УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Даворка Р. Јандрлић

Примена правила придруживања и метода подржавајућих вектора за предвиђање Т - ћелијских епитопа

Докторска дисертација

Београд, 2016

UNIVERSITY OF BELGRADE
FACULTY OF MATHEMATICS

Даворка Р. Јандрлић

Application of association rule and support vector machine technique for T - cell epitope prediction

Doctoral dissertation

Belgrade, 2016

Подаци о ментору и члановима комисије

Ментор

проф. др Ненад Митић, ванредни професор, Математички факултет, Универзитет у Београду

Чланови комисије

проф. др Ненад Митић, ванредни професор, Математички факултет, Универзитет у Београду

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

др Мирјана Павловић, виши научни сарадник, Институт за општу и физичку хемију, Универзитет у Београду

Датум одбране:

Иви и Николи

Подаци о докторској дисертацији

Наслов дисертације: Примена правила придруживања и метода подржавајућих вектора за предвиђање Т - ћелијских епитопа

Резиме: Истраживање података (eng. *Data mining*) је интердисциплинарно поље информатике које се бави аутоматским или полу-аутоматским откривањем знања у подацима. Основни задатак истраживања података је издвајање нетривијалних, претходно непознатих и потенцијално корисних образаца, односа и веза у подацима и статистички значајних структура из великих колекција података. Императив је да добијени резултати буду нови, ваљани, корисни и разумљиви. Технике за истраживање података укључују статистичке моделе, метематичке алгоритме и методе машинског учења. Брз развој у областима имунологије, геномике, протеомике, молекуларне биологије и другим сродним областима је условио велики пораст биолошких података. Без аутоматских метода за издвајање података готово је немогуће истражити и анализирати те податке.

Један од најактуелнијих проблема у имуноинформатици је проблем идентификовања Т – ћелијских епитопа. Идентификовање Т - ћелијских епитопа, а посебно доминантних Т - ћелијских епитопа широко заступљених у популацији, је кључан корак у добијању вакцина заснованих на епитопима као и откривању узорака аутоимуних болести. Вакцине засноване на епитопима су веома значајне у борби против инфективних болести, хроничних болести и различитих врста канцера. Експерименталне методе за идентификовање Т - ћелијских епитопа су веома скупе, временски јако захтевне, и такође нису применљиве за истраживања великих размера (нарочито када се ради о избору оптималне групе епитопа код вакцина за одређену популацију или индивидуу).

Информатичке и математичке методе су у овом подручју кључне како би омогућиле систематично истраживање и идентификовање Т - ћелијских епитопа на великом скупу података [16]. Везивање епитопа за молекуле главног хистокомпатибилног комплекса (eng. *major histocompatibility complex*, *MHC*) је основа адаптивног ћелијског имунитета јер Т ћелије препознају антиген само у форми пептида везаних за *MHC* молекуле [1]. Како су Т - ћелијски епитопи кључни, како за ћелијски, тако и за хуморални имунитет, постоји велики број

метода које предвиђају ове епитопе.

У оквиру ове дисертације, разматране су искључиво методе засноване на пептидној секвенци и везивању за *MHC* молекуле. Описане су постојеће методе за предвиђање Т - ћелијских епитопа, недостаци и ограничења постојећих метода и неке од најзначајнијих доступних база података са експериментално утврђеним Т - ћелијским епитопима. Помоћу метода истраживања података су развијени нови модели за предвиђање Т – ћелијских епитопа и урађена је систематична анализа заступљености Т – ћелијских епитопа у различитим структурама протеина на великом скупу података. Технике истраживања података које су коришћене у оквиру тезе су технике засноване на подржавајућим векторима, правилима придруживања и кластеровању k - срединама.

Развијени модели који су главни допринос ове дисертације су по карактеристикама упоредиви са тренутно најбољим постојећим методама, а у неким случајевима и бољи. Основа модела су технике подржавајућих вектора за проблем класификације и регресије. Направљени су и помоћни модели за бинарну класификацију засновани на техници груписања k - срединама, који су послужили за издвајање атрибута и припрему улаза за нове класификационе и регресионе моделе.

Други део тезе се односи на успостављање правила о заступљености Т - ћелијских епитопа у различим структурама протеинских секвенци. Техником правила придруживања су издвојени обрасци који могу да послуже за проналажење Т - ћелијских епитопа у протеинској секвенци. На основу ових образаца урађена је детаљна анализа над великим скупом података, добијеним применом постојећих алата за: предвиђање Т - ћелијских епитопа, неуређених/уређених структура протеина и хидропатије протеинских региона. Резултати спроведене анализе су потврдили да учесталост амино киселина у пептиду и одређене физичко хемијске особине имају директан утицај на класификовање епитопа. Током израде тезе развијен је и софтверски систем који има широку примену у предвиђању различитих карактеристика протеина.

Део резултата ове тезе је описан у радовима [71][82][45][42][43][44][72][73] који су објављени или су тренутно у процесу рецензије.

Дисертација је организована на следећи начин:

У глави 1 је приказан увод у проблем препознавања Т - ћелијских епитопа, значај математичких и информатичких методе за предвиђање Т - ћелијских епитопа, као и значај Т - ћелијских епитопа за имуни систем и основе функционисања имуног система.

У глави 2 су детаљно описане технике истраживања података које су коришћене у оквиру тезе за прављење нових модела.

У глави 3 је дат преглед постојећих метода за предвиђање Т - ћелијских епитопа и описан је начин рада постојећих модела и метода. Истакнути су недостаци постојећих метода који су били мотивација за прављење нових модела за предвиђање Т - ћелијских епитопа. Описане су неке од јавно доступних база са експериментално утврђеним *MHC* везујућим пептидима и Т - ћелијским епитопима.

У глави 4 су представљени новоразвијени модели за предвиђање епитопа. Развијени модели укључују нове шеме за представљање пептида у виду вектора који је погодан као улаз у моделе засноване на подржавајућим векторима.

У глави 5 су приказани резултати добијени применом развијених модела, њихово међусобно поређење и поређење са постојећим алатима за предвиђање Т - ћелијских епитопа.

У глави 6 су представљени резултати добијени истраживањем односа Т - ћелијских епитопа и уређених/неуређених региона у протеину помоћу правила придруживања. У оквиру ове главе су представљени сумарни резултати који су детаљније приказани у радовима [71][82][45] [44].

Глава 7 садржи закључке и могуће правце даљег рада.

Кључне речи: Подржавајући вектори, класификација, регресија, груписање к-срединама, правила придруживања, Т - ћелијски епитопи
Научна област: Рачунарство
Ужа научна област: Истраживање података, Биоинформатика
УДК број: [004.6:004.852.2]:576.8(043.3)

Dissertation Data

Dissertation title: Application of association rule and support vector machine technique for T cell epitope prediction

Abstract: Data mining is an interdisciplinary subfield of computer science, including various scientific disciplines such as: database systems, statistics, machine learning, artificial intelligence and the others. The main task of data mining is automatic and semi-automatic analysis of large quantities of data to extract previously unknown, nontrivial and interesting patterns. Rapid development in the fields of immunology, genomics, proteomics, molecular biology and other related areas has caused a large increase in biological data. Drawing conclusions from these data requires sophisticated computational analyses. Without automatic methods to extract data it is almost impossible to investigate and analyze this data.

Currently, one of the most active problems in immunoinformatics is T – cell epitope identification. Identification of T - cell epitopes, especially dominant T - cell epitopes widely represented in population, is of the immense relevance in vaccine development and detecting immunological patterns characteristic for autoimmune diseases. Epitope-based vaccines are of great importance in combating infectious and chronic diseases and various types of cancer. Experimental methods for identification of T - cell epitopes are expensive, time consuming, and are not applicable for large scale research (especially not for the choice of the optimal group of epitopes for vaccine development which will cover the whole population or personalized vaccines).

Computational and mathematical models for T - cell epitope prediction, based on MHC-peptide binding, are crucial to enable the systematic investigation and identification of T - cell epitopes on a large dataset and to complement expensive and time consuming experimentation [16]. T - cells (T - lymphocytes) recognize protein antigen(s) only when degradated to peptide fragments and complexed with Major Histocompatibility Complex (MHC) molecules on the surface of antigen-presenting cells [1]. The binding of these peptides (potential epitopes) to MHC molecules and presentation to T - cells is a crucial (and the most selective) step in both cellular and humoral adoptive immunity. Currently exist numerous of methodologies that provide identification of these epitopes.

In this PhD thesis, discussed methods are exclusively based on peptide sequence

binding to MHC molecules. It describes existing methodologies for T - cell epitope prediction, the shortcomings of existing methods and some of the available databases of experimentally determined linear T - cell epitopes. The new models for T - cell epitope prediction using data mining techniques are developed and extensive analyses concerning to whether disorder and hydropathy prediction methods could help understanding epitope processing and presentation is done. Accurate computational prediction of T cell epitope, which is the aim of this thesis, can greatly expedite epitope screening by reducing costs and experimental effort. These theses deals with predictive data mining tasks: classification and regression, and descriptive data mining tasks: clustering, association rules and sequence analysis.

The new-developed models, which are main contribution of the dissertation are comparable in performance with the best currently existing methods, and even better in some cases. Developed models are based on the support vector machine technique for classification and regression problems. A new approach of extracting the most important physicochemical properties that influence the classification of MHC-binding ligands is also presented. For that purpose are developed new clustering-based classification models. The models are based on k-means clustering technique.

The second part of the thesis concerns the establishment of rules and associations of T - cell epitopes that belong to different protein structures. The task of this part of research was to find out whether disorder and hydropathy prediction methods could help in understanding epitope processing and presentation. The results of the application of an association rule technique and thorough analysis over large protein dataset where T cell epitopes, protein structure and hydropathy has been determined computationally, using publicly available tools, are presented. During the research on this theses new extendable open source software system that support bioinformatic research and have wide applications in prediction of various proteins characteristics is developed.

A part of this thesis is described in the works [71][82][45][42][43][44][72][73] that are published or submitted for publications in several journals. The dissertation is organized as follows:

In section1 is illustrated introduction to the problem of identifying T - cell epitopes, the importance of mathematical and computational methods in this area,

as well as the importance of T - cell epitopes to the immune system and basis for functioning of the immune system.

In section 2 are described in details data mining techniques that are used in the thesis for development of new models.

Section 3 provides an overview of existing methods for predicting the T - cell epitopes and explains the work methodologies of existing models and methods. It pointed out the shortcomings of existing methods which have been the motivation for the development of new models for the T - cell epitope prediction. Some of the publicly available databases with the experimentally determined MHC binding peptides and T - cell epitope are described.

In section 4 are presented new developed models for epitopes prediction. The developed models include three new encoding schemes for peptide sequences representation in the form of a vector which is more suitable as input to models based on the data mining techniques.

Section 5 reports results of presented new classification and regression models. The new models are compared with each other as well as with currently existing methods for T cell epitope prediction.

Section 6 presents the research results of the T - cell epitopes relationship with ordered and disordered regions in proteins. In the context of this chapter summary results are presented which are shown in more detail in the published works [71][82][45][44].

Section 7 concludes the dissertation with some discussion of the potential significance of obtained results and some directions for future work.

Keywords: Support vector machine, classification, regression, κ-mean clustering, association rules, T cell epitopes

Scientific field: Computer Science

Scientific discipline: Data mining, Bioinformatics

UDC number: [004.6:004.852.2]:576.8(043.3)

Предговор

Биоинформатика је интердисциплинарно поље науке, где се комбинују математика, информатика и биологија како би се интерпретирале и симулирале биолошке појаве. Циљ биоинформатике је боље разумевање генетске основе разних болести, специфичних прилагођавања организама или разлика у популацији. Примена техника истраживања података је значајна јер омогућава издвајање корисних резултата из велике количине необрађених података. Примери примене укључују препознавање образаца, анализу података, машинско учење и визуелизацију биолошких података. Истраживања у овој области укључују "равнање" протеинских секвенци, предвиђање гена, откривање и прављење лека, структурно "равнање" протеина, предвиђање структуре протеина, идентификовање протеин - протеин интеракција, изучавање геномских веза, моделовење еволуције и још многе друге процесе. Током неколико последњих деценија брз развој геномике, молекуларне биологије и других сродних области је утицао на пораст значајних количина података, које се не могу једноставно анализирати и интерпретирати без примене информатичких и математичких техника.

У оквиру ове тезе је представљено једно решење проблема идентификовања T - ћелијских епитопа помоћу метода истраживања података. T - ћелијски епитопи се користе у имунотерапијама и у прављењу вакцина заснованих на пептидима. Њихово идентификовање има кључну улогу у разумевању функционисања имуног система. Експерименталне методе за препознавање T - ћелијских епитопа су веома скупе и временски јако захтевне што је условило потребу за развојем рачунарских метода за симулирање неких процеса имуног система. У претходних двадесет година је развијен велики број рачунарских метода за предвиђање T - ћелијских епитопа, које се веома успешно користе као допуна експерименталним методама, редукују број потребних

експеримената и знатно скраћују време потребно за идентификовање Т - ћелијских епитопа. Тачност ових метода је кључна у ситуацијама када је неопходна брза имунизација, и када је време од сустишке важности, где није могуће извести експерименталне методе. Побољшање тачности ових метода је један од најважнијих циљева у имуноинформатици, што је и била мотивација за истраживање у овој области.

У оквиру истраживања, изузетно ми је помогао ментор др Ненад Митић, ванредни професор Математичког факултета у Београду, и др Мирјана Павловић, виши научни сарадник Института за општу и физичку хемију, којима се овом приликом посебно захваљујем. Поред многих драгоцених коментара, усмеравања и савета, захвална сам и на подршци и разумевању које ми је указано током израде тезе. Захвална сам и на свему што сам од њих научила, током заједничког рада у задњих неколико година, и на активном учествовању током истраживања у оквиру тезе. Веома сам захвална и др Милошу Бељанском, научном саветнику Института за општу и физичку хемију на свим корисним саветима и охрабривању током истраживања. Члану комисије др Гордани Павловић-Лажетић, редовном професору Математичког факултета у Београду, дугујем захвалност на свему што сам научила током основних и магистарских студија на њеним предметима, на позитивном и професионалном ставу. Изузетно задовољство ми је причињавала свака сарадња са професорком Лажетић.

Највећу захвалност дугујем својој породици, родитељима, брату и сестри, за сву подршку, љубав и разумевање. Мом супругу, Андрији, на првом месту, дугујем највећу захвалност за сву подршку, помоћ, бескрајну толеранцију, љубав и пажњу, на охрабривању у моментима када ми је то највише било потребно. Својој деци, ћерки Иви и сину Николи, који су били моја највећа мотивација и извор енергије током целог истраживања.

Београд, април 2016.

Даворка Јандрлић

Садржај

Списак слика

Списак табела

Поглавље 1

Увод

1.1 Значај информатичких и математичких метода за предвиђање Т - ћелијских епитопа

Основа за препознавање Т - ћелијских епитопа је способност везивања за молекуле *MHC* класа. Иако само везивање није потврда да ће пептид бити препознат као епитоп, оно је најселективнији и основни корак у препознавању Т - ћелијских епитопа. Поступак идентификовања Т - ћелијских епитопа се у пракси најчешће изводи у два корака [24]:

(а) прво се примењују математичке и информатичке методе за предвиђање *MHC* везујућих пептида ("in silico" предвиђања),

(б) над резултатима добијеним у кораку (а) се спроводе исцрпни и скупи експерименти, који треба да провере да ли се идентификовани пептиди заиста понашају као епитопи.

У претходном поступку се рачунарске методе користе као помоћ експерименталним методама. Јасно је да што су перформансе математичких и информатичких метода, које се користе у првом кораку, боље то је лабораторијски рад бржи, троши се мање времена и новца за експерименте који се спроводе у другом кораку. Понекад, у случају када је време кључан фактор (нпр. потребна је брза имунизација) и није могуће спровести експерименте, рачунарске методе комплетно замењују експерименталне. Такође, у случају анализе великих количина података неопходна је примена рачунарских метода без обзира на степен коришћења експеримената.

1.2 *MHC* полиморфизам и ограничења експерименталних метода

Молекули главног хистокомпатибилног комплекса (*MHC*) имају веома важну улогу у адаптивном имунитету, тако што везују пептиде настале разградњом сопствених и не-сопствених протеина. *MHC* молекули представљају ове пептиде на површини антиген презентујућих ћелија (eng. *antigen presenting cells*, *APC*) где их излажу Т лимфоцитима, који иницирају имуни одговор. Схватање начина препознавања пептида који се везују за молекуле *MHC* класа је веома важно за разумевање функционисања адаптивног имуног система, јер они представљају резервоар Т - ћелијских епитопа и могу се користити у имунотерапијама, трансплантацијама и прављењу вакцина заснованих на епитопима. *MHC* молекули у људској популацији (eng. *human leukocyte antigens*, *HLA*) су екстремно полиморфни. То значи да постоје хиљаде различитих *MHC* молекула. За тако велики број различитих молекула експерименталне методе за препознавање везујућих пептида је готово немогуће извести, што је условило потребу за коришћењем рачунарских метода. У претходних двадесет година је развијен велики број метода за предвиђање пептида који се везују за *MHC* молекуле. Неке од тих метода се веома успешно користе за идентификовање Т - ћелијских епитопа. Постојеће методе за предвиђање Т - ћелијских епитопа су детаљно описане у поглављу 3, и истакнути су њихови недостаци. Да би се наведени недостаци смањили (или чак уклонили) потребно је дефинисати нове, бар исто толико тачне, поуздане и брзе методе.

Главни циљ ове тезе је био побољшање перформанси постојећих метода за предвиђања *MHC* везујућих пептида (Т - ћелијских епитопа[1]). Теза је фокусирана на методе засноване на техникама истраживања података, које у обзир узимају само информације из амино киселинске секвенце пептида за тренирање модела. Да би се направили добри модели неопходно је адекватно математичко представљање улазних података у облик погодан за примену техника истраживања података. Потребно је и пажљиво изабрати податке за прављење модела, адекватно тестирање направљених модела и коначну проверу добијених резултата. Како би се успешно остварили сви захтеви императив је и

[1] У наставку тезе се подразумева да се предвиђање Т - ћелијских епитопа односи на предвиђање *MHC* везујућих пептида

добро познавање области у којој се технике истраживања података примењују. У наставку је истакнут значај примене техника истраживања података за предвиђање епитопа и описан је процес препознавања Т - ћелијских епитопа од стране имуног система.

1.3 Технике истраживања података и предвиђање Т - ћелијских епитопа

За предвиђање Т - ћелијских епитопа тренутно су најпопуларније технике истраживања података јер најбоље балансирају однос цене и перформанси. Неке од ових техника не захтевају велики број података за тренирање како би се направио добар модел за предвиђање, што је за алеле где не постоји велики број експериментално утврђених података веома важно. Најзначајније технике истраживања података које се користе у постојећим методама за откривање Т - ћелијских епитопа су:

- Технике засноване на стаблима одлучивања (eng. *Decision trees*).
- Технике засноване на неурнским мрежама (eng. *Neural networks*).
- Технике засноване на подржавајућим векторима (eng. *Support vector machine*).
- Статистички засноване технике.
- Технике засноване на правилима придруживања (eng. *Association rules*).

