* апликације и сав код се налази унутар ње. *Django* долази са механизмом који може да генерише основну директоријум структуру апликације. На слици 5.2 приказана је структура пројекта.

Као што је већ речено, поглед је тип веб стране у *Django* апликацији који има одређену функцију и придружен шаблон. Све веб стране и сав садржај се испоручују кроз поглед. Сваком погледу одговара једна *Python* функција. У овом случају имамо следеће погледе који су дефинисани у *views.py*:

- **index** - поглед који приказује насловну страну апликације,
- **netctlpan** - поглед који приказује *netctlpan* страну,
- **netctlpan_rezultati** - поглед који је задужен за достављање резултата након позива *netctlpan* алата,
- **netctlpan_instrukcije** - поглед који је задужен за приказивање инструкција за покретање *netctlpan-a*,



Слика 5.2: Приказ структуре пројекта

- **netmhc** - поглед који приказује *netmhc* страну,
- **netmhc_rezultati** - поглед који је задужен за достављање резултата након позива *netmhc* алата,
- **netmhc_instrukcije** - поглед који је задужен за приказивање инструкција за покретање *netmhc-a*,
- **netmhcpan** - поглед који приказује *netmhcpan* страну,
- **netmhcpan_rezultati** - поглед који је задужен за достављање резултата након позива *netmhcpan* алата,
- **netmhcpan_instrukcije** - поглед који је задужен за приказивање инструкција за покретање *netmhcpan-a*,
- **pickpocket** - поглед који приказује *pickpocket* страну,
- **pickpocket_rezultati** - поглед који је задужен за достављање резултата након позива *pickpocket* алата,

- **pickpocket_instrukcije** - поглед који је задужен за приказивање инструкција за покретање *pickpocket-a*,
- **netmhccons** - поглед који приказује *netmhccons* страну,
- **netmhccons_rezultati** - поглед који је задужен за достављање резултата након позива *netmhccons* алата,
- **netmhccons_instrukcije** - поглед који је задужен за приказивање инструкција за покретање *netmhccons-a*,
- **metapredikcija** - поглед који приказује страну за метапредикцију,
- **metapredikcija_rezultati** - поглед који је задужен за достављање резултата након позива алата за метапредикцију,
- **metapredikcija_instrukcije** - поглед који је задужен за приказивање инструкција за покретање метапредикције,
- **korisni_linkovi** - поглед који приказује страну са корисним линковима.

Битна одлика *Django* радног оквира је елегантна УРЛ схема која је омогућена захваљујући тзв. УРЛ конфигурацији (енгл. *URL configuration*). УРЛ конфигурација представља обичан *Python* модул и омогућава једноставно упаривање између УРЛ-ова који су описани регуларним изразима (енгл. *regular expression*) и функција (погледа).

Ако још једном погледамо структуру пројекта, видећемо два *urls.py* фајла - један који се налази ван *main* апликације и један који се налази унутар *main* апликације. Први се практично односи на УРЛ садржај читавог пројекта и укључује, између осталог, и УРЛ конфигурацију *main* апликације (у овом случају је то једина апликација у пројекту). Тиме што се у глобалном *urls.py* фајлу налази референца на *urls.py* фајл апликације омогућено је једноставно манипулисање УРЛ-овима и лако надовезивање на */main/* УРЛ.

```
1 urlpatterns = [
2     url(r'^main/', include('main.urls'))
3 ]
```

Код 5.1: Глобални *urls.py* - онај који је везан за цео пројекат

Django ће одабрати одговарајући поглед на основу УРЛ-а који му је прослеђен, односно дела УРЛ-а након имена домена. УРЛ образац (енгл. *URL pattern*) је само генерална форма УРЛ-а (нпр. */main/metapredikcija/*) и у овом случају описан је регуларним изразом. У фрагментима кода 5.1 и 5.2 приказано је како изгледају оба *urls.py* фајла.

```
1  urlpatterns = [
2      url(r'^$', views.index, name='index'),
3      url(r'^netctlpan/$', views.netctlpan, name='netctlpan'),
4      url(r'^netctlpan_rezultati/$', views.netctlpan_rezultati, name='
       netctlpan_rezultati'),
5      url(r'^netmhc/$', views.netmhc, name='netmhc'),
6      url(r'^netmhc_rezultati/$', views.netmhc_rezultati, name='
       netmhc_rezultati'),
7      url(r'^netmhcpan/$', views.netmhcpan, name='netmhcpan')
8
9      ...
10
11 ]
```