Технике засноване на стаблима одлучивања су моћне и популарне технике моделирања података. Посебна погодност ове технике је једноставно представљање модела података у виду правила. Стабло се састоји од три врсте чворова: корена, унутрашњих чворова и листова. Правила се једноставно интерпретирају читањем и праћењем путања од корена ка листу. Оптимизација самог стабла и подстабла није једноставан задатак, али чињеница да су технике засноване на стаблима погодне и за линеарне и за нелинеарне проблеме их чини погодним за проблем класификације Т - ћелијских епитопа. Класификација Т - ћелијских епитопа стаблима одлучивања се заснива на

правилима где се класификују обрасци на основу секвенци са већ добро установљеним правилима. Мотиви на специфичним позицијама су преведени у правила која су укључена у чворове стабла. Структура резултујућег стабла укључује релевантне атрибуте, у конкретном случају то могу бити особине амино киселина које су укључене у представљање пептида. Примена модела заснованих на техници стабла одлучивања подразумева пропуштање пептидне секвенце кроз серију чворова, а резултати транзиције из чвора у чвор се користе за коначно предвиђање [40][39]. Развијен је велики број алгоритама за формирање стабла одлучивања [101]. Треба напоменути да је развојем *Random forest* [36] алгоритма ова техника постала изузетно популарна за решавање проблема у области биоинформатике.

Вештачке неуронске мреже представљају много погоднију технику за проналажење веза између атрибута улазних података, као и за представљање нелинеарних података. Модели зановани на неуронским мрежама су изузетно добри за класификационе проблеме као и за откривање сложених образаца. Ова техника је исцрпно коришћена за предвиђање Т - ћелијских епитопа [17][9][59][61][118][77]. Тренутно најбоље методе за предвиђање Т - ћелијских епитопа су засноване управо на овој техници [61][118]. Једино ограничење модела заснованих на овој техници је што захтевају фиксан улаз, тј. дужина пептида за који се врши предвиђање мора да буде иста као и дужина пептида за коју је прављен модел.

Технике засноване на подржавајућим векторима су у основи засноване на статистичким методама минимизације проблема структуралног ризика и векторским просторима. Веома успешну примену имају за решавање проблема класификовања слика, регресиону анализу и препознавање образаца. Слично као и технике засноване на неуронским мрежама могу да решавају и линеарне и нелинеарне проблеме, али се ова техника у решавању других врста проблема врло брзо показала као супериорнија у односу на вештачке неуронске мреже. И овде је главни недостатак примене ове технике што захтева улаз фиксне величине. Проблем се превазилази прављењем засебних модела за различите дужине пептида. Технике засноване на подржавајућим векторима су изузетно погодне за проблеме где је број примера за учење мали, а број атрибута велики. Неки од примера примене технике подржавајућих вектора за проналажење Т -

ћелијских епитопа су описани у радовима [74][41][105][22][112].

Скривени Марковљеви модели (eng. *Hidden Markovs Models*) је статистички заснована техника, која је постала је изузетно популарна у последњој деценији. Примену у предвиђању Т - ћелијских епитопа има управо јер надомешћују наведене недостатке претходно описаних метода. Модели направљени овом техником не захтевају фиксан улаз, што значи да је довољно направити само један модел за предвиђање епитопа произвољне дужине [64]. Детаљан опис технике се може наћи у [90], а интезивно се користи већ дуги низ година за проналажење сродних удаљених секвенци, издвајање познатих домена у новим секвенцама и равнање протеинских секвенци.

Техника правила придруживања у основи проверава све трансакције и проналази интересантна правила и обрасце у ставкама трансакције за које су задовољени услови да су минимална подршка и поверење веће од унапред дефинисане границе. Ова техника свакако није упоредива са техникама заснованим на неуронским мрежама и подржавајућим векторима, али оно што је чини значајном у решавању проблема у оквиру ове тезе је што трансакција не мора да буде фиксне дужине, а то значи да је примењива на пептиде различитих дужина. Примери примене ове технике за проналажење Т - ћелијских епитопа су описани у [70][115].

Свака од техника захтева припрему података у погодан облик за улаз у модел, најчешће у виду вектора где се пептид представља нумеричким дескрипторима особина амино киселина које улазе у његов састав. Избор одговарајуће технике за прављење модела као и добар избор атрибута за представљање података је кључан корак у прављењу доброг модела. Модели развијени у оквиру тезе користе технику засновану на подржавајућим векторима. Мотивација за избор техника је била: супериорност ове технике у односу на друге за нелинеарне проблеме, могућност прављења доброг модела и са мањим скупом података за учење и великим бројем атрибута и подршка и за проблеме класификације и регресије.

Технике истраживања података коришћење у оквиру тезе за развој предложених модела и добијање корисних образаца су детаљно описане у глави 2, док су постојеће, тренутно најактуелније, методе за откривање Т - ћелијских епитопа детаљно описане у глави 3.

1.4 Имуни систем - основни појмови

Антиген (skr. *Ag*, од првобитног eng. *Antibody generator*) је молекул кога препознаје имунолошки систем организма, док је епитоп регион или фрагмент антигена који се везује за одговарајуће рецепторе на *Ag* - везујућим ћелијама имунолошког система. Имунолошки систем чине организована ткива која бране организам од страних молекула, инфективних микроорганизама и њихових токсина. Постоје два типа имунолошког одговора:

(1) Урођени имунитет који је неспецифичан и без имунолошке меморије и чини прву линију одбране од страних микроорганизама.

(2) Адаптивни (стечени) имунитет који чине хуморални имунитет и ћелијски имунитет.

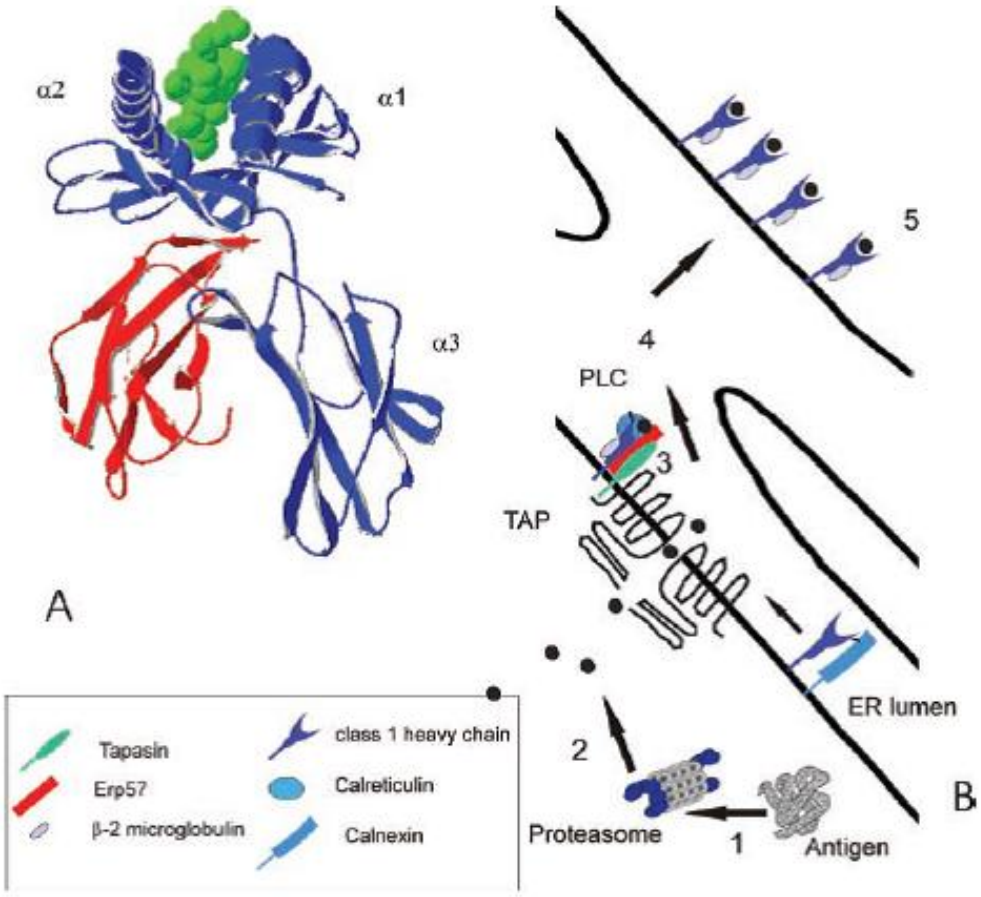
Адаптивни имунитет се јавља код кичмењака, специфичан је за одређени антиген, има имунолошку меморију и јавља се касније у току имунолошког одговора него урођени. Дели се на хуморални и ћелијски имунитет. Хуморални имунитет се тако назива јер се молекули протеина (антитела), који су главни носиоци овог типа имунитета, налазе у телесним течностима. Стварају их ћелије које се зову Б лимфоцити или Б ћелије. Антитела препознају антигене, неутралишу инфекције изазване микроорганизмима, тако што их уништавају различитим механизмима одбране. Хуморални имунитет је главни механизам одбране од микроорганизама који нападају ћелије споља, и усмерен је, углавном на просторне (нелинеарне или дисконтинуалне) епитопе антигена. Ћелијски имунитет (или ћелијама посредовани имунитет) се заснива на Т - лимфоцитима, и усмерен је на линеарне епитопе антигена. Грана ћелијског имунитета (*Th*, *Tr*) има улогу да регулише, како адаптивни, тако и урођени имунитет и одлучује какав тип имунолошког одговора тело индукује на одређени патоген. Усмерен је углавном на антигене из спољне средине, као што су *Ag* бактерија, (егзогени пут уношења *Ag*) које ћелије (назване „професионалне *Ag* - приказивачке ћелије"), уносе ендоцитозом, деградирају и „представљају" на ћелијској површини. Друга грана ћелијског имунитета су цитотоксични Т лимфоцити (*Tc*). Овај пут је усмерен, углавном на контролу сопствених, унутарћелијских протеина и елиминацију утрошених протеина (ендогени пут уношења *Ag*). Ако вирус

инфицира ћелију, вирални пептиди (епитопи) ће бити представљени преко овог пута, омогућујући *Tc* лимфоцитима да препознају и убију инфицирану ћелију. И Б и Т лимфоцити носе на ћелијској мембрани рецепторне молекуле (код Б лимфоцита су то антитела, а код Т лимфоцита Т - ћелијски рецептори, скр. *TCR*, од eng. *T-cell receptor*).

Имунолошки одговор чини препознавање антигена, активација лимфоцита и ефекторна фаза елиминације антигена. Адаптивни имунолошки одговори су иницирани препознавањем специфичних антигена. Адаптивни имунолошки систем сисара је еволуирао тако да излаже фрагменте (епитопе) протеина, који потичу од микробних патогена (антигена), као и сопствене протеине (као сталну контролу сопственог имунитета) ћелијама имунолошког система. Ове ћелије се деле на антиген - приказивачке, ефекторне и регулаторне. Фрагменти протеинске секвенце (епитопи) су пептиди, дужине до 25 амино киселина који се ослобађају из интактних протеина преко протеолитичких механизама који се одвијају у специјализованом органелама антиген - показивачких ћелија. У наредном кораку се преносе на површину ћелија у комплексу са протеинима главног хистокомпатибилног комплекса организама, да би их (у комплексу) препознале ефекторне ћелије имунолошког система. Ћелије имунолошког система које препознају комплексе су помажући/регулаторни (eng. *helper/regulatory*, скр. *Tr* или *Th*) Т лимфоцити који носе ознаке Т4 или $CD4^+$ и цитотоксични Т лимфоцити који носе ознаке Т8 или $CD8^+$. Молекули главног хистокомпатибилног комплекса су генски региони или фамилије гена. Састоје се од две подкласе *MHC* I и *MHC* II)[89]. Код човека носе назив *HLA* I и *HLA* II. Њихове комбинације представљају индивидуалну ткивну и имунолошку специфичност организма, која је генетски дефинисана (генским алелима класе *MHC* I и II). Постоји пет типова гена *HLA* молекула класе I : HLA-A, HLA-B, HLA-C, HLA-E и HLA-G, а за *HLA* молекуле класе II постоје три локуса: HLA-DP, HLA-DQ и HLA-DR. *HLA* генски алели су кодоминантни и у једном људском организму су најчешће изражени кроз 6 различитих молекула класе *MHC* I и 12 или више молекула *MHC* класе II. *HLA* локус је најполиморфнији познати генски систем. *HLA* алели представљају једну од више форми ДНК секвенце, а везују велики спектар различитих пептида, извучених из 1000 до 10.000 протеинских секвенци - антигена.

1.4.1 Препознавање Т - ћелијских епитопа

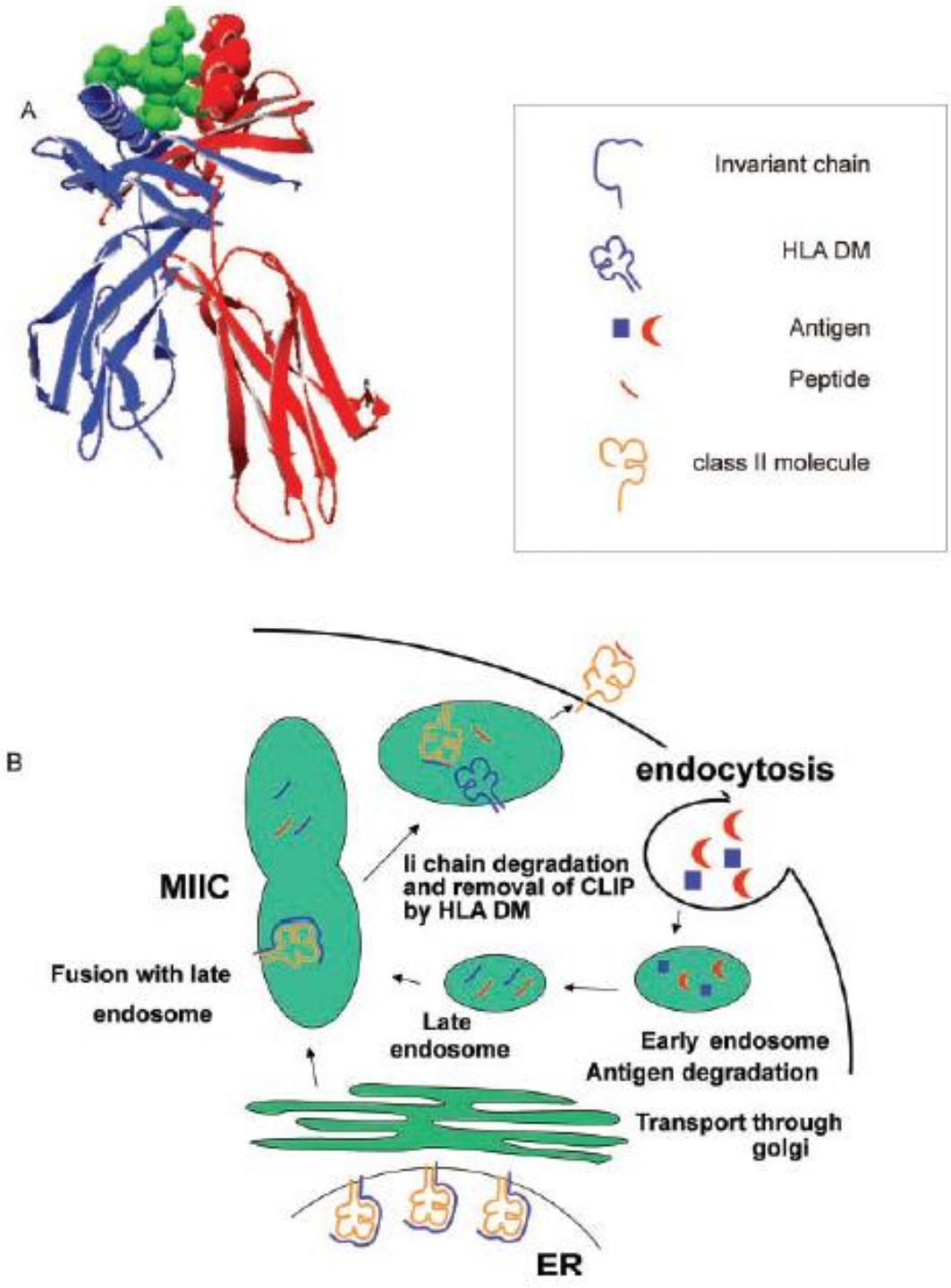
На слици 1.1 и слици 1.2 је приказана молекуларна структура *HLA* класе I и II и пут обраде антигена у обе класе.



Слика 1.1: *А, 3Д структура молекула MHC класе I. Хетеродимер класе I се понаша као платформа за коју се антигенски пептид везује (зелена боја). В, Процесирање антигена кроз MHC I пут. 1, Протеински антиген се деградира у цитоплазми у акцији протеосоматске протеазе (eng.*proteosome-protease*). 2, Пептиди генерисани у претходном кораку се транспортују у лумен на ER на ATP - зависан начин кроз акцију TAP хетеродимера. 3, Ланци настали повезивањем са молекулима класе I су мета ER и стабилизују се у интеракцији са 'пратиоцима'.4, Када се пептид са довољним афинитетом везивања повеже са хетеродимером класе I, цео комплекс се транспортује на површину ћелије, где га препознају $CD8^+$ Т ћелије [89].*

Путеви процесирања антигена класе *MHC* I и II се разликују, као и сама структура ових молекула. Самим тим је и процес препознавања Т - ћелијских епитопа који се везују за ове две класе другачији. Информатички приступ у идентификовању Т - ћелијских епитопа се заснива на добром моделовању

путева процесирања антигена.



Слика 1.2: *А, 3Д структура молекула МНС класе II. В, Пут процесирања антигена молекулима класе II. Егзогени антиген се у ендоцитози деградира у два корака: први је фаза ране а други фаза касне ендозоме [89].*

Поглавље 2

Методе истраживања података

Истраживање података (eng. *Data mining*) је интердисциплинарно поље информатике које се бави аутоматским или полу-аутоматским откривањем знања у подацима. Основни задатак истраживања података је издвајање нетривијалних, претходно непознатих и потенцијално корисних образаца, односа и веза у подацима и статистички значајних структура из великих колекција података. Императив је да добијени резултати буду нови, ваљани, корисни и разумљиви. Број података из године у годину расте, и постоје базе података које су величине неколико терабајта. Без аутоматских метода за издвајање података готово је немогуће истражити и анализирати податке. Током последњих година су развијени многобројни алгоритми и методе за истраживање података. Они укључују статистичке моделе, метематичке алгоритме и методе машинског учења. Истраживање података не укључује само велике колекције података и управљање подацима, већ и различите анализе и предвиђања. Постоји неколико различитих методологија за приступ овим проблемима. У оквиру ове тезе су, за предвиђање Т ћелијских епитопа, коришћене следеће технике: класификација, регресија, правила придруживања и кластеровање методом k -средина, које су и детаљно објашњење у наставку.