Код 5.2: Део *urls.py* кода који је везан за саму апликацију

Ако би на пример корисник затражио страну /main/netctlpan/, *Django* би прво учитао *Python* модул *ime_sajta.url* јер му је то задато у глобалним подешавањима. Затим би нашао променљиву *urlpatterns* и онда би ишао редом кроз регуларне изразе којима су УРЛ-ови дефинисани и покушавао да нађе поклапање са задатом адресом. Прва линија у фрагменту кода 5.2 одговарала би делу *main/*. При даљем обиласку овог низа би наишао на поклапање са целом траженом адресом - линија 3. Када наиђе на поклапање са траженом адресом, он у том случају зна да треба да позове *netctlpan()* метод и да га изврши на начин на који је приказан у фрагменту кода 5.3.

```
1 def netctlpan(request):
2     return render(request, 'main/netctlpan.html')
```

Код 5.3: *NetCTLpan* поглед

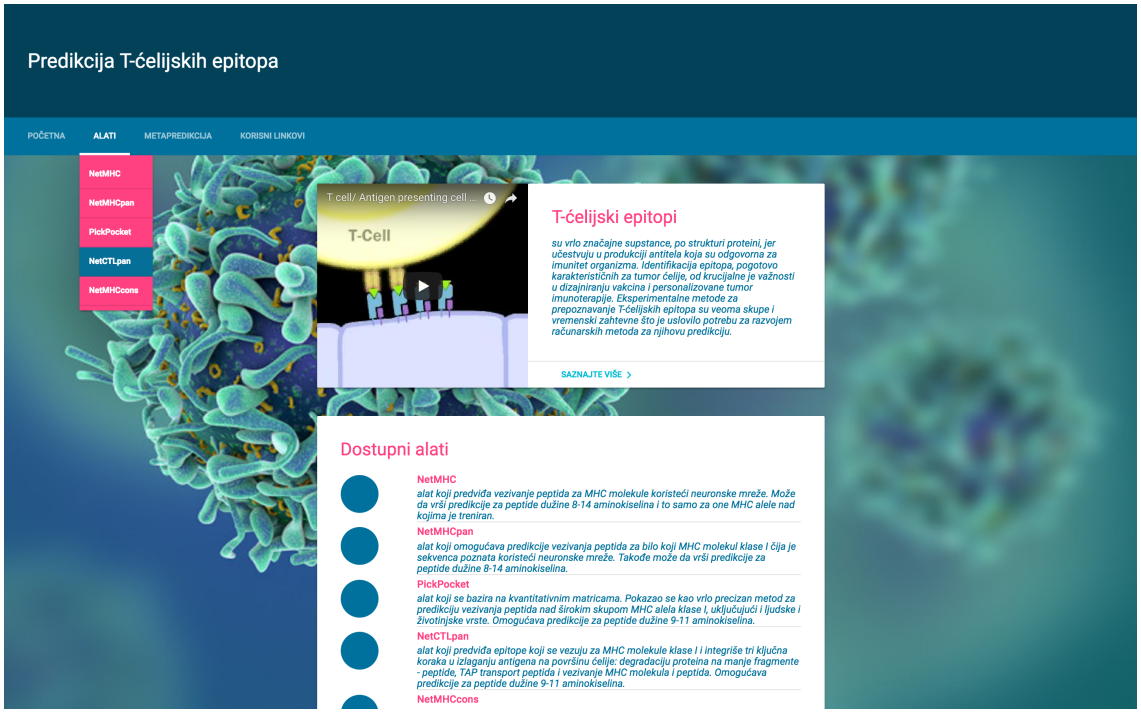
Ако бисмо се одлучили да у нашим погледима пишемо код који се односи на дизајн стране, нарушили бисмо основни концепт раздвајања садржаја од презентације. Из тог разлога направљен је *templates* директоријум у оквиру

main апликације у ком су смештене *HTML* стране. Оне представљају шаблоне о којима је претходно било речи и њих сада погледи могу да користе. У фрагменту кода 5.3 приказан је једноставан поглед који само дохвата и рендерује *netctlpan.html* страну. Ово је основни принцип на основу ког апликација ради.