2.1 Класификација података

Класификација је једна врста предиктивног моделирања. Прецизније, класификација представља придруживање предефинисаног скупа класа, унапред познатог, новим улазним подацима. Улазни податак у процес

класификације је скуп слогова који се најчешће називају примери или инстанце. Сваки слог је облика (x, y) где је x скуп атрибута, а y је ознака класе (предвиђени или излазни атрибут). Дакле, свака инстанца улазног скупа је представљена својим атрибутима, док је комплетан улазни скуп представљен матрицом чије су врсте инстанце а колоне атрибути тих инстанци. Уколико је излазни атрибут категоричка вредност у питању је класификација, ако је пак излазни атрибут нумеричка вредност онда је у питању регресија. Уобичајено је да се улазни подаци у процесу класификације деле на два подскупа. Један подскуп чине подаци за учење (тренирање) класификационог модела (класификатора), а други скуп чине подаци за тестирање добијеног модела. Подаци за учење чине скуп слогова (x, y), где је y позната класа, а задатак је направити класификациони модел (пронаћи функцију одлучивања) који пресликава скуп атрибута x у одговарајућу предефинисану класу y. Циљ је прављење таквог модела који ће скупу слогова са познатим атрибутима и непознатим класама придружити одговарајући класу што прецизније. На основу података за тестирање се одређује тачност модела. Прављење класификационог модела је проблем учења под надзором. Тачност модела се одређује мерама за оцену описаним у поглављу 2.1.1. Процес класификације се углавном састоји из две фазе:

(1) **Фаза учења**: У овој фази се прави модел на основу направљеног подскупа података за учење.

(2) **Фаза тестирања**: У овој фази се модел примењује на податке из скупа за тестирање и придружује улазним подацима класу y. На основу поређења придружене класе са стварном класом се рачуна тачност добијеног модела.

Излаз из класификационог модела може бити:

- Дискретна вредност, нпр. када се доноси бинарна одлука (0 или 1), тј. да ли податак припада одређеној класи или не. У питању је чврста (eng. *hard*) класификација.

- Нумеричка вредност, у овом случају нумеричка вредност се придружује свакој инстанци за сваку од могућих класа из скупа података за тестирање. Ова нумеричка вредност може даље да се преведе у дискретну,

или скуп дискретних вредности где ће се касније утврдити граница припадности класи. Најчешћа примена оваквог вида класификације је код ретких класа, где је оригинална дистрибуција класа прилично неуравнотежена, а идентификовање неких класа је значајније од других. У питању је мека (eng. *soft*), класификација. Најчешће се улазном податку придружује вредност која је између 0 и 1, као мера припадности класи [111] [110].

У зависности од броја класа, класификација може бити:

- **Бинарна**, када су дефинисане само две могуће класе.
- **Вишекласна**, када је дефинисано више од две могуће класе.

У зависности од тога да ли се класе могу преклапати или не, класификација може бити:

- **Једнозначна** (eng. *single-label*), када једном податку може бити додељена тачно једна класа.
- **Вишезначна** (eng. *multi-label*), када једном улазном податку може бити додељен произвољан број (> 1) класа.

Уколико се у току класификације посматрају самостално без икакве структуре која дефинише односе између њих, тада се ради о нехијерархијској организацији. Када је број различитих класа јако велики, јавља се потреба за организовањем класа ради тачнијег и прецизнијег претраживања, и тада се ради о хијерархијској класификацији [11]. Класа се тада дели у мање подкласе.

2.1.1 Оцена квалитета модела класификације

За неке моделе се може рећи да су бољи од других у смислу тачности, док се за друге може сматрати да су бољи уколико су једноставнији или разумљивији. Ипак постоје мере којима се оцењује квалитет модела. Важно је проценити колико је добро модел успео да уопшти решење проблема на основу скупа података за учење. Најчешћи проблем који се јавља је "превише прилагођен" модел (eng. *overfitting*). Превише прилагођен модел је онај који се добро понаша над подацима за учење, али не и над подацима за тестирање. У том случају

није добро уопштено решење проблема. Узрок овог проблема је врло често шум који се јавља у улазним подацима а који није елиминисан пре прављења модела. Такође, један од разлога може да буде недостатак репрезантитавних података у скупу за учење или недовољан број података. Други проблем је када се модел понаша недовољно добро и над подацима за тренирање и тестирање (eng. *underfitting*). Оба проблема су уско везана са сложеношћу направљеног модела. Идеална сложеност је она где модел има најмању грешку генерализације (најмање погрешно класификованих података у скупу за тестирање). Како је у процесу прављења модела једино познат скуп података за учење, не може се одредити грешка генерализације унапред. За тестирање модела се мора користити независтан скуп података, а то су подаци који нису коришћени у процесу учења. Процес оцене квалитета модела се састоји у поређену унапред познате класе са оном коју је предложио модел. На тај начин се добија непристрасна оцена грешке генерализације модела. Могући исходи код бинарне класификације су:

- Стварно позитивни, (eng. *true positives*, TP)
- Стварно негативни, (eng. *true negatives*, TN)
- Лажно позитивни, (eng. *false positives*, FP)
- Лажно негативни, (eng. *false negatives*, FN).

Перформансе класификатора могу да се представе и *матрицама конфузије* (eng. *confusion matrix*). Свака од предефинисаних класа представља по једну врсту и једну колону у матрици. Збир вредности по колонама представља број инстанци које је модел придружио тој класи, док је збир вредности у врстама број стварних инстанци те класе (видети слику 2.1). Елементи главне дијагонале су исправно класификоване инстанце. Дефинисане мере за оцену квалитета класификационог модела, на основу матрице конфузије [29] су следеће:

- Прецизност (eng. *Precision*, (P)) оцењује колики проценат примера за тестирање је исправно класификован:

$$P = \frac{TP}{TP + FP}$$

Слика 2.1: *Матрица конфузије*

- Одзив (eng. *Recall*, R) оцењује колико је модел успешан у покривању класе односно колико примера за тестирање из дате класе модел може да препозна:

$$R = \frac{TP}{TP + FN}$$

- F-мера (eng. *F measure*, F) представља комбинацију претходне две мере: прецизности и одзива, у једној мери, представљеној као њихова хармонијска средина:

$$F = \frac{2 * P * R}{P + R}$$

- Тачност (eng. *Accuracy*, Acc) представља проценат тачно класификованих података, а корисна је само уколико су класе исте или сличне величине:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

- Степен грешке (eng. *error rate*, ER) представља проценат погрешно класификованих података:

$$ER = \frac{FP + FN}{TP + TN + FP + FN}$$

Све приказане мере су на нивоу једне класе (у конкретним примерима на нивоу позитивне класе), и рачунају се појединачно за обе класе. Квалитет класификационог модела може да се посматра и на глобалном нивоу када се ове мере усредњавају. Усредњавање може да буде урађено на два начина, као:

- **Макро просек**, када се свакој класи придаје исти значај. Израчунавају се одговарајуће мере за сваку од класа а затим се добијене суме поделе са бројем класа.

- **Микро просек**, где се фаворизује једна класа, која има највећи број инстанци на следећи начин:
  - Израчунавају се вредности за TP, TN, FP, FN за сваку од класа појединачно.
  - Израчунавају се нове вредности $\overline{TP}$, $\overline{TN}$, $\overline{FP}$, $\overline{FN}$ као суме свих TP, TN, FP, FN за појединачне класе.
  - Израчунавају се нове мере (тачност, прецизност, одзив, F - мера) за тако добијене $\overline{TP}$, $\overline{TN}$, $\overline{FP}$, $\overline{FN}$.

Мере за оцену квалитета модела на основу *ROC* (eng. *Receiver Operating Characteristic*) кривих [32][114]:

- *ROC* крива је алтернатива мери тачности (*Acc*), која за разлику од тачности не представља једну конкретну вредност већ криву која се не интерпретира једном статистичком оценом.
- *AROC* представља површину испод *ROC* криве, као једну нумеричку вредност која представља сумарну оцену перформанси модела. Најчешће се користи за оцену квалитета класификатора који као резултат враћају вероватноћу припадности класи, а израчунава се на следећи начин:

$$A_{ROC} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{PN} \int_0^N TP dFP$$

  где P представља број инстанци позитивне класе, а N број инстанци негативе класе. Случајни класификатор (eng. *random classifier*) има AROC вредност 0.5, док савршени класификатор има вредност 1. У пракси би добар класификатор требало да узима вредности између 0.5 и 1, пожељно је да буде што ближе 1.

Да би се направио добар модел и проценио његов квалитет примењују се још и следећи поступци:

- **Издвајање мањег скупа** (eng. *Holdout method*) из основног скупа података са познатим (придруженим) класама. Овај поступак је већ поменут у уводном делу о класификацији. Подела скупа података на подскупове за тренирање и тестирање може бити у различитом односу

(50:50, 70:30, 80:20). Тада се квалитет модела оцењује на основу мера тачности утврђених на тестном скупу. Овај приступ има неколико ограничења: скуп података за тренирање модела се редукује, тиме направљени модел не може бити толико добар колико модел који би био направљен на комплетном скупу података. Квалитет модела јако зависи од композиције скупа за тренирање и тестирање: ако је скуп за тренирање мањи модел ће бити лошији, ако је скуп података за тестирање мањи оцена квалитета модела ће бити мање поуздана.

- **Издвајање случајних узорака** (eng. *Random Sampling*). Овај поступак подразумева примену претходног поступка више пута како би се побољшале порформансе класификатора. Ако је на пример Acc тачност модела у i - тој итерацији, укупна тачност модела се рачуна као $Acc_{sum} = \sum_{i=1}^{k} Acc_i/k$. И овде је проблем као у претходном случају то што се не узима комплетан скуп података за тренирање модела. Такође може да се деси да се неке инстанце много чешће јављају у скупу за тренирање него друге.

- **Унакрсна провера** (eng. *Cross - Validation*, CV) је алтернатива претходном приступу. У овом приступу се сваки слог користи исти број пута за тренирање модела и само једном за тестирање. Поступак се може објаснити на следећи начин: претпоставимо да је основни скуп података подељен на два једнака дела. У првом кораку, један од подскупова се користи за учење а други за тестирање. У другом кораку, подскупови имају замењене улоге онај који се користио за учење се користи за тестирање и обрнуто. Овај поступак је назван 2-унакрсна провера. Укупна грешка модела се рачуна као средња вредност грешака појединачних модела. k-унакрсна провера је уопштење претходног примера, и најчешће се користе 5 или 10 – унакрсне провере. У k-унакрсној провери један подскуп се користи за тестирање, а остакак $k-1$ подскупова за тренирање модела. Поступак се понавља k пута, тако да се сваки од k подскупова користи за тестирање тачно једном. Специјалан случај унакрсне провере је када је $k = N$ где је N укупан број слогова (eng. *Leave-one-out*, LOO). У сваком кораку се тачно један слог оставља за тестирање. Предност овог приступа је у томе што се користи скоро цео скуп података за учење. Недостатак је осим што је ова метода са рачунарске тачке гледишта скупа,

и то што се тестни скуп састоји од само једног слога због чега варијација у перформансама добијених модела може бити велика.

- **"Повећавање"** (eng. *Bootstrap*) основног скупа за тренирање. У претходним примерима се подразумева да скуп података за учење модела нема понављања. Дакле нема дуплих слогова ни у тренинг ни тест скупу. У овом приступу се бира скуп података за тренирање, а затим се изабрани подаци враћају у основни скуп где опет могу бити изабрани у следећој итерацији бирања тренинг скупа. Ако оригинални скуп података има N инстанци, може се показати да овом методом скуп података за тренирање садржи у просеку 63.2% података из основног скупа. Ова апроксимација следи из вероватноће да се изабере један слог из основног скупа која је $1-(1-\frac{1}{N})^N$. За довољно велико N вероватноћа се апроксимира са $1-e^{-1}=0.632$. Слогови који нису укључени у избор тренинг скупа на овај начин, се бирају за тестни скуп. Модел добијен на овако направљеном тренинг скупу се тестира на тестном скупу, добијена тачност модела је ε_i. Процедура избора узорака за тренирање се понавља b пута. Постоји неколико варијација ове методе, у смислу како се укупна тачност модела рачуна. Најчешће коришћени приступ је *.632 bootstrap*, где су укупна тачност модела рачуна комбиновањем свих добијених b тачности ε_i, на подмоделима, и тачности добијене на тренинг скупу који садржи оригиналне слогове са већ познатим лабелама acc_s, на следећи начин:

$$Accuracy, acc_{boot} = \frac{1}{b}\sum_{i=1}^{b}(0.632 \times \varepsilon_i + 0.368 \times acc_s)$$

Постоји велики број примера употребе класификације, јер ознака класе може да представља произвољну особину. Неки од примера примене су следећи:

- Циљани избор маркетинга (eng. *Customer Target Marketing*)
- За дијагнозу болести у медицини
- Мултимедијална анализа података
- Анализа билошких података
- Категоризација документа

- Анализа друштвених мрежа

Методе истраживања података које се најчешће користе за класификационе проблеме су: методе засноване на стаблима одлучивања (eng. *decision trees*), методе засноване на правилима (eng. *rule based*), методе засноване на растојању (eng. *nearest neighbour*), методе засноване на неуронским мрежама (eng. *neural networks*), методе засноване на подржавајућим векторима (eng. *support vector machine*) и статистички засноване методе. У наставку су детаљно описане методе које су коришћење у тези.

2.1.2 Класификација техником подржавајућих вектора

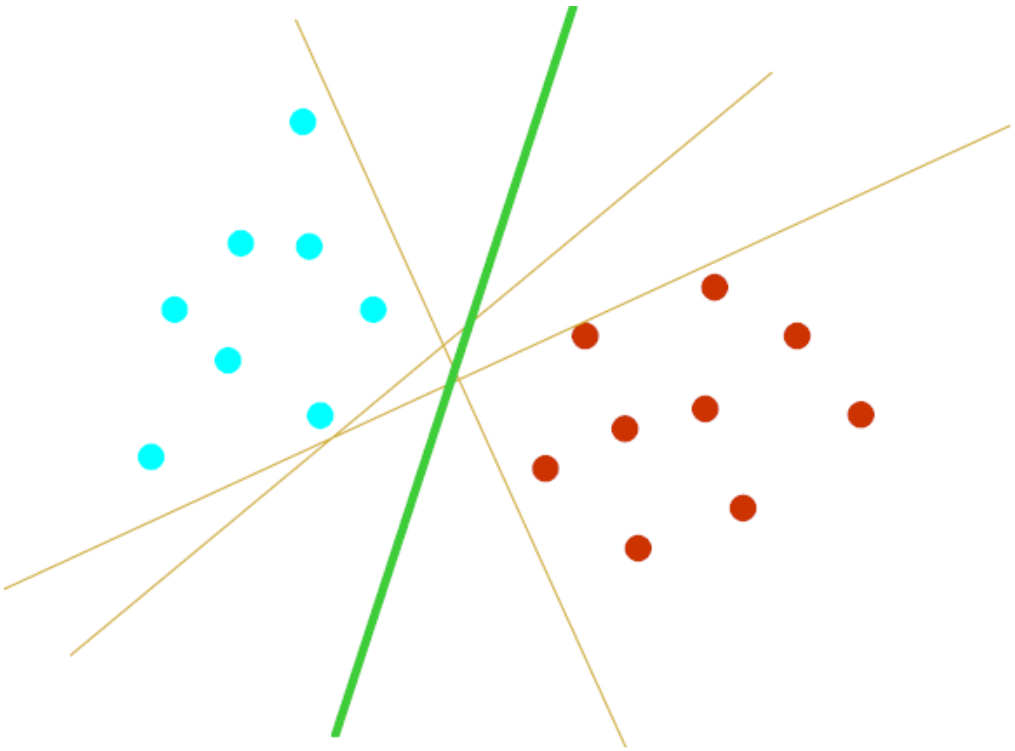
Оригиналан алгоритам заснован на техници подржавајућих вектора (eng. *Support vector machine*, SVM) увели су *Vapnik* и *Chervonenkis* 1963. године [108]. *Vapnik* је у свом раду у [13] предложио први начин за прављење нелинеарног класификатора увођењем кернел функција, а *Cortes* у [19] је први пут предложио увођење меке маргине. Техника је детаљно објашњена у наставку.

Линеарно раздвојиви подаци

Класификациони проблем се може ограничити на разматрање проблема две класе без губљења општости. У том случају је циљ пронаћи функцију која раздваја две класе из доступних примера за учење. Односно циљ је прављење класификатора који ће се добро понашати и над подацима који нису учествовали у учењу. У случају линеарно раздвојивих података (eng. *linearly separable*), пример приказан на слици 2.2, хиперраван која представља границу одлучивања је права. Сви примери који су са једне стране те праве припадају једној класи а сви примери који су са друге стране праве припадају другој класи. Нека је дат скуп података:

$$\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}, x \in R^N, y \in \{-1, +1\}$$

Бесконачно много правих задовољава услов. Треба изабрати једну од тих правих која представља најбоље решење односно максимизира растојање од праве и најближих тачака обе класе.



Слика 2.2: *Пример линеарно раздвојивог скупа.*

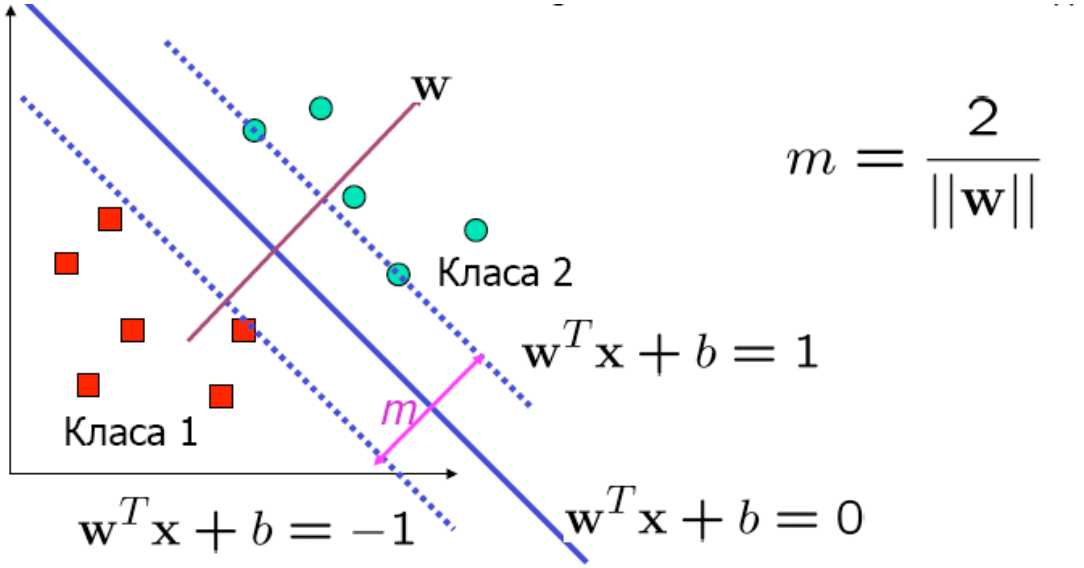
Јасно је да ће раван која је сувише близу елемената (тачака) за обуку бити осетљивија на шум, тако да је могуће да у том случају последица буде лошији модел над подацима ван скупа за обуку (тестним примерима). Насупрот томе, хиперраван која је највише удаљена од свих елемената за обуку би требало да се покаже као добра и у општем случају (на примерима за тестирање). Према томе, оптимална хиперраван ће бити раван са највећом маргином, при чему се маргина дефинише као минимално растојање између елемената за обуку и површи одлучивања (тачака које су близу потенцијалне линије раздвајања). Резултат тога је да је раздвајајућа хиперраван потпуно одређена специфичним подскупом елемената за обуку, који се зову подржавајући вектори (eng. *support vectors*), по чему је метода и добила име. Будући да нам је циљ максимизација маргине, можемо је изразити у функцији тежинског вектора w и тежинског прага-помераја b хиперравни. Једначина хиперравни $\Pi(w, b)$ се може дефинисати изразом $w^T x+b = 0$ и она је потпуно одређена параметрима (w, b). x је елемент за учење (има их укупно N). Параметар w представља коефицијент правца и одређује смер хиперравни, док је b померај, и одређује удаљеност хиперравни од центра координатног система. Растојање између тачке x и равни

$\Pi(w, b)$ је дата изразом:

$$d(x, \Pi) = r = \frac{w^T x + b}{\|w\|}$$

Овако задато растојање може бити и негативно. Маргина m се дефинише као ширина раздвајања између класа коју треба максимизовати (видети слику 2.3). Канонске вредности за w и b одређују се тако да је раздаљина најближих тачака (подржавајућих вектора) једнака 1 по апсолутној вредности. Растојање подржавајућих вектора од резултујуће равни биће $r_1 = \frac{1}{\|w\|}$ а дебљина маргине биће $m = 2r_1 = \frac{2}{\|w\|}$. Да би раздаљина најближих тачака од хиперравни била 1 по апсолутној вредности први услов који се поставља је $|w^T x_i + b| = 1$. Ова раван је позната као канонична хиперраван. Растојање од најближег примера за учење до равни износи:

$$\frac{|w^T x + b|}{\|w\|} = \frac{1}{\|w\|}$$



Слика 2.3: *Пример линеарно раздвојивог скупа, максимизација маргине* m.