Поред *HTML* страна, веб апликацији су обично потребни *CSS* фајлови, слике, *JavaScript* библиотеке. Они се сматрају статичким фајловима и налазе се у оквиру *static* директоријума, а који се налази у оквиру *main* апликације. *Django* на паметан начин зна да увек ту треба да тражи статичке фајлове, исто као што зна да увек у оквиру директоријума *templates* треба да тражи шаблоне.

5.4 Рад апликације

Дакле, основни циљ апликације је развој заједничког интерфејса за покретање различитих предикционих алата и имплементација новог тзв. метапредиктора који ће да комбинује постојеће алате и омогући бољи увид у резултате.



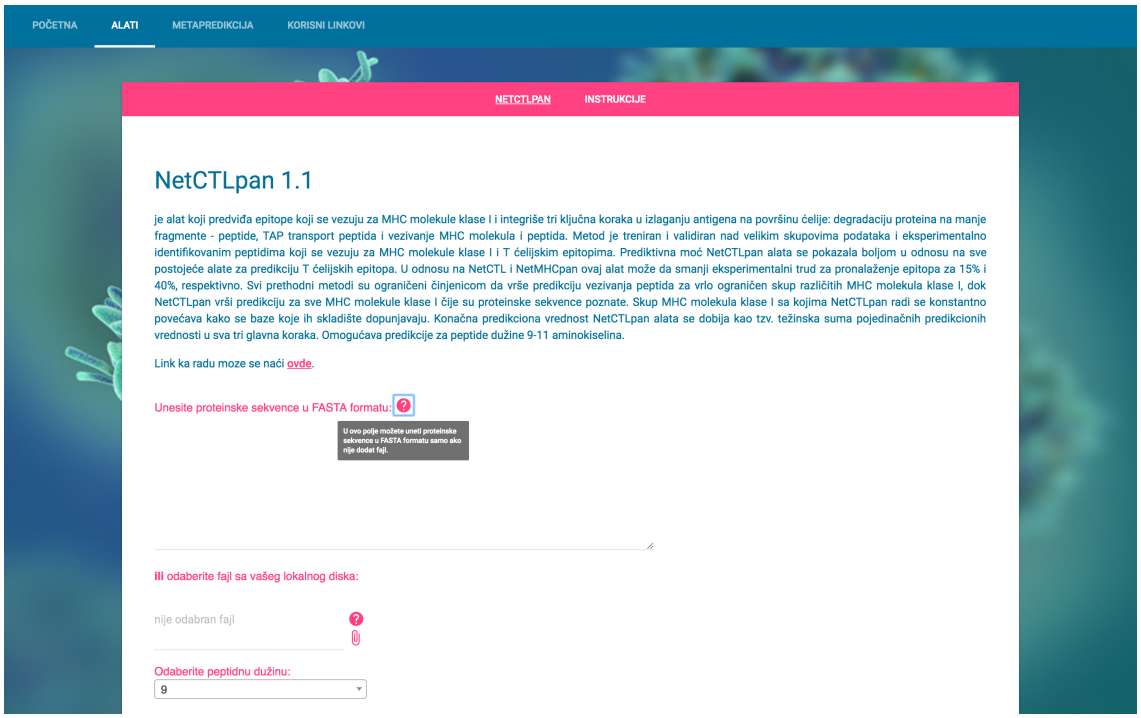
Слика 5.3: Део насловне стране апликације

Као што се види на слици 5.3, са насловне стране је могуће прећи на сваку од страна појединачних алата. У наставку ће бити објашњен сам рад апликације

на примеру једног од алата - *NetCTLpan-a*, са освртом на неке разлике код осталих алата. Посебно ће бити размотрен део који је везан за метапредикцију.

NetCTLpan

Када се отвори *NetCTLpan* страна, кориснику се приказује формулар са пољима која служе да корисник унесе различите параметре који су карактеристични за сам алат. На основу тих параметара се гради командна линија сваког алата. На слици 5.4 може се видети део *NetCTLpan* формулара.



Слика 5.4: Део *NetCTLpan* формулара

Кориснику се нуди да унесе протеинске секвенце у *FASTA* формату у текстуално поље *или* да дода фајл у истом формату који ће служити као улазни фајл у алат. За разлику од свих осталих алата код којих је то имплементирано, *NetCTLpan* је једини алат који не прима фајл са подацима у пептидном формату. Затим, кориснику се нуди могућност да одабере дужину пептида за коју ће се вршити предикција. И овде је важно нагласити да *NetCTLpan*, као и *PickPocket*, у току једног извршавања могу да врше предикције за пептиде фиксне дужине, док је осталим алатима могуће проследити више пептидних дужина у исто време. Кориснику се такође нуди опција да одабере *HLA* тип

за који ће се вршити предикција. Допуштен је и унос прага за приказивање резултата алата (који се односи на комбиновани предикциони скор), као и прага за слабе и јаке епитопе. Такође, могуће је бирати начин сортирања резултата, као и филтрирање резултата, у смислу да се као финални излаз приказују само препознати епитопи, не и сви остали процесирани пептиди. Поред приказа резултата алата, кориснику је омогућено и преузимање оригиналног излазног фајла из сваког алата. На слици 5.5 могу се видети излазни резултати из *NetCTLpan-a*.

Ovde možete preuzeti izlazni fajl

Ime proteina: 143B_BOVIN_(P29 . HLA alel: HLA-A*02:01. Broj peptida: 237. Broj MHC liganda: 0 identifikovano.

peptid	hla tip	ime sekvence	mhc vrednost	tap vrednost	sečenje proteina	kombinovana vrednost	rang	tip
AFDEAIAEL	HLA-A*02:01	143B_BOVIN_(P29	0.402	1.1	0.97574	0.64904	3	-
QLLRDNLTL	HLA-A*02:01	143B_BOVIN_(P29	0.425	1.041	0.97391	0.67015	3	-
LLSVAYKNV	HLA-A*02:01	143B_BOVIN_(P29	0.472	0.475	0.48171	0.59226	4	-
ALNFSVFYY	HLA-A*02:01	143B_BOVIN_(P29	0.232	3.042	0.97083	0.52649	5	-
SSWRVISSI	HLA-A*02:01	143B_BOVIN_(P29	0.249	0.932	0.96898	0.49032	6	-
MQPTHPIRL	HLA-A*02:01	143B_BOVIN_(P29	0.234	0.895	0.95489	0.47122	6	-
AVTEQGHEL	HLA-A*02:01	143B_BOVIN_(P29	0.203	1.302	0.97058	0.45393	6	-

Слика 5.5: Део *NetCTLpan излаза*

Важно је напоменути да, уколико корисник унесе лоше параметре или лошу пептидну секвенцу на улазу, добиће одговарајуће поруке и биће упућен на нови покушај покретања алата са исправним параметрима.

На страни *Инструкције* дата су детаљна објашњења за покретање сваког појединачног алата, а затим и појашњења излазних резултата. На слици 5.6 могу се видети инструкције за покретање *NetCTLpan-a*. Што се осталих алата тиче, формулари су врло слични овом. Разликују се у појединим параметрима, али нема неке значајније разлике у самој имплементацији.

Да бисмо објаснили механизам који омогућава описану функционалност, неопходно је описати начин на који се подаци из формулара шаљу серверу, односно специфичном погледу који је задужен за покретање алата на серверској страни и слање одговора назад клијенту. У ове сврхе коришћен је *AJAX* механизам. *AJAX* или асинхрони *JavaScript* и *XML* (енгл. *Asynchronous JavaScript And XML*) омогућава да веб стране буду ажуриране асинхроно док размењују

Слика 5.6: Део инструкција за покретање *NetCTLpan-a*

податке са веб сервером у позадини, односно да буду ажурирани делови стране, а да се притом не учитава цела страна.

Притиском на дугме *У реду* шаље се *AJAX* позив одговарајућем погледу (у овом случају је то поглед *netctlpan_rezultati*). Поглед на основу параметра који су му послати кроз *AJAX* позив формира командну линију и екстерни програм *NetCTLpan* се позива из погледа са претходно формираном командном линијом. Резултат извршавања сваког програма се уписује у фајл, а затим се тај фајл парсира специфично за сваки алат. Тако парсиран фајл се затим учитава у *pandas.DataFrame* - табеларну структуру података која даље бива обрађена и пребачена у *JSON* (енгл. *JavaScript Object Notation*) формат. *JSON* је текстуално базиран отворени стандард који је дизајниран за људима разумљиву размену података и пре свега се користи за размену података. Тако формиран *JSON* се шаље назад клијенту и на основу тих података се ажурира део стране који је задужен за приказивање резултата програма.