Маргина у овом случају износи $\frac{2}{\|w\|}$. За сваки тачку (x_i, y_i), услов раздвајања може да се формулише на следећи начин:

Дефиниција 2.1 *Наћи* w *и* b *тако да се максимизује* m*, уз услов* $w^T x_i + b \geq 1$ *ако је* $y_i = 1$*, односно* $w^T x_i + b \leq -1$ *ако је* $y_i = -1$*.*

Како важи $min\|w\| = max\frac{1}{\|w\|}$, проблем се може формализовати и као:

Дефиниција 2.2 *Наћи w и b тако да се минимизује $f(w) = \frac{1}{2}\|w\|^T\|w\|$ (услов максималне маргине) уз услов $y_i(w^T x_i + b) \geq 1$ (услов раздвајања).*

Математичко решење проблема

Проблем максимизације маргине се своди на познати квадратни оптимизациони проблем са линеарним условима. Како је функција $f(w)$ квадратна, значи да има јединствени глобални минимум. За решавање проблема ове врсте користи се класична Лагранжова оптимизациона техника. Минимизација функције $f(w) = \frac{1}{2}\|w\|^2$ под ограничењем, решава се увођењем Лагранжове функције:

$$L_p(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N}(\alpha_i \left[y_i(w^T x_i + b) - 1\right]$$

На овај начин се долази до оптимизационог проблема без ограничења који се решава:

- минимизацијом L_p с обзиром на основне промењљиве w и b, и
- максимизацијом L_p с обзиром на споредне промењљиве $\alpha_i \geq 0$ (тзв. Лагранжове мултипликаторе).

Представљени проблем се назива *Лагранжов основни проблем.* Да би се поједноставио основни проблем, елиминишу се основне промењљиве (w, b) користећи $\partial f/\partial z = 0$. Диференцирањем $L_p(w, b, a)$ с обзиром на w и b и изједначавањем извода са нулом добија се:

$$\frac{\partial L_p(w, b, a)}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\frac{\partial L_p(w, b, a)}{\partial b} = 0 \Rightarrow 0 = \sum_{i=1}^{N} \alpha_i y_i$$

Развијањем израза L_p добија се

$$L_p(w, b, a) = \frac{1}{2}w^T w - \sum_{i=1}^{N} \alpha_i y_i w^T x_i - b\sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i$$

Користећи вредности за w добијене из првог услова диференцирања, први израз у L_p може да се запише у облику:

$$w^T w = w^T \sum_{i=1}^{N} \alpha_i y_i x_i = \sum_{i=1}^{N} \alpha_i y_i w^T x_i = \sum_{i=1}^{N} \alpha_i y_i (\sum_{j=1}^{N} \alpha_j y_j x_j)^T x_i = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j {x_i}^T x_j$$

На исти начин може да се напише други члан у изразу L_p, док је трећи члан у изразу једнак нули на основу другог оптимизационог услова. Спајањем добијених израза добија се:

$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j {x_i}^T x_j$$

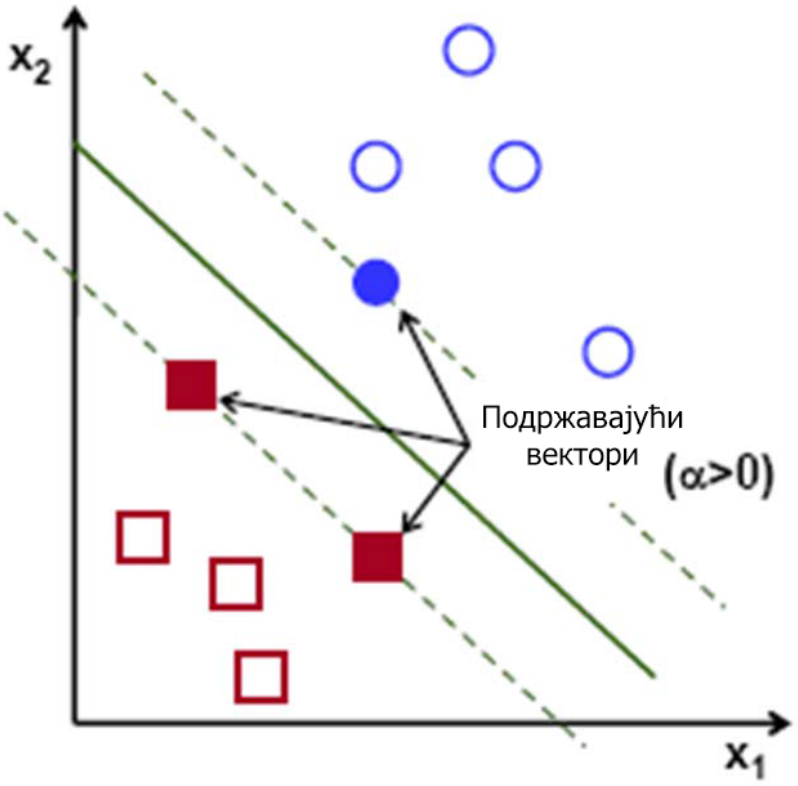
Сада је потребно да важе много једноставнија ограничења: $\alpha_i \geq 0$ и $\sum_{i=1}^{N} \alpha_i y_i = 0$. Овај проблем је познат као *Лагранжов дуални проблем*. Основни Лагранжов проблем је на овај начин трансформисан у много једноставнији проблем максимизације $L_D(\alpha)$ који зависи само од Лагранжових мултипликатора α а не и од променљивих w и b. Лагранжов основни проблем расте са димензионалношћу јер w има један коефицијент за сваку димензију, док је дуални проблем сразмеран са количином података за учење (постоји један Лагранжов мултипликатор за сваки пример). У оквиру израза $L_D(\alpha)$, подаци за учење се појављују само у форми скаларног производа $x_i^T x_j$. Управо ово својство се користи када је потребно вршити класификацију у вишедимензионом простору. За сваки пример (инстанцу) из скупа за учење мора да важи следећа једнакост:

$$\alpha_i \left[y_i (w^T x_i + b) - 1 \right] = 0, \forall i = 1, ..., N$$

За сваки пример из тренинг скупа је или $\alpha_i = 0$ или $y_i(w^T x_i + b) - 1 = 0$.

- Тачке за које важи $\alpha_i > 0$ припадају једној од две хиперравни које дефинишу највећу маргину јер само за њих важи да је $y_i(w^T x_i + b) - 1 = 0$. Ове тачке представљају подржавајуће векторе, приказане на слици 2.4.
- За све остале тачке важи да је $\alpha_i = 0$.

Претходне једнакости повлаче да само подржавајући вектори дефинишу оптималну хиперраван $\frac{\partial L_p(w,b,\alpha)}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i$. Вредности за b



Слика 2.4: *Пример подржавајућих вектора на линеарно раздвојивом скупу.*

се проналазе на основу комплементарног услова који важи за подржавајуће векторе. Закључак је да ако се читав скуп података замени само подржавајућим векторима хиперраван остаје иста.

Линеарно нераздвојив скуп података

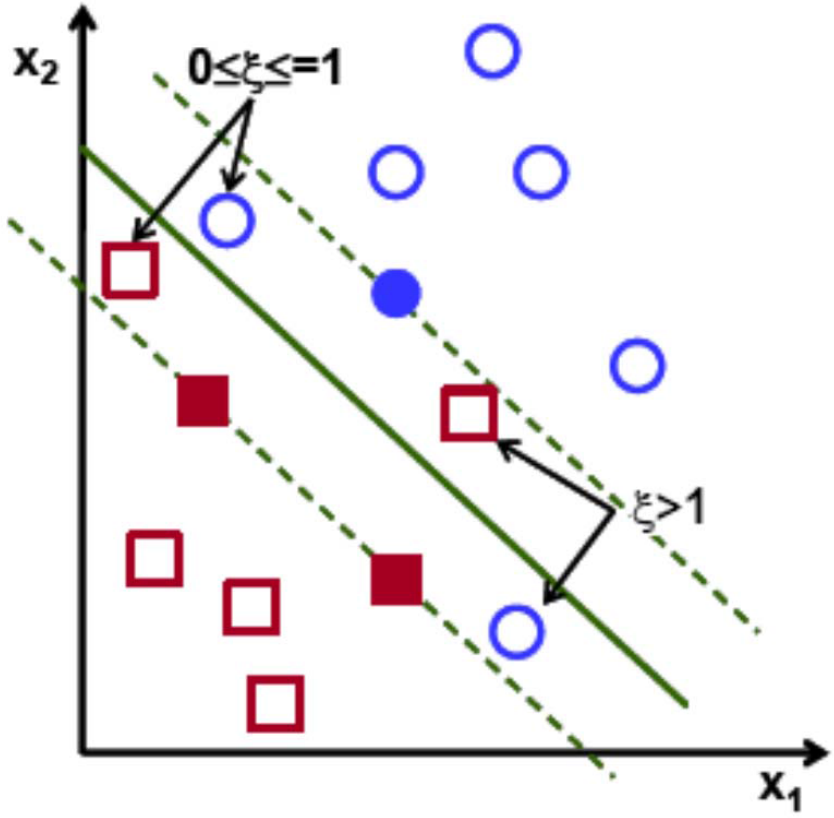
У случају линеарно нераздвојивог скупа података се уводе додатне ослабљене променљиве (eng. *slack variables*) ξ_i. За уведене променљиве важе ослабљена ограничења једначине каноничне хиперравни.

$$|y_i(w^T x_i + b)| \geq 1 - \xi_i, \forall i = 1, ..., N$$

Додатне ослабљене променљиве мере одступање од идеалног случаја.

- У случају када важи $0 \leq \xi \leq 1$ узорак се налази на исправној страни хиперравни раздвајања, али на растојању мањем од маргине (видети слику 2.5).
- У случају када важи $\xi > 1$ узорак се налази на погрешној страни

хиперравни.



Слика 2.5: *Пример линеарно нераздвојивог скупа, увођење ослабљених променљивих. Класификација меком маргином.*

Сада се оптимизациони проблем проналажења оптималне хиперравни мења, и циљ је пронаћи хиперраван којом се погрешна класификација своди на минимум. Односно, уводи се функција циља:

$$\Phi(\xi) = \sum_{i=1}^{N} I(\xi_{i-1})$$

$$I(\xi) = \begin{cases} 0 & \text{ako je } \xi \leqslant 0 \\ 1 & \text{ako je } \xi > 0 \end{cases}$$

$\Phi(\xi)$ представља укупан број погрешно класификованих примера. Проблем минимизације ове функције представља НП комплетан проблем (eng. *Nondeterministic Polynomial*) због нелинеарности индикационе функције $I(\xi)$. Из тог разлога се $\Phi(\xi)$ апроксимира следећом сумом:

$$\Phi'(\xi) = \sum_{i=1}^{N} \xi_i$$

која представља ограничење са горње стране у броју погрешно класификованих примера и уврштава се у функцију циља коју треба минимизовати:

$$f(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$

са условом да важи $y_i(w^T x_i + b) \geq 1 - \xi_i, \forall i = 1, ..., N$ и $\xi_i \geq 0, \forall i = 1, ..., N$. Параметар C предствља неку врсту компромиса између капацитета и броја погрешно класификованих примера. Што је већи параметар C то је број погрешно класификованих примера мањи. Што је мања вредност параметра C то су решења мање комплексна. Овај параметар се из тога разлога тумачи као регуларизациони параметар. Погодна вредност овог параметра се обично утврђује емпиријски унакрсном провером. Полазећи од сличне процедуре као у случају линеарно раздвојивог скупа података, добија се дуални проблем у следећем облику:

$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

са ограничењима $\sum_{i=1}^{N} \alpha_i y_i = 0$ i $0 \leq \alpha_i \leq C, \forall i = 1, ..., N$. Дакле, долази се до истог оптимизационог проблема као у случају линеарно раздвојивог скупа података. Изузетак су ограничења $\alpha_i \geq 0$ која су замењена много стриктнијим ограничењима $0 \leq \alpha_i \leq C$. Оптимално решење за тежински вектор је облика:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

Тежински праг може да се нађе избором примера из скупа за учење за које важи $0 \leq \alpha_i \leq C$ $(\xi_i = 0)$, и решавањем услова:

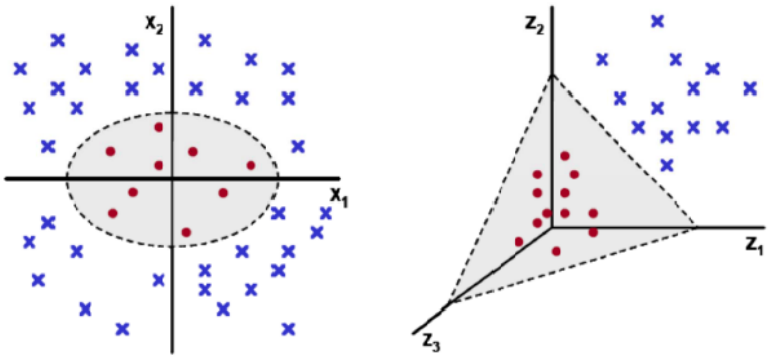
$$\alpha_i \left[y_i(w^T x_i + b) - 1 + \xi_i \right] = 0$$

Значај и снага метода подржавајућих вектора је заснована на чињеници да оне представљају ефикасну имплементацију принципа датог у Коверовој теореми:

Теорема 2.1 *Вероватније је да ће комплексан проблем класификације облика бити линеарно раздвојив уколико је нелинеарно пресликан у вишедимензиони простор, него у оригиналном нижедимензионалном простору.*

Методе засноване на подржавајућим векторима раде у две фазе. Прва фаза представља нелинеарно пресликавање примера за учење у вишедимензиони простор који је сакривен од улаза и излаза. Друга фаза је конструкција оптималне хиперравни раздвајања у вишедимензионалном простору.

$$\varphi : R^2 \to R^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) = \left(x_1^2, \sqrt{2}\, x_1 x_2, x_2^2\right)$$



Слика 2.6: *Нелинеарно пресликавање примера из дводимензионалног скупа података у тродимензионални простор.*

Недостаци ове методе су следећи:

- Статистички, јер рад у вишедимензионом простору је отежан због проблема димензионалности и додатног ризика од преприлагођавања.
- Рачунарски, јер рад у простору веће димензије захтева и више рачунарске снаге, чиме се ограничава и величина проблема који може бити разматран.

Функције језгра (Кернел функције)

Набројани недостаци се ефикасно решавају могућношћу упрошћавања, захваљујући класификацији на основу највеће маргине. Пресликавање у вишедимензиони простор је само имплицитно, јер сва решења зависе само од скаларног производа $\langle x_i, x_j \rangle$ елемената за обуку. Тачније, операције

у вишедимензионалном простору $\varphi(x)$ не морају експлицитно да се изводе уколико се нађе функција $K(x_i, x_j)$ таква да важи $K(x_i, x_j) = \langle\varphi(x_i), \varphi(x_j)\rangle$. Функција $K(x_i, x_j)$ се зове функција језгра или кернел функција. На слици 2.6 је дат пример прецликавања података из дводимензионалног простора R^2 у простор димензије 3 R^3 функцијом φ. Ако се за кернел функцију узме $K(x_i, x_j) = (x_i^T x_j)^2$ циљ је да се пронађе нелинеарно пресликавање $\varphi(x_i)$ такво да важи $(x_i^T x_j)^2 = \varphi(x_i)^T \varphi(x_j)^2$. Развијањем израза за кернел функцију се добија:

$$K(x_i, x_j) = (x_i^T x_j)^2 = ((x_{1,1}, x_{1,2})^T (x_{2,1}, x_{2,2}))^2 = (x_{1,1}x_{2,1} + x_{1,2}x_{2,2})^2 =$$

$$= x_{1,1}^2 x_{2,1}^2 + 2x_{1,1}x_{2,1}x_{1,2}x_{2,2} + x_{1,2}^2 x_{2,2}^2 =$$

$$= (x_{1,1}^2, \sqrt{2}x_{1,1}x_{1,2}, x_{1,2}^2)^T (x_{2,1}^2, \sqrt{2}x_{2,1}x_{2,2}, x_{2,2}^2)$$

где је $x_{i,k}$ k - та координата примера x_i. На тај начин користећи кернел функцију $K(x_i, x_j) = (x_i^T x_j)^2$ имплицитно се ради у вишедимензионом простору дефинисаним пресликавањем $\varphi(x)$. Скаларни производ $\varphi(x_i)^T \varphi(x_j)^2$ може да се израчуна у оригиналном простору R^2, без потребе за пројектовањем у R^3 захваљујући кернелу $(x_i^T x_j)^2$. У општем случају, под претпоставком да вектор x припада простору R^D први корак је нелинеарна пројекција вектора x на вишедимензиони имплицитни простор $\varphi(x) \in R^{D_1}(D_1 > D)$ у којем је вероватније да ће класе бити раздвојиве. Хиперраван раздвајања у новом простору R^{D_1} биће дефинисана са:

$$\sum_{j=1}^{D_1} w_j \varphi_j(x) + b = 0$$

Да би се елиминисао праг b вектору примера се додаје у имплицитном простору још једна константна димензија $\varphi_0(x) = 1$. Тада резултујућа хиперраван постаје:

$$w^T \varphi(x) = 0$$

Оптимална хиперраван, односно хиперраван са максималном маргином, на

основу свега претходног постаје:

$$w = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i)$$

Спајањем израза за оптималан тежински вектор и једначину хиперравни добија се:

$$w^T \varphi(x) = 0 \Rightarrow (\sum_{i=1}^{N} \alpha_i y_i \varphi(x_i))^T \varphi(x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i)^T \varphi(x_i) = 0$$

Како је $\varphi(x_i)^T \varphi(x_i) = K(x_i, x_j)$, оптимална хиперраван постаје:

$$\sum_{i=1}^{N} \alpha_i y_i K(x_i, x_j) = 0$$

Ово значи да се класификација непознатих примера врши израчунавањем пондерисане суме кернел функција с обзиром на подржавајуће векторе x_i, јер само подржавајући вектори имају ненулте дуалне променљиве α_i. Дуалне променљиве α_i се у имплицитном простору рачунају исто као и пре, с тим да се скаларни производ $\varphi(x_i)^T \varphi(x_i)$ замењује кернел функцијом $K(x_i, x_j)$. Лагранжов дуални проблем за нелинеаран SVM има облик:

$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i^T, x_j)$$

са ограничењима $\sum_{i=1}^{N} \alpha_i y_i = 0$ и $0 \leq \alpha_i \leq C, \forall i = 1, ..., N$. За кернел функцију може бити изабрана било која функција за коју постоји имплицитно пресликавање, односно кернел који може да се изрази као скаларни производ два вектора. Свака кернел функција мора да задовољава Мерсеров услов [69].