Још један важан аспект у развоју сваке апликације јесте њен кориснички интерфејс. У изради корисничког интерфејса коришћена је MDL (енгл. *Material Design Lite*) библиотека [15]. *MDL* библиотека састављена је од нових и побољшаних верзија уобичајених компоненти интерфејса, као што су дугмад (енгл.

buttons), текстуална поља (енгл. *text fields*), поља за потврду (енгл. *check boxes*) итд. Бесплатна је за преузимање и коришћење и подржана је независно од веб прегледача и радног оквира.

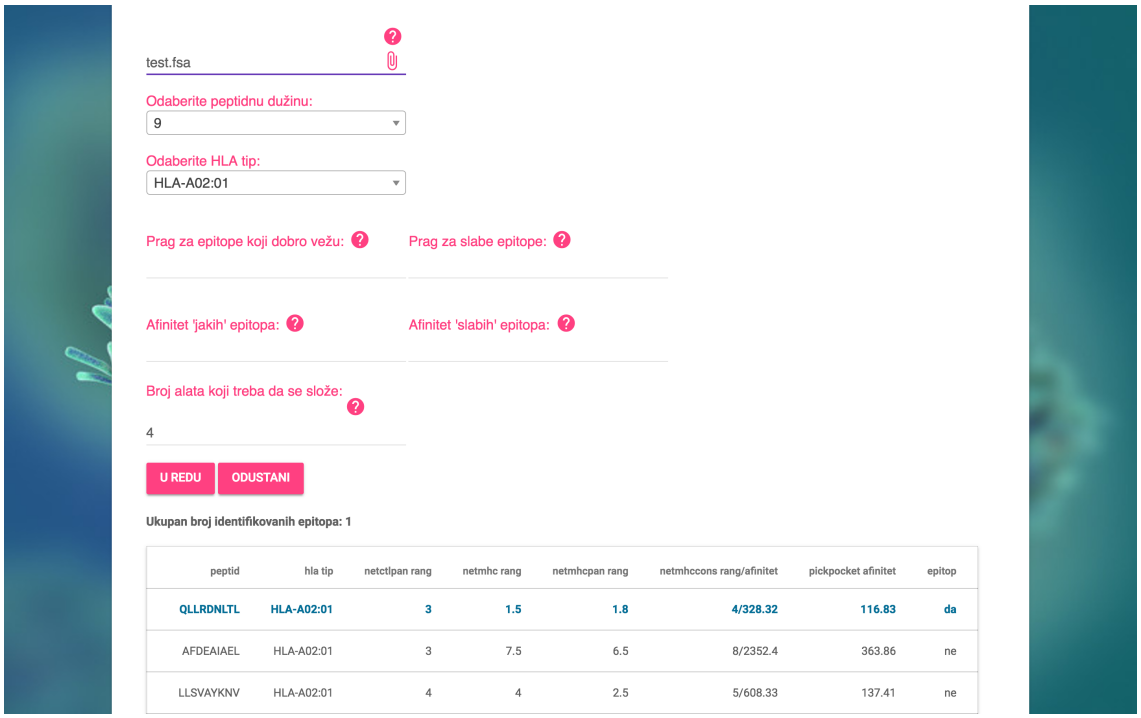
MDL компоненте састављене су од *CSS-a*, *JavaScript-a* и *HTML-a*. При изради веб страна, компоненте се комбинују и заједно доприносе њиховој атрактивности, конзистентности и функционалности. Стране које су развијене коришћењем *MDL-a* придржавају се основних приципа у модерном веб развоју као што су портабилност на различитим прегледачима и независност од уређаја на ком се апликације чије су оне део извршавају.

Метапредикција

Након што је приказано покретање појединачних алата, у овом поглављу биће дат опис новог предиктора, тзв. „метапредиктора". У метапредикцију су укључени свих 5 претходно описаних алата, који се комбинују једноставним консензусом. Примера ради, ако од 5 предиктора 3 за неку секвенцу аминокиселина у протеину предвиде да представља потенцијални епитоп, резултат метапредиктора је да је та секвенца епитоп, у супротном да није. Подразумевано, неки пептид је епитоп ако бар један од ових 5 алата каже да јесте, мада је овај параметар могуће мењати. На слици 5.7 може се видети део стране за метапредикцију. У наставку ће бити речи о начину на који је комбиновање ових алата омогућено.

На самом почетку је било потребно формирати скуп *HLA* типова који су подржани од стране свих 5 алата. Дакле, метапредикција може да ради над ограниченим скупом *HLA* типова, односно само над оним који су заједнички за све алате. Као улаз може да прими само секвенце у *FASTA* формату (јер *NetCTLpan* не подржава пептидни улаз). Ово ограничење није велико јер се лаким претпроцесирањем података од пептидног улаза може направити *FASTA* формат. За једно покретање, метапредиктору се може дати само једна пептидна дужина, због ограничења *NetCTLpan-a* и *PickPocket-a* о ком је било речи у поглављу 4.5. У наставку ће посебна пажња бити посвећена мерама на основу којих се резултати алата комбинују.

Наиме, *NetMHC*, *NetMHCpan* и *NetCTLpan* као излаз за сваки пептид дају и ранг о ком је било речи у поглављу 4.1. С обзиром да аутори ових алата препоручују да се ранг узима као мера на основу које се за дати пептид одлучује да ли је епитоп или не, природно је ранг одабран као заједничка мера. Међутим,



Слика 5.7: Део стране за метапредикцију

NetMHCcons поред ранга узима у обзир и афинитет, док *PickPocket* уопште не даје ранг пептида у свом излазу, већ за сваки пептид даје $1 - \log_{50000}(aff)$, односно логаритамски трансформисану вредност афинитета која је била споменута у поглављу 4.1. Пошто се пептиди чији је афинитет мањи од 500 сматрају слабим епитопима, а пептиди чији је афинитет мањи од 50 сматрају јаким епитопима, ова мера је у недостатку ранга код *PickPocket* узета у обзир. Од логаритамски трансформисане вредности афинитета лако је израчунат афинитет за сваки пептид у случају *PickPocket-a*. *NetMHCcons* рангира пептиде и на основу ранга и на основу афинитета. Из ових разлога, обе поменуте мере се нуде као параметри кориснику при метапредикцији, а ако их не унесе, подразумеване вредности су, као и до сада, 2 и 0,5 за ранг, као и 500 и 50 за афинитет за слабе и јаке епитопе, респективно. И као последњи, не мање важан параметар, јесте број алата чији резултати за дати пептид треба да се сложе како би пептид био проглашен епитопом. Као излазне вредности, за сваки пептид дат је резултат сваког предикционог алата, као и колона која каже да ли је дати пептид метапредиктором препознат као епитоп.

Развој веб апликације је подразумевао да, након што су имплементиране функције за позиве разматраних алата, све оне буду интегрисане у једном по-

гледу, а њихови резултати обједињени узимајући у обзир претходно описане мере.

За потребе тестирања апликације коришћени су подаци из *IEDB* базе (енгл. *Immune Epitope Database and Analysis Resource*). *IEDB* нуди претрагу експериментално потврђених Т ћелијских епитопа. Могуће је преузети ове податке са адресе [11] и прилагодити их форматима које ови алати користе.

Глава 6

Закључак

Овај рад осмишљен је са намером да се неки од најпопуларнијих алата за предикцију Т-ћелијских епитопа у научној заједници обједине у једној апликацији. Омогућено је њихово самостално покретање кроз веб интерфејс, које је тако имплементирано да корисницима пружа сву удобност у раду са овим алатима. Кориснику је дата слобода да сам подешава параметре који одговарају његовом истраживању и дат му је детаљан опис за коришћење и покретање алата, као и објашњење излазних резултата. Имплементиран је и нови метапредиктор који користити предикције свих разматраних алата и доноси коначну одлуку на основу различитих гласачких шема (консензуса). Метапредиктор је заправо покушај да се резултати разматраних алата интегришу, као и да се обезбеди сигурнија и поузданија предикција.

Систематична процена перформанси алата које веб апликација обухвата, укључујући и постојеће алате и метапредиктор, била би веома корисна као круна овог истраживања. Међутим, приликом разматрања овог корака наиђено је на више проблема. То су пре свега неадекватне документације скупа података и самих метода које алати имплементирају, а затим и недоступност самих скупова података над којима су методе трениране и евалуиране.