Теорема 2.2 *Под претпоставком да је $k \in L_\infty(\chi^2)$ такав да је дефинисан оператор интеграције $T_k : L_2(\chi^2) \to L_2(\chi^2)$, тада је*

$$T_k f(\cdot) = \int_x K(\cdot, x) f(x) d\mu x K(\cdot, x) f(x) d\mu x$$

позитиван. Нека је $\Psi_j L_2(\chi^2)$ сопствена функција од T_k са сопственим

вредностима $\lambda_j \neq 0$ *и нека важи да је норма* $\|\Psi_j\|_{L_2} = 1$ *затим нека је са* $\overline{\Psi_j}$ *означена њена комплексна конјугована функција. Тада*

1. $(\lambda_j(T))_j \in l_1$
2. $\Psi_j \in L_\infty(\chi)$ *и* $sup_j\|\Psi_j\|_{L_\infty}$
3. $K(x_i, x^{'}) = \sum_{i=1}^{N} \lambda_j \overline{\Psi_j(x)}\Psi_j(x^{'})$ *важи за скоро све* $(x_i, x^{'})$

Ова теорема у ствари значи да ако важи:

$$\int_{X\times X} K(x_i, x^{'})f(x)f(x^{'})dxdx^{'} \geq 0, \forall f \in L_2(\Psi)$$

тада се кернел функција $K(x_i, x^{'})$ може изразити преко скаларног производа у неком другом простору више димензије. Из ове теореме следе и последице о композицији функција језгра, које такође задовољавају Мерсеров услов.

- Последица 1: Линеарна комбинација кернел функција такође може бити кернел функција. Ово тврђење је директна последица линеарности интеграла.
$$K(x, x^{'}) = c_1K_1(x, x^{'}) + c_2K_2(x, x^{'})$$
- Последица 2: Ако је $s(x, x^{'})$ симетрична функција у односу на своје аргументе на $X \times X$, онда следећа функција може бити кернел функција:
$$K(x, x^{'}) = \int_x s(x, z)s(x^{'}, z)dz$$

Теорема 2.3 *Транслационо инваријантно језгро* $K(x, x^{'}) = K(x - x^{'})$ *је прихватљиво као кернел функција за методе подржавајућих вектора ако и само ако важи да је Фуријеова трансформација:*

$$F[k](w) = (2\pi)^{-\frac{d}{2}} \int_x e^{-i\langle w,x\rangle}k(x)dx$$

не-негативна [100].

У даљем тексту су наведени примери функција језгра које задовољавају Мерсеров услов:

- Хомогена полиномијално језгро уведено у [6] где је урађено експлицитно пресликавање и утврђено да за $p \in N$

$$K(x, x^{'}) = \langle x, x^{'} \rangle^{p}$$

јесте кернел функција. Степен полинома је параметар који дефинише корисник. Последица овога је и да нехомогено језгро

$$K(x, x^{'}) = (\langle x, x^{'} \rangle + c)^{p}$$

за $p \in N$ и $c > 0$ такође јесте прихватљива кернел функција.

- Функција хиперболичког тангенса такође задовољава услове кернел функције[1]

$$K(x, x^{'}) = tanh(\vartheta + \Phi\langle x, x^{'} \rangle)$$

- Радијално засноване функције (eng. *Radial basis function*, RBF):

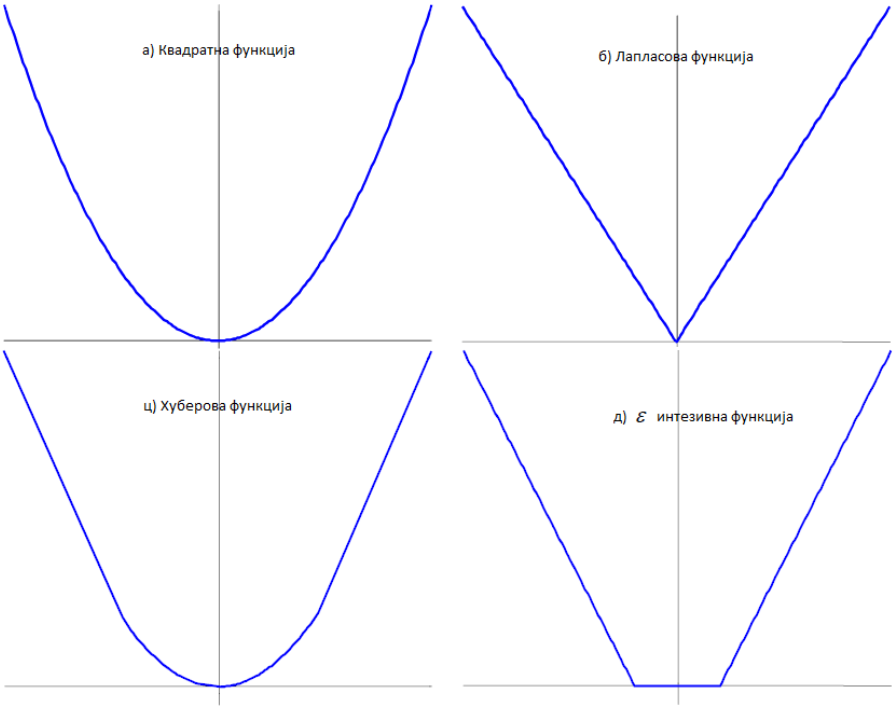
$$K(x, x^{'}) = e^{-\frac{\|x - x^{'}\|}{\sigma^2}}$$

Ширина σ је параметар који одређује сам корисник, али су број радијално заснованих функција и њихови центри аутоматски одређени на основу броја подржавајућих вектора и њихових вредности.

2.1.3 Регресија техником подржавајућих вектора

Методе подржавајућих вектора SVMs се такође могу применити и на регресионе проблеме (eng. *Support vector regression*, SVR) увођењем функције губитка (eng. *loss function*) [99]. Функција губитка мора да укључи меру растојања. Једноставности ради, и овде је прво описан линеаран проблем а затим проширен на нелинеаран случај увођењем кернел функција. Иако је наизглед тешко повезати проблем регресије са већ описаним проблемом, увођењем одговарајуће математичке нотације проблем регресије може да се повеже са претходно описаним поступком метода подржавајућих вектора. У проблему регресије су

[1] Овај кернел задовољава Мерсеров услов само за неке параметре [3]



Слика 2.7: *Примери функција губитка*

примери (инстанце) задате на следећи начин:

$$D = \{(x_1, y_1), ..., (x_n, y_n)\}, x \in R^N, y \in R$$

са линеарном функцијом

$$f(x) = \langle w, x \rangle + b$$

Оптимална регресиона функција се проналази минимизирањем израза:

$$\Phi(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i^- + \xi_i^+),$$

где је C као и у претходном случају параметар који корисник дефинише, а ξ_i^- и ξ_i^+ су допунске променљиве које представљају доњу и горњу границу грешке. Примери најчешће коришћених функција губитка су дати на слици 2.7

- Функција губитка на слици 2.7(а) одговара рачунању грешке методом најмањих квадрата (eng. *Least square error*).
- Функција губитка на слици 2.7(б) је Лапласова (eng. *Laplacian loss function*) која је мање осетљива на елементе ван граница(eng. *outliers*).

- Хуберова функција (слика 2.7(ц)) је најпогоднија када је непозната дистрибуција података.
- Као алтернативу овим трима функцијама Вапник је предложио нову функцију (слика 2.7(д)) која се може добити из ретког скупа подржавајућих вектора.

Узимањем ξ-интезивне функције губитка Лагранжов оптимизациони проблем се своди на:

$$L_\xi(y) = \begin{cases} 0, & \text{Ако } |f(x) - y| < \xi \\ |f(x) - y| - \xi, & \text{иначе} \end{cases}$$

А решење се може записати у облику:

$$\max_{\alpha\alpha^*} W(\alpha, \alpha^*) = \max_{\alpha\alpha^*} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle + \sum_{i=1}^{l}\alpha_i(y_i - \xi) - \alpha_i^*(y_i + \xi)$$

са ограничењима:

$$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, .., l$$

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0$$

Решавањем добијене једначине са ограничењима, се добијају Лагранжови мултипликатори. А из регресионе функције следи да је:

$$\overline{w} = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i$$

$$\overline{b} = -\frac{1}{2}\langle \overline{w}, (x_r + x_s)\rangle$$

Да би *Karush-Kuhn-Tucker* (ККТ) [57] услови били задовољени, мора да важи:

$$\overline{\alpha_i \alpha_i^*} = 0, i = 1, .., l.$$

Одатле следи да су подржавајући вектори тачке за које важи да је тачно један Лагранжов мултипликатор већи од нуле. За $\xi = 0$ се добија L_1 функција

губитка, где је оптимизациони проблем поједностављен:

$$min_{\beta}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\beta_i\beta_j\langle x_i, x_j\rangle - \sum_{i=1}^{l}\beta_i y_i$$

са следећим ограничењима:

$$-C \leq \beta_i \leq C, i = 1, .., l.$$

$$\sum_{i=1}^{l}\beta_i = 0$$

За задату регресиону функцију тада важи:

$$\overline{w} = \sum_{i=1}^{l}\beta_i x_i$$

$$\overline{b} = -\frac{1}{2}\langle \overline{w}, (x_r + x_s)\rangle$$

Проблем се аналогно решава избором неке друге функције губитка.

Нелинарна регресија

Проблем нелинеарне регресије се решава слично као проблем нелинеарне класификације, потребно је погодно представити податке. Подаци се пресликавају у простор веће димензије где се може применити поступак линеарне регресије. Увођењем кернел функција се елиминише проблем димензионалности. Нелинерно SVR решење проблема увођењем кернел функције, за ξ интезивну функцију губитка је дато једначином:

$$\max_{\alpha\alpha^*} W(\alpha, \alpha^*) = \sum_{i=1}^{l}\alpha_i^*(y_i - \xi) - \alpha_i(y_i + \xi) - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j)$$

са следећим ограничењима:

$$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, .., l$$

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0$$

Решавањем ове једначине са задатим ограничењима се добијају Лагранжови мултипликатори α_i, α_i^*, а регресиона функција је задата изразом:

$$f(x) = \sum_{SVs}(\alpha_i - \alpha_i^*)K(x_i, x) + \bar{b}$$

где је

$$\langle w, x\rangle = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x_i, x_j)$$

$$\bar{b} = -\frac{1}{2}\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(K(x_i, x_r) + K(x_i, x_s))$$

Оптимизациони услови се слично примењују и за друге функције губитка, изменом скаларног производа у кернел функцији. ξ - интезивна функција губитка је боља јер за разлику од Хуберове или квадратне функције губитке не захтева да све тачке скупа буду подржавајући вектори, већ број подржавајућих вектора може да се редукује и може бити редак скуп.

2.2 Кластеровање података

Постоје случајеви када имамо нпр. n примера (инстанци, слогова) без придружених класа. Тада може бити значајно груписање примера у значајне/смислене подгрупе. Ако се почетни скуп података дели у k група (кластера), свака од тих k група може даље да се третира као класа за себе. Број k може бити унапред познат, или га треба емпиријски одредити. Са уведеном мером *сличности* међу датим инстанцама, задатак груписања (кластеровања) је оптимизациони проблем максимизовања мере сличности унутар групе, односно минимизовање сличности инстанци које не припадају истој групи. Уколико су сви атрибути инстанци нумерички, онда се за мере сличности узима удаљеност инстанци. Основне методе кластеровања се могу поделити на следеће:

- **Засноване на растојању** (eng. *Distance based*). Сви атрибути су нумерички и инстанца је представљена као вектор чије су компоненте атрибути инстанце. За меру сличности се узима Еуклидско растојање или друга растојања која су заснована на особинама тачака а не њиховим положајем у простору:

  - *Jaccard* растојање,
  - Косинусно растојање и
  - *Edit* растојање.

  Ако постоје и категорички атрибути онда је мера сличности број заједничких атрибута.

- **Засноване на подгруписању** (eng. *Partition based*). Инстанце се групишу у k подгрупа. Свакој од могућих k подгрупа се додељује оцена. Како означавање свих могућих подгрупа није једноставан задатак примењују се различите хеуристике да би се убрзала претрага.

- **Хијерархијске** (eng. *Hierarchical*). Постоје два приступа хијерархијског груписања. Први полази од формирања једног кластера где се налазе све инстанце. Затим се почетни кластер дели на два која су доста удаљена у смислу мере сличности. Поступак се наставља све док свака инстанца не буде сама у кластеру. Други приступ је обрнут, полази се од тога да свака инстанца припада засебном кластеру. У следећем кораку се групишу по два кластера која су претходно утврђеном мером сличности најближа. Поступак се наставља све док све инстанце не буду придружене једном кластеру.

- **Засноване на моделима** (eng. *Model-based*). Сваки кластер се посматра као колекција података у којима важи мултиваријантна нормална дистрибуција. Затим се рачуна вероватноћа да сваки пример/инстанца припада том кластеру.

- **Засновано на густини** (engl. *Density-based*). Треба установити кластер произвољног облика према густини инстанци у изабраном региону, који се затим рекурзивно повећава додавањем суседних примера све док се не достигне одређена густина и док има суседних примера.

Мера сличности тј. близине у кластеровању, се дефинише као геометријска удаљеност. Пожељно је нормализовати или скалирати атрибуте у примерима, јер атрибути у различитим димензијама могу бити у различитим мерним јединицама те не могу да се упореде. Алтернатива је додељивање тежина

атрибутима према њиховој важности. У овој тези је коришћена метода груписања k-срединама [53] [63] , која је описана у следећим корацима:

(1) Прво се одреди (обично произвољно) k - средина (k - центара).

(2) Затим се понављају следећа два корака све док процес конвергира или док се не достигне унапред утврђен број итерација:

   (а) Пролази се кроз све примере и они се додељују кластеру чијем центру су најближи.

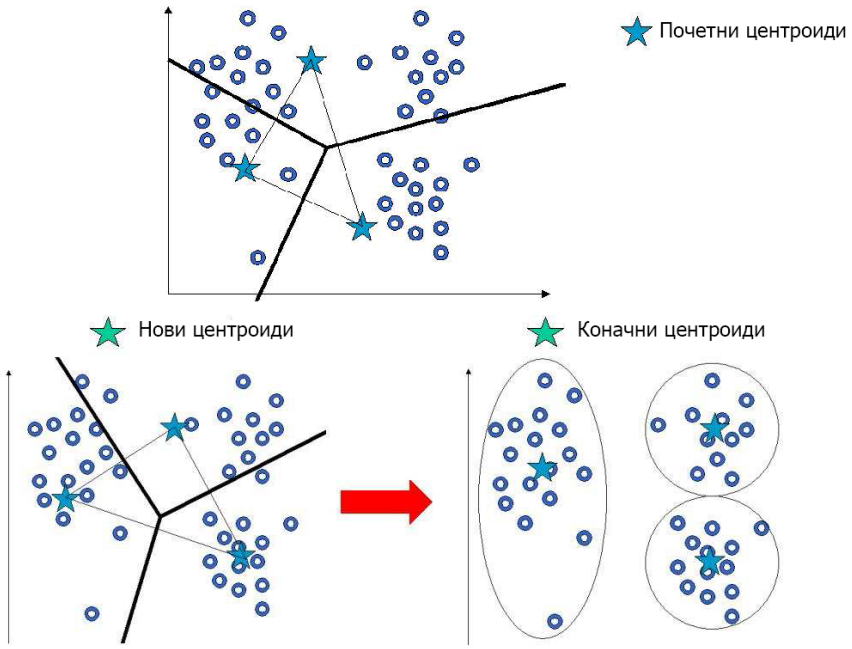
   (б) За сваки од k направљених кластера се поново прерачунају центри на основу инстанци које су му придружене.

У кораку два се за рачунање најближих примера најчешће користи Еуклидско растојање $\sqrt{\sum_{i=1}^{n}\left([d_1]_{f_i}-[d_2]_{f_i}\right)^2}$, између два n димензиона вектора атрибута d_1 и d_2 уколико су атрибути нумеричке вредности. У кораку $(2)\,(b)$ центар сваког од k кластера се поново рачуна и узима се средња вредност $\left\langle \left(\sum_{d\in C}[d]_{f_1}\right)/|C|,...,\left(\sum_{d\in C}[d]_{f_n}\right)/|C|\right\rangle$ израчуната за све тачке (примере) d који припадају кластеру C. Алгоритам k - средина је приказан на слици 2.8. На случајан начин су иницијално изабрана три центроида што је приказано на горњем делу слике. Такође су приказане границе између кластера (граничне линије су симетрале правих добијених повезивањем центара кластера). На слици у доњем левом углу су приказани нови израчунати центри. После још једне итерације рачунања растојања свих тачака од нових центроида добијени су нови кластери (слика доле десно).

Недостатак ове методе кластеровања је што није осмишљена за кластере који се преклапају. Други недостатак је што центри могу лако да се поремете због елемената ван граница. Набројани недостаци се превазилазе алгоритмима за *максимизацију очекивања* (eng. *Expectation Maximization* (EM))[14].

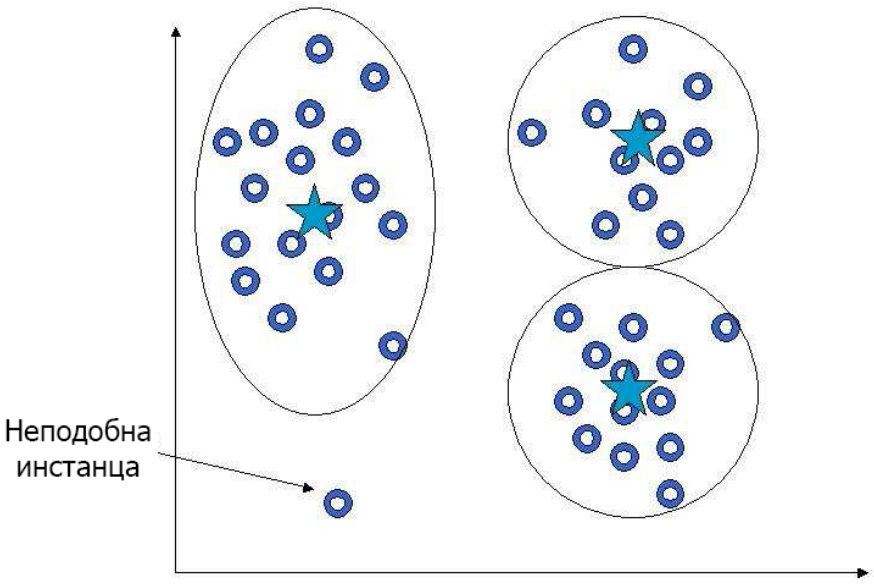
2.2.1 Откривање одступања

Откривање одступања (eng. *deviation detection*) је поступак проналажења инстанци које одступају од осталих инстанци у скупу података. Овај поступак је супротан ономе што по дефиницији подразумева кластеровање. На слици 2.9 је илустрован овај концепт и јасно је да инстанца издвојена испод великог



Слика 2.8: *Илустрација алгоритма кластеровања к-срединама*

левог кластера не припада ни једном кластеру. Избацивањем те инстанце (када се инстанца третира као шум у подацима) се формирају компактни кластери. Међутим постаје и случајеви када су баш инстанце које одступају од интереса. Многобројне статистичке [33] и технике истраживања података [2], [66],[65],[56], [92] се баве овим проблемом.



Слика 2.9: *Илустрација проналажења неподобних инстанци*

2.3 Правила придруживања

Техника правила придруживања проналази интересантна правила и/или корелације односа између различитих ставки огромних скупова података. Ова истраживачка техника је широко примењена у многим сферама пословне праксе - од анализе потрошачких навика, преко управљања људским ресурсима, до развоја језика. Омогућава откривање скривених образаца у великим скуповима података. Типичан и широко распрострањен пример коришћења правила придруживања је анализа потрошаке корпе. Сви производи које купац наручи или купи током одређене активности представљају један запис (слог), односно чине једну трансакцију (eng. *itemset*). Сваки елемент трансакције има одређену вредност атрибута. У правилима придруживања се вредност атрибута назива ставка (eng. *item*). У процесу проналажења правила придруживања постоје две фазе: проналажење честих скупова и генерисање правила придруживања на основу добијених резултата. Основне мере које се користе за издвајање правила су подршка (eng. *support*) и поверење (eng. *confidence*). Подршка осликава учесталост са којом се скуп одређених елемената трансакције појављују у скупу података. Рачуна се као проценат трансакција (слогова) који садрже дати скуп (као подскуп) од укупног броја трансакција. Ако скуп ставки има подршку већу од прецизираног прага (eng. *minsup*), каже се да је он подржан (eng. *supported*) или чест (eng. *frequent*). Поверење осликава импликативност (узрочност, повезаност) које је присутно у правилу, односно представља условну вероватноћу да су ставке на десној страни правила присутне ако су присутне ставке на левој страни правила.

Увођење математичке нотација

Каже се да трансакција T садржи ставку x ако је $x \in T$. Такође, скуп ставки X се јавља у трансакцији T ако важи да је $X \subseteq T$. Нека је дат скуп свих трансакција D и скуп свих ставки X. Кардиналност скупа трансакција се означава са $|D|$. Нека је број ставки X у D се означава са $count^D(X)$ и представља број трансакција које садрже ставке X. Ниво подршке скупа ставки X у D се означава са $support^D(X)$ и представља проценат трансакција у D које

садрже X тј.:

$$support^D(X) = \frac{T \in D | X \subseteq T}{|D|}$$

Правила придруживања су парови које представљамо $X \Rightarrow Y$, где су X и Y два скупа ставки за које важи $X \cap Y = \emptyset$. Скуп ставки X се назива још и претходник правила. Скуп ставки Y се назива последица правила. Мере подршка и поверење се представљају следећим дефиницијама:

Дефиниција 2.3 *Ниво подршке правила $X \Rightarrow Y$ у скупу D се дефинише као проценат трансакција у D које садрже $X \cup Y$:*

$$support^D(X \Rightarrow Y) = support^D(X \cup Y)$$

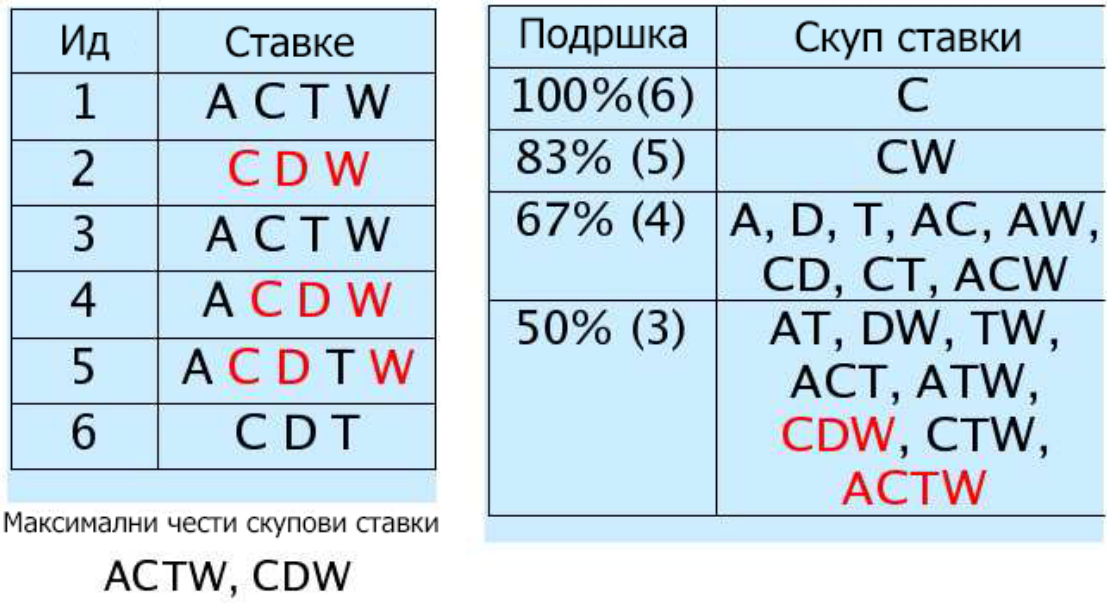
Дефиниција 2.4 *Ниво поверења правила $X \Rightarrow Y$ у скупу D се дефинише као проценат трансакција у D које садрже X и такође садрже Y у односу на све трансакције које саджже X:*

$$confidence^D(X \Rightarrow Y) = \frac{support^D(X \cup Y)}{support^D(X)} = \frac{count^D(X \cup Y)}{count^D(X)}$$

Правила придруживања се дефинишу: за дати скуп података (слогова) и дефинисаних прагова *minsupp* и *minconf*, треба пронаћи сва правила $X \Rightarrow Y$ тако да важи $support^D(X \cup Y) \geq minsupp$ и $confidence^D(X \Rightarrow Y) \geq minconf$. Правило $X \Rightarrow Y$ се интерпретира на следећи начин: ако трансакција садржи скуп ставки X онда највероватније садржи и скуп ставки Y. Границе за миниминалну подршку и минимални ниво поверења одређује корисник према томе која правила су од интереса. За дати границу *minsupp*, за скуп ставки X се каже да је чест (eng. *frequent itemset*) у скупу D ако важи $support^D(X) \geq minsupp$. Такође, за чест скуп ставки X се каже да је максималан у скупу података D ако ниједан надскуп у D није чест. Јасно је да су сви подскупови честог скупа такође чести. За чест скуп ставки X се каже да је *затворен* ако ниједан од његових надскупа нема исту подршку. Слика 2.10 приказује пример издвајања честих скупова. У примеру је приказана база података која садржи 5 трансакција. У табели са десне стране су приказане ставке које су честе у односу на задату границу подршке. За границу $minsupp = 50\%$ све приказане ставке су честе, где су ставке ACTW и CDW су нејвећи чести скупови. Да би се пронашла сва значајна правила придруживања у неком скупу података која

задовољавају границе поверења и подршке *minconf* и *minsupp* примењују се следећи кораци:

- Генеришу се сви чести скупови ставки.
- Генеришу се правила из честсих скупова и елиминишу се она правила која не задовољавају услов за минимално поверење.



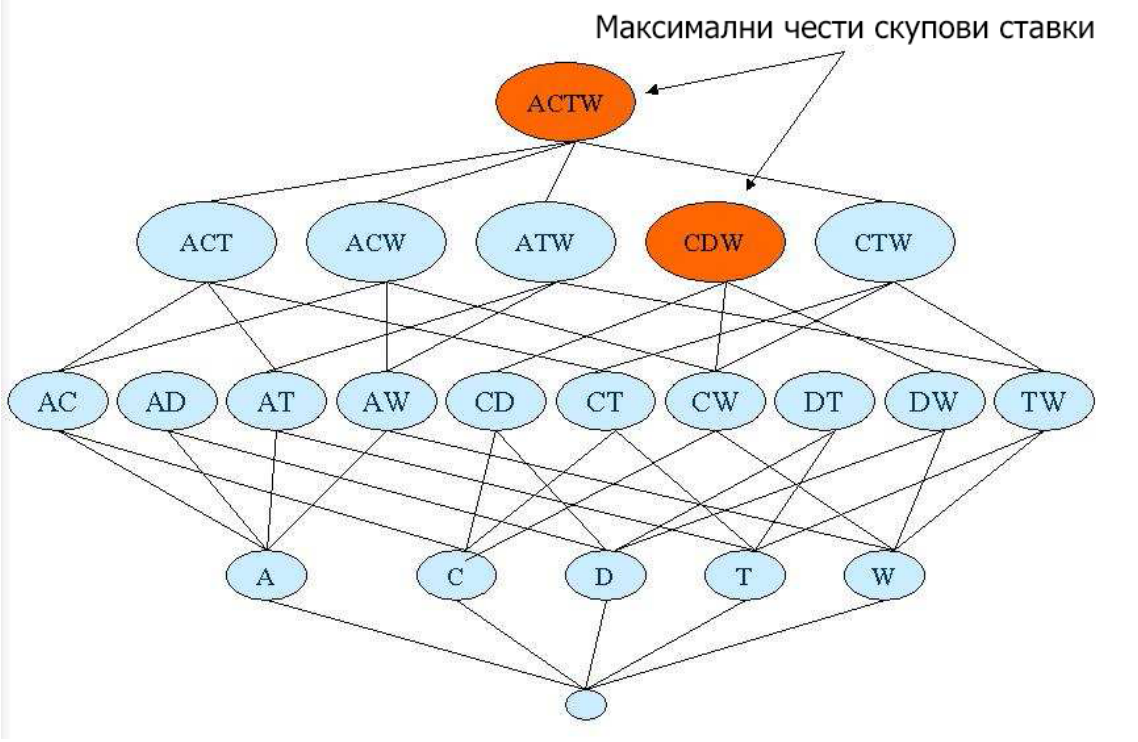
Слика 2.10: *Пример честог скупа ставки, где је граница за минималну подршку* $minsupp = 50\%$

Први корак захтева експоненцијалну претрагу у односу на величину скупа ставки, и са рачунарске тачке гледишта је веома скупа операција. Са тим у вези, кључни корак у издвајању правила придруживања је решавање првог корака. У истраживању података је овај проблем интезивно проучаван и развијен је велики број алгоритама за решавање овог проблема. Први ефикасан приступа је Априори (eng. *Apriori*) алгоритам [4], који је био инспирације за многе друге ефикасне приступе [15] [31] [116]. Други приступ је био оријентисан ка проналажењу метода за издвајање максималног [27] и затвореног скупа ставки [117]. Ефикасност Априори алгоритма се заснива на знању да ако један скуп ставки није чест онда ни његови подскупови нису чести. Чести скупови се праве у итеративном поступку. Нека су са L_k означени чести скупови добијени у k - тој итерацији. Онда је за добијање честих скупова дужине $k+1$ довољно само

спајање добијених скупова L_k у итерацији k, уместо тестирања свих могућих кандидата дужине $k+1$. Спајање подскупова се дефинише на следећи начин:

$$\{i | i_1 \in L_k, i_2 \in L_k, i \subseteq (i_1 \cup i_2, |i| = k+1, (\nexists i' \subset i, |i'| = k) \wedge (i' \notin L_k)\}$$

Ниво подршке се рачуна за сваки од ових скупова и проверава се да ли је скуп чест или није. Други корак у издвајању правила придруживања је јефтинији у смислу рачунарских операција. За познати чест скуп ставки, правила придруживања се приказују као на слици 2.11.



Слика 2.11: *Пример представљања честог скупа ставки на основу максималних честих скупова приказаних на претходној слици.*

Сваки чвор на слици је јединствен чест скуп ставки, чији ниво подршке је већи од унапред дефинисане границе. Ивице повезују чворове које су директно у вези подскуп-надскуп. После овог корака, за сваки чвор X добијен из скупа $X \cup Y$ се генирушу кандидати правила $X \Rightarrow Y$, и проверава се њихов ниво поверења. У примерима приказаним на сликама 2.10 и 2.11 за максималан скуп ставки $\{CDW\}$ важи:

- $count^D(CDW) = 3$,
- $count^D(CD) = 4$,

- $count^D(CW) = 5$,
- $count^D(DW) = 3$,
- $count^D(C) = 6$,
- $count^D(D) = 4$ и
- $count^D(W) = 5$.

За сваки од генерисаних скупова се могу генерисати правила и израчунати њихов ниво поверења:

- $confidence^D(CD \Rightarrow W) = 3/4 = 75\%$,
- $confidencet^D(CW \Rightarrow D) = 3/5 = 65\%$,
- $confidence^D(DW \Rightarrow C) = 3/3 = 100\%$,
- $confidence^D(C \Rightarrow DW) = 3/6 = 50\%$,
- $confidence^D(D \Rightarrow CW) = 3/4 = 75\%$ и
- $confidence^D(W \Rightarrow CD) = 3/5 = 60\%$.

Из ових правила се сада лако могу пробрати она која задовољавају минимални дефинисани ниво поверења.

	$Y \subset T$	$Y \not\subset T$
$X \subset T$	TP	FP
$X \not\subset T$	FN	TN

Табела 2.1: *Табела контигената*

Мере нивоа подршке и поверења служе за дефинисање правила која су од интереса. Ипак ове мере нису довољне да се издвоје правила која су од интереса увек. У различитим применама користе се додатни услови да се одреде да ли је правило заиста значајно. Други приступи се користе за раздвајање правила која су већ укључена у нека од постојећих правила [55], која ограничавају издвајање правила на само она која задовољавају већ установљени шаблон. Већ је објашњено да се правило $X \Rightarrow Y$ интерпретира као "Ако се X јавља

у скупу података T онда се и Y највероватније јавља у скупу података T". Ако се правило посматра као предикција онда се дефинише табела контигената 2.1. У складу са табелом контигената, уводе се мере које издвајају правила и представљају ниво неочекиваности правила: *Lift*, *Gini*, J - мера, и друге. Мера *Lift* се најчешће користи за издвајање правила а узима у обзир статистичку зависност скупа ставки на левој и десној страни правила. Израчува се као:

$$Lift = \frac{P(Y|X)}{P(Y)}$$

Јасно је да за вредности 1 ($P(X)P(Y) = P(X,Y)$) скуп ставки X и Y су независне. Вредности ове мере веће од 1 указују на то да се лева и десна страна правила јављају чешће него што је очекивано.

Поглавље 3

Предвиђање Т - ћелијских епитопа

Два су основна биоинформатичка приступа за идентификовање Т - ћелијских епитопа: директно и индиректно предвиђање [26]. Директни приступи су углавном били засновани на препознавању кратких линеарних мотива у протеинима које препознају Т - ћелије. Индиректан приступ је фокусиран на препознавању пептида који се везују (eng. *binders* - везујући пептиди) за молекуле главног хистокомпатибилног комплекса, који представља кључан догађај у селекцији Т - ћелијских епитопа. Методе за директно предвиђање епитопа су се показала као лошије, са недовољно добром тачношћу предвиђања [26] и тренутно је доступан само један предиктор који је заснован на директном приступу: CTLpred[1]. Насупрот томе, постоји велики број расположивих метода које индиректно предвиђају Т - ћелијске епитопе. Предмет ове тезе су индиректне методе предвиђања *MHC* везујућих пептида. У наставку су детаљно описане постојеће методе и њихова методологија, као и предложени нови приступ. Све методе користе знања, обрасце, правила и везе установљене анализом експериментално утврђених епитопа. Данас постоји велики број јавно доступних база, које се разликују у начину на који су подаци прикупљени, количини, квалитету и поузданости података.

[1] http://www.imtech.res.in/raghava/ctlpred/

3.1 Базе експериментално утврђених Т - ћелијских епитопа и *MHC* везујућих пептида

У оквиру тезе су разматране и укратко описане само базе података које се најчешће помињу као основни извор података за прављење модела за предвиђање Т - ћелијских епитопа. Неке од њих су:

- **IEDB** (eng. *Immune Epitope DataBase*), је највећа и најпоузданија слободно доступна база Т - ћелијских епитопа и *MHC* везујућих лиганада. База садржи податке и о Б - ћелијским епитопима, а поседује и велики број алата за предвиђање специфичних региона протеина које су у вези са антигенским регионима. База има укупно 455 000 слогова о експериментално утврђеним *MHC* везујућим и не-везујућим пептидима. Осим експерименталних података у бази се налазе и подаци добијени предвиђањем помоћу IEDB алата[2]. База је настала у оквиру истраживања Националног института за алергију и инфективне болести[3], а налази се на локацији http://www.iedb.org/,

- **AntiJen** база садржи квантитативне податке о везујућим пептидима и лигандима, *TCR-MHC* комплексима, Т - ћелијским епитопима, *TAP*, Б - ћелијским епитопима и молекуларним протеин-протеин интеракцијама. Подаци у бази су сакупљени из објављених експерименталних студија. Цела база има 24 000 слогова. Ниједан слог у бази није добијен предвиђањем. Налази се на локацији http://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm, а настала је на *Edward Jenner* институту за истраживање вакцина[4].

- **MHCBN** (eng. *MHC Binding and Non-binding peptides*) база садржи детаљне информације о *MHC* везујућим пептидима и Т - ћелијским епитопима. Тренутна верзија base је 4.0. Број расположивих експерименталних података је знатно мањи него у IEDB бази, и тренутно садржи око 25 860 везујућих и невезујућих пептида. Налази се на локацији http://www.imtech.res.in/raghava/mhcbn/. Сматра се мање поузданом

[2] http://tools.iedb.org/main/
[3] http://www.niaid.nih.gov/
[4] http://www.jenner.ac.uk/home

базом него претходне две. Подаци су сакупљани из објављених радова, и не прецизира се тачно да ли су подаци везани само за експерименталне студије.

- **SYFPEITHI** је база T - ћелијских епитопа и *MHC* везујућих пептида која је заснована само на подацима сакупљеним из разних објављених радова. Налази се на локацији http://www.syfpeithi.de/. База садржи и све познате мотиве, а на истој локацији се налази и истоимени предиктор за предвиђање T - ћелијских епитопа. Предиктор узима у обзир за предвиђање све постојеће мотиве из ове базе, али и све сакупљене информације о амино киселинама које могу бити сидра (детаљније објашњење је дато у поглављу 3.2.1).

- **HIV Molecular Immunology Database**. Подаци који су укључени у ову базу су везани искључиво за HIV протеине. Прикупљани су из литература везаних за HIV имунологију. Специфични одговори Б - и T - ћелија су сумирани и означени. Сакупљени подаци су подељени у 3 секције: *CTL*, T помажуће ћелије и антитела. У оквиру сваке секције су подаци означени тако да је дат и епитоп и протеин коме припада епитоп заједно са позицијом на којој се налази. База се налази на локацији http://www.hiv.lanl.gov/content/immunology/index.