Даљи развој би подразумевао детаљно и исцрпно анализирање и поређење перформанси метапредиктора у односу на перформансе појединачних алата користећи унапред одабрани скуп експериментално потврђених Т-ћелијских епитопа.

Литература

[1] Morten Nielsen Massimo Andreatta. *NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets*. https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0288-x посећено 21.9.2017.

[2] Aidan MacNamara Ulrich Kadolsky Charles R. M. Bangham Becca Asquith. *T-Cell Epitope Prediction: Rescaling Can Mask Biological Variation between MHC Molecules*. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000327 посећено 21.9.2017.

[3] *CBS*. http://www.cbs.dtu.dk/services/ посећено 21.9.2017.

[4] *Django*. https://www.djangoproject.com/ посећено 21.9.2017.

[5] Atanas Patronov Irini Doytchinova. *T-cell epitope vaccine design by immunoinformatics*. http://rsob.royalsocietypublishing.org/content/royopenbio/3/1/120139.full.pdf посећено 21.9.2017.

[6] Wikipedia the free encyclopedia. *Antibody*. https://en.wikipedia.org/wiki/Antibody посећено 21.9.2017.

[7] Wikipedia the free encyclopedia. *Antigen*. https://en.wikipedia.org/wiki/Antigen посећено 21.9.2017.

[8] Wikipedia the free encyclopedia. *Dissociation constant*. https://en.wikipedia.org/wiki/Dissociation_constant посећено 21.9.2017.

[9] Wikipedia the free encyclopedia. *Lymphocyte*. https://en.wikipedia.org/wiki/Lymphocyte посећено 21.9.2017.

[10] Wikipedia the free encyclopedia. *Major histocompatibility complex*. https://en.wikipedia.org/wiki/Major_histocompatibility_complex посећено 21.9.2017.

[11] *IEDB*. http://www.iedb.org/result_v3.php?cookie_id=bd2eba посећено 21.9.2017.

[12] Linus Backert Oliver Kohlbacher. *Immunoinformatics and epitope prediction in the age of genomic medicine*. https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-015-0245-0 посећено 21.9.2017.

[13] Karosiene E Lundegaard C Lund O Nielsen M. *NetMHCcons: a consensus method for the major histocompatibility complex class I predictions.* https://www.ncbi.nlm.nih.gov/pubmed/22009319 посећено 21.9.2017.

[14] Zhang H Lund O Nielsen M. *The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding.* https://www.ncbi.nlm.nih.gov/pubmed/19297351 посећено 21.9.2017.

[15] *Material Design Lite*. https://getmdl.io/index.html посећено 21.9.2017.

[16] Claus Lundegaard Kasper Lamberth Mikkel Harndahl Søren Buus Ole Lund Morten Nielsen. *NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2447772/ посећено 21.9.2017.

[17] Mette V Larsen Claus Lundegaard Kasper Lamberth Soren Buus Ole Lund Morten Nielsen. *Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction.* https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-424 посећено 21.9.2017.

[18] Thomas Stranzl Mette Voldby Larsen Claus Lundegaard Morten Nielsen. *NetCTLpan: pan-specific MHC class I pathway epitope predictions*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875469/pdf/251_2010_Article_441.pdf посећено 21.9.2017.

[19] Ruth E.Soria-Guerra Ricardo Nieto-Gomez Dania O.Govea-Alonso Sergio Rosales-Mendoza. *An overview of bioinformatics tools for epitope prediction: Implications on vaccine development*. http://www.sciencedirect.com/science/article/pii/S1532046414002330 посећено 21.9.2017.

[20] Rajat K. De Namrata Tomar. *Immunoinformatics*. http://www.springer.com/gp/book/9781493911141 посећено 21.9.2017.

[21] *Линк ка GitHub репозиторијуму апликације*. https://github.com/milicakojicic/Master-rad.git.

[22] Саша Малков. *Архитектура веб апликација.* `http://poincare.matf.bg.ac.rs/~smalkov/files/pveb.r338.2016/public/predavanja/PVeb.2014.04%20-%20arh.apl,veb.2.0.pdf` посећено 21.9.2017.