- **MHCPEP** је база са близу 13 000 пептида за које је познато да се везују за *MHC* молекуле. Подаци су сакупљани из разних објављених радова али и из експерименталних студија. База припада *Walter and Eliza Hall* Институту[5] који се бави истраживањем различитих врста канцера. Сваки слог у бази садржи пептид, *MHC* молекул за који се везује (уколико овај податак постоји), опис/назив експерименталне методе којом је добијен, ознаку изворног протеина и референцу на рад из кога су подаци преузети. Сви подаци у бази су везани само за канцер протеине.

- **FIMM** база више није доступна, а подаци из ове базе су укључени у IEDB базу. Садржај базе и начин прикупљана података је детаљно описан у [97].

- **Ligand** је база *MHC* везујућих пептида сакупљених из неколико

[5] http://www.wehi.edu.au

објављених радова [96]. База је доступна на локацији http://hlaligand.ouhsc.edu/.

3.2 Методологија рада постојећих метода за предвиђања Т - ћелијских епитопа

Методе за индиректно предвиђање Т - ћелијских епитопа се могу даље поделити у две основне категорије. Прву категорију чине методе засноване на издвајању образаца из секвенци пептида који се везују за молекуле *MHC* класа, док се методе друге категорије заснивају на 3Д структури протеина која моделује везивање пептида и *MHC* молекула. Прва група метода и алата укључује процедуре које су засноване на мотивима, квантитативним матрицама и техникама истраживања података: стаблима одлучивања (eng. *decision trees*, DT), вештачким неуронским мрежама (eng. *Artificial neural networks*, ANNs), скривеним Марковљевим моделима (eng. *Hidden Markov models*, HMMs), и методама подржавајућих вектора (eng. *Support vector machines*, SVMs). Друга категорија је изван опсега истраживања ове тезе.

3.2.1 Методе засноване на мотивима

За настанак ових метода од значаја су била прва знања везана за пептиде који се везују за молекуле *MHC* класа. Установљено је да су пептиди који се везују за конкретан *MHC* молекул функционално сродни и деле амино киселине са сличним особинама на различитим позицијама у примарној протеинској секвенци. Ове амино киселине су назване сидра (eng. *anchor*). Познато је да одређене амино киселине из пептида формирају бочне ланце са комплементарним амино киселинама из специфичних *MHC* молекула. Име сидро су и добиле јер се јако везују за ланце *MHC* молекула и највише доприносе самом везивању. Само на основу овог знања је направљено пуно алата за предвиђање *MHC* везујућих пептида [104]. SYFPEITHI је једна од најпознатијих база података са сакупљеним експериментално утврђеним везујућим мотивима. Ова база је последњи пут ажурирана 2006. године. У међувремену је утврђено да и амино киселине које припадају не - сидро позицијама значајно доприносе самом везивању пептида, тако да су се ови

модели, показали неподобним за предвиђање везујућих пептида. Ипак ови модели као такви нису били скроз одбачени, и иако су застарели, они су највише коришћени за предвиђање епитопа [81]. Методологија рада ових модела се заснива на претраживању амино киселинске секвенце пептида где се траже мотиви који се поклапају са мотивима из библиотеке мотива [93]. *MHC* везујући мотиви за анализирани пептид се могу добити поређењем са познатим везујућим или не-везујућим пептидима [5]. Познати примери употребе метода заснованих на мотивима су:

- Предвиђање епитопа који се везују за HLA-DR алеле протеина *Plasmodium falciparum* [21],
- EPIPREDICT је алат заснован на мотивима, а направљен је у сврху предвиђање епитопа који се везују за молекуле *MHC* класе II у људским протеинима који учествују у нетолеранцији на глутен [50],
- MOTIF [20] алат укључује све прикупљене мотиве за алел HLA-A*0201, и може да предвиђа само епитопе који се везују за тај алел,
- EPIMER је такође алат чији се модел заснива на мотивима, направљен је на *Brown*[6] Универзитету и коришћен за предвиђање епитопа повезаних са HIV-ом [68] [28],
- SYFPEITHI [93] је један од најкоришћенијих алата за предвиђање из ове групе. Заснован је на свим прикупљеним мотивима из истоимене базе. Методологија рада овог предиктора је следећа: пептид се анализира и пореди са познатим мотивима, ако се у пептиду пронађу мотиви који се јављају у епитопима пептид се сматра епитопом. Оцена афинитета везивања се додељује према следећим правилима: свакој амино киселини се додељује мера на основу тога да ли она припада сидро позицији, помоћној сидро позицији, или је нека од преферираних амино киселина. Идеална сидро амино киселина добија 10 поена, неуобичајена сидро амино киселина 6 до 8 поена, помоћна сидро амино киселина 4 до 6 поена и префериране амино киселине од 1 до 4 поена. Амино киселине које се смтрају непожељним добијају од -1 до -3 поена.

[6] https://www.brown.edu/

Тачност модела заснованих на мотивима је између 60 и 70%. Разлог томе је што се сви везујући пептиди не могу повезати са идентификованим мотивима. У већини случајева је веза између предвиђених и експериментално утврђених афинитета везивања веома слаба. Спроведено је више истраживања која су показала неподобност ових алата. Једно од таквих истраживања су спровели и *Andersen* и његови сарадници [7], где су анализирали афинитете везивања епитопа добијених предвиђањима, уз помоћ SYFPEITHI и BIMAS алата (описан је у следећој групи модела), са експериментално утврђеним епитопима онкогених и вирусних протеина. Аутори су показали да наведени алати предвиђају велики број лажно позитивних епитопа, док неке праве епитопе уопште не предвиђају као епитопе. Недовољно добра тачност постојећих алата и метода је условила настанак нових унапређених метода за предвиђање епитопа.

3.2.2 Методе засноване на матрицама повезаности

Модели засновани на матрицама повезаности (квантитативним матрицама) су настали као унапређење већ постојећих модела заснованих на мотивима. Коришћењем статистичких метода, свакој амино киселини се додељује вредност-мера која представља везу између позиције на којој се налази у пептиду и улога у повезивању. На овај начин се прави матрица коефицијента димензије $l \times 20$, где l предтсваља дужину пептида који се анализира а 20 је величина алфабета $\alpha = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$ (број основних амино киселина). За предвиђање се даље подразумева да свака амино киселина на свакој позицији у пептиду даје известан допринос (нумеричка вредност у матрици) у коначном рачунању енергије везивања (eng. *binding energy*). Коначан резултат се рачуна на неки од следећих начина: сумирањем елемената матрице, множењем елемената матрице, узимањем средње вредности елемената, итд. Добијена вредност се пореди са већ унапред дефинисаном границом (eng. *treshold*). Граница одлучивања се утврђује над експерименталним подацима. У општем случају коефицијенти матрице се добијају рачунањем учесталости амино киселина на одговарајућим позицијама код познатих пептида – везивача или квантитативних података (афинитету везивања) о *MHC* везујућим пептидима. Првим приступом се добија вероватноћа везивања пептид за молекуле *MHC* класа, у другом

приступу где се у обзир узимају квантитативне мере се предвиђа афинитет везивања пептида за молекуле *MHC* класа.

Неки од познатих модела заснованих на матрицама везивања су:

- EpiMatrix[7] је комерцијалан предиктор и није слободно доступан.
- SYFPEITHI - основна верзија предиктора је заснована на мотивима, која је касније унапређена у предиктор заснован на матрицама[8],
- BIMAS[9] предиктор је такође у основи заснован на мотивима јер се квантитативна матрица формира од постојећих познатих мотива. Аутори овог алата су искористили већ постојеће мере за сваку амино киселину [80]. Мотиви су сакупљани из различитих извора, база и до тада објављених радова.
- RANKPEP[10] предиктор прво користи методу засновану на мотивима како би анализирани пептид био поравнат са постојећим мотивима, али узима у обзир и матрицу коефицијената која је добијена рачунањем учесталости свих амино киселина на свакој позицији у пептидима за које је познато да се везују за молекуле *MHC* класа [94].
- SMM[11] (eng. *Stabilised Matrix Method*) предиктор се разликује од претходно описаних у начину рачунања коефицијената квантитативне матрице. Коефицијенти се добију применом метода машинског учења где је циљ да се пронађу коефицијенти који најбоље објашњавају посматране примере за учење. Коефицијенти се добијају када се минимизује разлика између предвиђеног афинитета и измерене (стварне) вредности афинитета. SMM предиктор се показао као најбољи предиктор у овој групи у истраживању спроведеном у [60] и [12].
- ARB[12] (eng. *Average relative binding*) је предиктор заснован на мултипликативној квантитативној матрици. Формирање матрице је засновано на претпоставци да свака амино киселина даје свој допринос у

[7] http://www.epivax.com/epimatrix/
[8] http://www.syfpeithi.de
[9] http://www.bimas.cit.nih.gov/molbio/hla_bind
[10] http://bio.dfci.harvard.edu/RANKPEP
[11] http://tools.immuneepitope.org/analyze/html/mhc_binding.html
[12] http://tools.immuneepitope.org/analyze/html/mhc_binding.html

укупном афинитету везивања. Нпр. амино киселина R на позицији i има допринос R_i. Израчунат је допринос сваке од 20 амино киселина на свакој позицији, на основу експерименталних података о афинитету везивања. Доприноси сваке амино киселине су представљени матрицом. Предвиђање се врши тако што се за посматрани пептид из матрице издвоје одговарајући доприноси и измноже. Добијена вредност представља афинитет везивања.

Набројане методе се разликују у начину на који рачунају коефицијенте матрице [62]. У основи поступак је сличан, модели су тренирани статистичким методама које анализирају колико често се нека амино киселина појављује на одређеној позицији у пептиду који се везује насупрот учесталости појављивања на тој позицији у пептиду који се не везује за молекуле *MHC* класа.

За разликовање слабих и јаких епитопа и за откривање колинеарности амино киселина у посматраном пептиду су погоднији модели засновани на PSSM (eng. *Position Specific Scoring Matrix*) матрицама, где се из скупа "поравнатих" пептида предвиђа везивање за низ молекула *MHC* класа I и II. Равнање потенцијалног везујућег пептида, са већ познатим везујућим мотивима се врши према сличности саме секвенце као и према структурној сличности. Од наведених предиктора SMM спада у PSSM методе. SMM је настао 2003. године као предиктор за HLA*A2 везујуће пептиде. У основи овог приступа је већ описан поступак заснован на квантитативним матрицама, који се комбинује са коефицијентима придруженим амино киселинама које се појављују у паровима (суседним) у везујућим пептидима. Ипак, SMM предиктор се показао као боље решење само у малом броју случајева [12]. Осим SMM предиктора, за идентификовање слабих мотива су предложени разни нови приступи углавном везани за један специфичан алел. Метода названа *Gibbs sampler* је прилагођена за тражење образаца код слабих мотива за алел HLA-DR4(B1*0401)[76]. *Rajapakse* је у свом истраживању [91] предложио нови приступ, заснован на мулти-објектном еволутивнном алгоритму за проналажење консензусних слабих мотива за један алел. *Guan* и *Doytchinova* су укључили и рачунање мултиваријантне статистике како би побољшали њихове моделе засноване на матрицама [30].

Велики недостатак метода за предвиђање T - ћелијских епитопа заснованих

на матрицама је што не узимају у обзир ефекат корелације амино киселина у пептиду тј. када допринос једне амино киселине у пептиду зависи од друге амино киселине на другој позицији. За укључивање узајамне корелације амино киселина у пептиду су погоднije друге методе засноване на техникама истраживања података. Методе засноване на техникама истраживања података могу бити и квантитативне и квалитативне. У првом случају метода као резултат даје квантитативну оцену афинитета везивања. У другом случају метода даје квалитативну оцену, разматрани пептид јесте везујући (потенцијални епитоп) или није везујући (неепитоп). Још важнији недостаци набројаних метода су следећи:

1. већина ових метода је направљена за предвиђање епитопа који се везују за један или два алела, што их чини непогодним за предвиђање промискуитетних епитопа, и

2. методе су развијене над подацима из ограниченог и малог скупа везујућих (не-везујућих) пептида што их чини мање поузданим.

3.2.3 Методе засноване на техникама машинског учења

У циљу превазилажења свих набројаних ограничења развијен је велики број метода заснованих на техникама истраживања података. У табелама 3.1 и 3.2 је дат преглед већине постојећих предиктора заснованих на техникама истраживања података. У табели 3.1 су приказане неке од најзначајнијих метода за квалитативно предвиђање. Излаз из таквих модела је: 1 - јесте или 0 - није епитоп, док је у табели 3.2 дат преглед најзначајнијих метода које предвиђају афинитет везивања. Иако су експериментални подаци о афинитету везивања епитопа готово увек задати у IC_{50} мери, не постоји јединствена мера за представљање излаза из различитих квантитативних методе.

Квалитативне методе

- **ANNPred** - Модел који је у основи овог предиктора је заснован на вештачким неуронским мрежама. Како неуронске мреже захтевају за тренирање велики број података то је овај предиктор направљен за 30 алела за које је постојало бар 40 експериментално утврђених епитопа. Сви

Назив	Метода	Шема енкодирања	Карактеристике
ANNPred[13]	ANN	Sparse encoding	Accuracy: 87.3% ± 5.9%
nHLAPred[14]	ANN/PSSM	Sparse encoding	Accuracy: 93.6% ± 2.92%
Zhu et al.[119],	Decision tree	N/A	Accuracy: 0.8
KISS[15]	SVM	Heckerman et al.[34]	AUC: 0.86 0.90
POPI[16]	SVM	Physicochemical properties	Accuracy: 60%
SVMHC[17]	SVM	Sparse encoding	MCC: 0.85

Табела 3.1: *Преглед квалитативних метода за предвиђање T - ћелијских епитопа MHC класе I*

подаци за тренирање модела су узети из MHCBN3.1 базе експериментално утврђених везујућих (eng. *MHC binding ligands*) и не-везујућих пептида. У случајевима где није постојао довољан број не-везујућих пептида, негативан скуп пептида је допуњен случајним бирањем нонамера из SWISS-PROT[18] базе. SWISS-PROT база не садржи експерименталне податке о не-везујућим пептидима, већ је овде искоришћена претпоставка да су случајно изабрани пептиди не-везујући. У прављењу модела су учествовали само пептиди величине 9 (нонамери). Пептидна секвенца је енкодирана ретким енкодирањем и добијени вектор се директно прослеђује као улаз у ANN.

- **nHLAPred** предиктор представља комбинацију ANNPred предиктора и методе засноване на квантитативним матрицама. За 30 алела за које је направљена подршка у ANNPred-у се користи ANNPred, а за још 37 алела се користе квантитативне матрице. За 17 алела од тих 37 су направљене нове квантитативне матрице а за преосталих 20 се користе оне из BIMAS алата (описан у претходном подпоглављу). Квантитативне матрице су направљене на следећи начин: свакој амино киселини у пептиду на одговарајућој позиција се придружује тежина, која се добија када се израчуна учесталост те амино киселине на одговарајућој позицији у скупу свих везујућих пептида, а затим подели са учесталошћу добијеном за исту амино киселину и позицију у скупу не-везујућих пептида. Квантитативне матрице су збирне матрице, где се вредност која се придружује пептиду рачуна као сума тежина његових амино киселина. На пример за пептид

[18] http://web.expasy.org/docs/swiss-prot_guideline.html

$p = ILKEPVHGV$ вредност која се придружује пептиду је

$$Score = I(1) + L(2) + K(3) + E(4) + P(5) + V(6) + H(7) + G(8) + V(9)$$

Пептиди којима је на овај начин придружена вредност већа него унапред дефинисана граница се сматрају везујућим (епитопима), док они са мањом вредношћу од утврђене границе се сматрају не-везујућим (неепитопима). Матрице које користе алати BIMAS и ProPred1 су мултипликативне, тј. вредност која се придружује пептиду се рачуна на следећи начин:

$$Score = I(1) * L(2) * K(3) * E(4) * P(5) * V(6) * H(7) * G(8) * V(9)$$

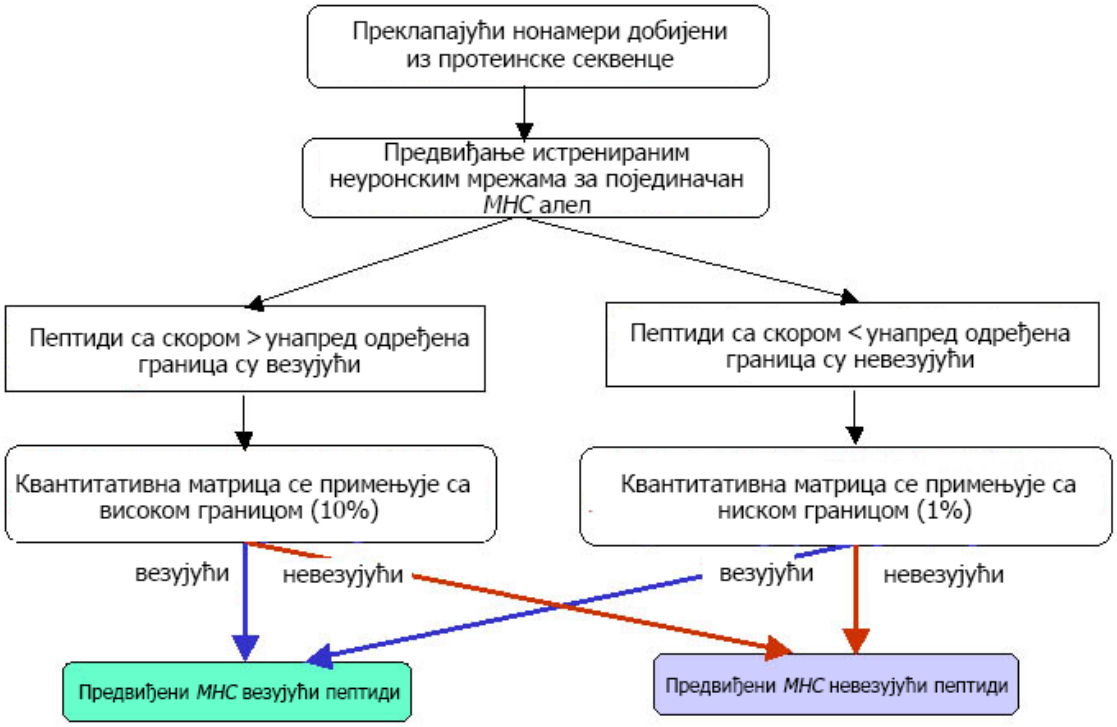
Граница за раздвајање везујућих од не-везујућих пептида се добија у два корака:

(1) Из свих протеина из SWISS-PROT базе су издвојени преклапајући нонамери.

(1) На основу направљене матрице свим генерисаним нонамерима се придружују одговарајуће вредности. На основу придружене оцене се сортирају сви пептиди у опадајући поредак и првих 1% се сматрају везујућим те се та мера узима као граница (негде се узима 2% или више[19]). На слици 3.1 је приказан принцип рада овог предиктора.

- *Zhu* и сарадници[119] су направили моделе засноване на стаблима одлучивања за 16 различитих алела *MHC* класе I. Податке о везујућим пептидима су узели из шест различитих база: MHCPEP3.1, SYFPEITHI3.1, FIMM3.1, MHCBN3.1 [10], AntiJen3.1 и Ligand 3.1 верзије из марта 2003. године, као и из приватног извора (*A. Sette*, необјављени резултати; *K. Udaka*, необјављени резултати) [119]. Модели су направљени само за бинарно поређење (као разлог је наведена немогућност поређења са различитим алатима и искористивост мера афинитета које су за сваку базу различите или чак недоступне). Већина алела за које су

[19]Напомена: утврђивање границе није стриктно, и базира се на теоријској основи и тврђењу да међу 100 случајно изабраних пептида највише 1% до 2% ће се показати као епитопи [16]. Ово тврђење се често користи за произвољан избор не-епитопа, у недостатку експерименталних података.



Слика 3.1: *Методологија рада nHLAPred предиктора.*

сакупљени експериментални подаци нема више од 95 везујућих пептида (од 16 анализираних алела таквих је 12). У недостатку експериментално утврђених не-везујућих пептида су искористили тврђење: "Највише 1% од свих могућих нонамера ће се везати за одговарајући алел" [106]. Из тог разлога су узели случајно генерисане пептиде из KEGG базе[20] [51] водећи рачуна само о томе да се разликују од познатих везујућих пептида. Овај начин бирања не-везујућих пептида није неуобичајен и коришћен је и у другим истраживањима [23]. Шема енкодирања која је коришћена за репрезентацију пептида и припрему улаза у модел стабла одлучивања за овај модел није позната. Оно што је интересантно у овом приступу је да су подаци добијени из база за један алел коришћени у тренирању модела за други алел. Тиме је повећана тачност модела добијеног првобитно тренирањем само над подацима за одговарајући алел. Мера ентропије је коришћена за одбацивање редундантних података из скупа свих везујућих пептида за тренирање. Ентропија се рачуна на следећи начин: ако је скуп свих везујућих пептида S кардиналности N, прави се матрица C димензије

[20] http://www.genome.jp/kegg/kegg1.html

20×9 која садржи број појављивања сваке од раличитих амино киселина на одговарајућој позицији. Са n_{ij} је у матрици означен број појављивања амино киселине i на позицији j у свим пептидима из скупа S. Коначно се ентропија рачуна према следећем изразу:

$$Entropy(S) = -\sum_{i=1}^{20}\sum_{j=1}^{9}(\frac{n_{ij}}{N} \times log\frac{n_{ij}}{N})$$

Пептиди су избацивани из скупа S тако да се максимизује мера ентропије. У свакој итерацији је биран подскуп величине $S-1$ такав да се максимизује ентропија. Овај поступак је примењиван онолико пута колико је било потребно да се достигне тачност модела од 80%.

- **KISS** предиктор је трениран над подацима из SYFPEITHY3.1, MHCBN3.1, LANL3.1 и IEDB3.1 [41]. У основи је модел заснован на техници подржавајућих вектора. Предиктор не врши предвиђање за појединачне алеле, већ су алели груписани у супертипове према сличности *MHC* молекула, тиме је омогућено предвиђање и за алеле за које не постоје или постоји мали број експериментално утврђених података. У основи је тврђење да ако се пептид везује за један молекул *MHC* класе врло вероватно ће се везивати и за други структурно сличан молекул, за који не постоје експериментални подаци о везивању [34]. *Heckerman* и сарадници су утврдили да постоје информације које су значајне за све везујуће пептиде, у следећем смислу:

  1. Информације везане за специфичне алеле, које су значајне за прављене модела за предвиђање епитопа који се везују за појединачан алел.
  2. Информације везане за специфичан супертип, које су заједничке за све алеле једног супертипа, и користе се за моделе за предвиђање везивања епитопа за тај супертип.
  3. Информације заједничке за све везујуће пептиде.

  У њиховом истраживању су закључили да постоје информације које су заједничке за различите супертипове, односно да груба подела по супертиповима није адекватна. Такође су закључили да хемијске особине

неких амино киселина у везујућим пептидима имају кључну улогу у везивању за различите алеле. Закључак је и да би узимање у обзир утицаја суседних амино киселина било од значаја иако то у свом истраживању нису спровели. Модел у основи KISS предиктора узима као улаз не само енкодирану пептидну секвенцу већ уређени пар $peptide/allele(p, a)$ чији су атрибути производи атрибута p и a.

- **POPI** је први предиктор који узима у обзир физичко хемијске особине амино киселина везујућих и не-везујућих пептида [105]. Разматрана је 531 физичко хемијска особина и биран је подскуп оптималних m физичко хемијских особина. Показало се да је оптимално $m = 23$, а тачност добијеног модела је 64.72%. Број везујућих пептида над којима је трениран модел је 428 из хумане *MHC* класе I (*HLA* I). Подскуп од $m = 23$ физичко хемијске особине је добијен фактор анализом. Проблем максимизације тачности предикционог проблема избором мањег подскупа m особина од n расположивих за улаз у SVM класификатор, еквивалентан је проблему оптимизације бинарног комбинаторног проблема [35]:

$$C(n, m) = \frac{n!}{m!(n-m)!}$$

Решавањем бинарног проблема је утврђено да је оптималан број фактора $m = 23$.

- **SVMHC** предиктор у основи има модел заснован на техници подржавајућих вектора. Направљен је за 26 алела *MHC* класе I [22]. За тренирање модела су издвојени пептиди из две базе: SYFPEITHI3.1 и MHCPEP3.1. Пептиди у овим базама су сакупљани из различитих извора и сматра се да је SYFPEITHI база у том смислу квалитетнија. Како ниједна од ове две базе не садржи експериментално утврђене неепитопе, за тренирање су бирани на случајан начин пептиди из ENSEMBL[21] базе, као и у случају прављења ANNPred предиктора, подразумева се да су у питању неепитопи.

[21] http://www.ensembl.org/

Квантитативне методе

У табели 3.2 је дат преглед предиктора који предвиђају афинитет везивања пептида за одређени молекул *MHC* класе.

Назив	Метода	Шема кодирања	Карактеристике
NetMHC[22]	ANN	Sparse encoding/BLOSUM50	AUC: 0.914
NetMHCpan[23]	ANN	Sparse encoding/BLOSUM50	Pearson: 0.77
MHCPRED[24]	QSAR regression	-	q: 0.3-0.8
SVRMHC[25]	SVM	Sparse encoding/11 physicochemical properties	q:0.6-0.7%

Табела 3.2: *Преглед квантитативних метода за предвиђање T - ћелијских епитопа MHC класе I*

- **NetMHC**[26] и **NetMHCpan**[27] су предиктори развијени у оквиру Центра за анализу биолошких секвенци CBS[28](eng. *Center for Biological Sequence Analysis*) Техничког универзитета у Данској. Показали су се као најтачнији и најпоузданији предиктори, а подржавају и предвиђање за највећи број алела [118][52]. NetMHCpan може да врши предвиђање и за алеле за које не постоје експериментално утврђени подаци. Оба предиктора су заснова на вештачким неуронским мрежама. Пептиди се у овим моделима енкодирају и ретким и BLOSUM енкодирањем. Једна шема се користи за припрему улаза у једну неуронску мрежу а друга за другу неуронску мрежу. На крају се врши линеарна комбинација добијених модела и добија се коначан резултат предвиђања. Ови предиктори се стално унапређују и изнова тренирају новим подацима. Доступни су у види самосталних апликација, као и преко веб сервера не неколико локација[29].

- **MHCPRED**[30] је предиктор заснован на адитивној методи [25]. Предвиђање афинитета везивања пептида се заснива на претпоставци да свака амино киселина на свакој позицији доприноси афинитету везивања. Израчунава се допринос сваке амино киселине а затим се те вредности сумирају. У рачунању доприноса појединачне амино киселине се узимају

[26] http://www.cbs.dtu.dk/services/NetMHC/
[27] http://www.cbs.dtu.dk/services/NetMHCpan/
[28] http://www.cbs.dtu.dk/
[29] http://www.cbs.dtu.dk/services/; http://tools.iedb.org/mhci/
[30] http://www.ddg-pharmfac.net/mhcpred/MHCPred/

у обзир и утицаји суседних бочних амино киселина, на следећи начин:

$$bindingffinity = const + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2}$$

Модел који је у основи овог предиктора је заснован на парцијалној регресији најмањих квадрата (eng. *Partial Least Square*, PLS).

- **SVRMHC** у основи овог предиктора је модел заснован на регресији подржавајућим векторима. Припрема података за улаз у модел је заснована на ретком енкодирању пептида за неке алеле односно енкодирањем са 11-физичко хемијских особина за друге алеле [112]. База из које су добијени подаци за прављење модела је AntiJen3.1.

Методе засноване на вештачким неуронским мрежама (ANNs) се успешно користе још од 1995. године за предвиђање T - ћелијских епитопа. Прву примену су имале за алел HLA-A*02:01. Да би се направио ANN модел (или неки други модел заснован на техникама машинског учења) потребно је трансформисати секвенцу пептида у погодан облик за улаз у модел. Пептидна секвенца се представља нумеричким дескрипторима у виду вектора који се затим задају као улаз у неколико слојева вештачких неурона. Готово све методе засноване на техникама истраживања података као улаз изимају само секвенцу пептида. Изузетак су NetMHCpan/NetMHCIIpan тзв. "пан" предиктори који узимају у обзир и псеудо секвенцу алела за који се врши предвиђање. "Пан" предиктори на тај начин дозвољавају предвиђање и за алеле за које експериментални подаци нису познати. Аутори ових предиктора су изучавали саму структуру HLA алела и утврдили серију амино киселина које се учестало везују за пептиде псеудо секвенци. Већина модела из ове групе захтева да пептидна секвенца буде фиксне дужине, па су направљени засебни модели за све различити дужине пептида. Изузетак су методе засноване на HMMs, где пептидна секвенца представља директан улаз у модел. Начин представљања пептида у погодан, нумерички облик, за улаз у моделе свих набројаних предиктора је приказан у табелама 3.1 и 3.2 у колони "Шеме енкодирања". Може се приметити да су најчешће шеме енкодирања пептидне секвенце: ретко енкодирање (eng. *sparse encoding*), енкодирање матрицама супституције (eng.

BLOSUM[31] *encoding*) и/или енкодирање физичко хемијским особинама амино киселина које улазе у састав пептидне секвенце. Поступак припреме података различитим шемама енкодирања пептида:

- Ретко енкодирање је једноставно, али вероватно и најчешћи начин представљања пептида. Примењено је код ANNPred/nHLAPred, NetMHC, NetMHCpan, SVMHC, SVRMHC предиктора. Како на свакој позицији у пептидној секвенци може да се нађе било која од 20 амино киселина, свака позиција се представља бинарним записом у виду вектора величине 20. Све компоненте вектора су 0 изузев једне 1 на позицији која зависи од тога која амино киселина се енкодира. На овај начин се сваки пептид величине l преводи у вектор димензије $l \times 20$. Пример ретког енкодирања је приказан у табели 3.3.

Амино киселина	Код
A	$\langle 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0\rangle$
R	$\langle 0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0\rangle$
N	$\langle 0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0\rangle$
D	$\langle 0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0\rangle$
C	$\langle 0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,00,0,0,0\rangle$
Q	$\langle 0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0\rangle$
E	$\langle 0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0\rangle$
G	$\langle 0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0\rangle$
H	$\langle 0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0\rangle$
I	$\langle 0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0\rangle$
L	$\langle 0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0\rangle$
K	$\langle 0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0\rangle$
M	$\langle 0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0\rangle$
F	$\langle 0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0\rangle$
P	$\langle 0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0\rangle$
S	$\langle 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0\rangle$
T	$\langle 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0\rangle$
W	$\langle 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0\rangle$
Y	$\langle 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0\rangle$
V	$\langle 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1\rangle$

Табела 3.3: *Пример ретког енкодирања*

- Енкодирање матрицама супституције. Најчешће се користе блок матрице супституције, међутим није неуобичајено да се користе и друге

[31] BLOck SUpstitution Matrix

супституционе матрице. Ове матрице као елементе имају вредности које представљају меру сличности амино киселина на одговарајућим позицијама, и користе се за утврђивање веза и сличности протеинских секвенци у не-сродним протеинима. Употреба ових матрица има два различита циља: или да оцени сличност пептидне секвенце са већ експериментално утврђеним епитопима, како би се на основу добијене мере оцене пептид класификовао као епитоп или неепитоп, или се елементи матрице узимају као компоненте вектора којим се представља пептид и припрема за улаз у неку од набројаних метода. На слици 3.2 је приказана BLOSUM62 матрица:

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	8	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Слика 3.2: *Блок супституциона матрица BLOSUM62*

Најчешће коришћене блок матрице супституције су BLOSUM50, BLOSUM42 и BLOSUM62. Једна од првих супституционих матрица је PAM матрица, и она се често користи у исте сврхе као BLOSUM матрица.

- Енкодирање физичко хемијским особинама. Свака амино киселина се представља као вектор нумеричких вредности изабраних физичко хемијских особина као што су: хидрофобност, хидрофилност, поларност, и друге. Пример енкодирања физичко хемијским осибинама је илустрован

на слици 3.3, где се узима M физичко хемијских особина за представљање пептида.

Слика 3.3: *Илустрација поступка енкодирања пептида ФХ особинама.*

У исцрпним истраживањима о хидрофобности/хидрофилности епитопа у уређеним и неуређеним структурама протеина [71][82], сумарни резултати су приказани у другом делу тезе, је утврђено да су епитопи богатији хидрофобним амино киселинама, као и да је њихова заступљеност у неуређеним деловима протеина знатно слабија него у уређеним или прелазним регионима. Сви добијени резултати указују да физичко хемијске особине амино киселина које улазе у састав пептида могу бити добар избор за атрубуте вектора којим ће бити представљен пептид за улаз у неку од метода истраживања података.

3.3 Недостаци постојећих метода за предвиђање T - ћелијских епитопа

Већина набројаних предиктора, изузев предиктора из CBS и IEDB групе алата, нису доступни у виду самосталних апликација, већ само у виду веб сервера. Главни недостатак ових предиктора је што се не могу лако тестирати, и

непогодни су за примену над великим бројем протеина. Чак и када се само један протеин проследи као улаз, време одзива и чекања на резултате је јако дуго или се добије само порука о грешци. Још важније, предиктори користе податке из различитих извора (база података) без вођења рачуна о квалитету тих података, експерименталним методама којима су везујући пептиди добијени, да ли су у питању природно обрађени епитопи, или везујући пептиди, итд. Исти пептиди су често означени као везујући и не-везујући [85]. Узимање не-везујућих пептида из произвољних база података без да су експериментално утврђени као не-везујући носи додатне ризике за добијање неквалитетних модела, и чини их мање поузданим. Треба нагласити и да је већина набројаних предиктора заснована на моделима направљеним на малом скупу пептида за појединачне алеле, што их такође чини мање поузданим.

У следећем поглављу су описани нови модели засновани на техници подржавајућих вектора за класификацију и регресију. Модели користе три нове шеме енкодирања пептидне секвенце. Да би се избегли сви претходно наведени проблеми и неконзистентност података, коришћени су искључиво експериментално потврђени подаци из IEDB3.1 базе, која се сматра најпоузданијим извором експериментално утврђених *MHC* везујућих и не-везујућих пептида.

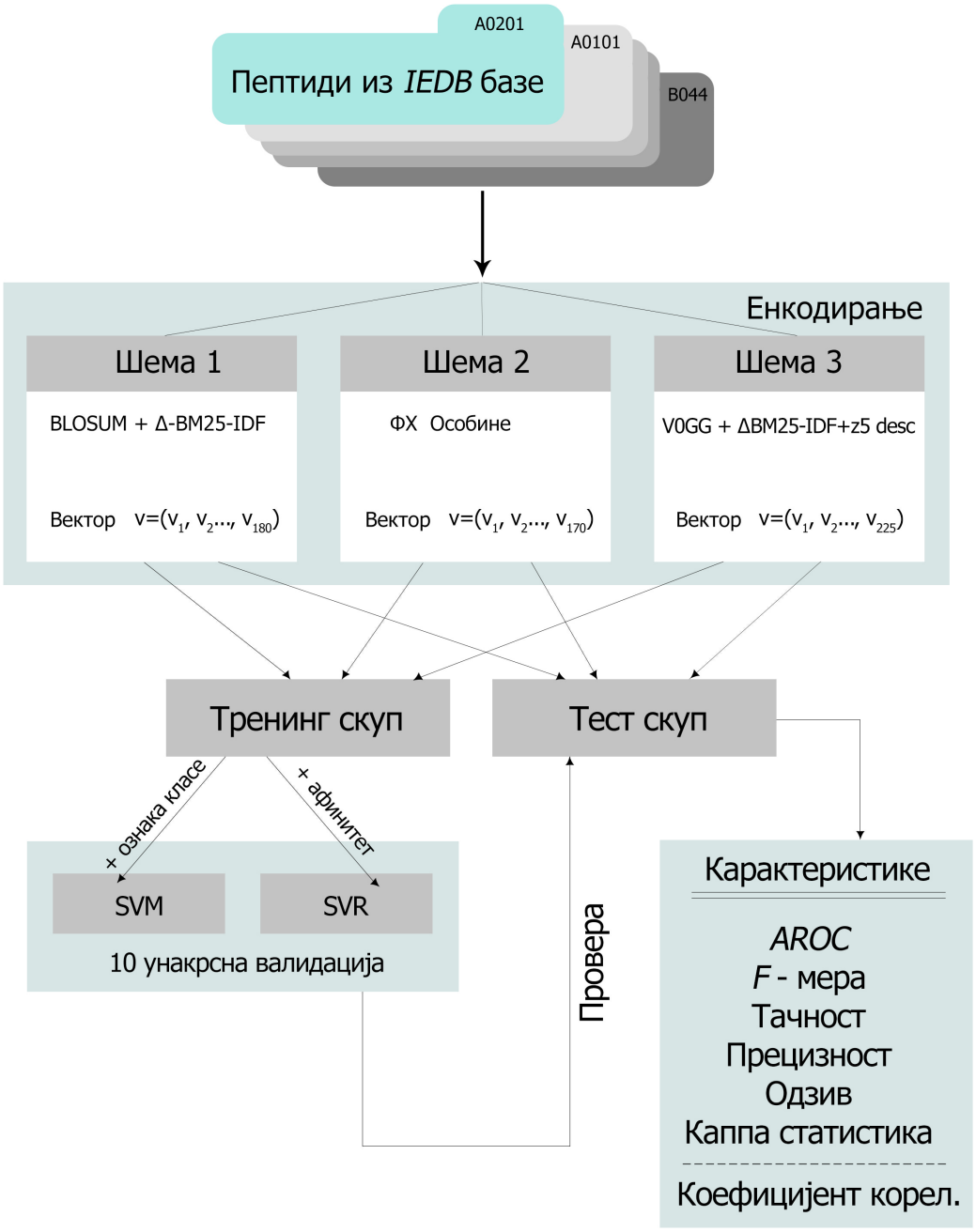
Поглавље 4

Нове методе за предвиђање Т - ћелијских епитопа

Главни допринос ове дисертације су нови развијени модели за предвиђање Т - ћелијских епитопа. Развијени модели се разликују у начину представљања пептида као улаза у модел. За припрему података и представљање пептида у облик погодан за улаз у моделе су направљене три нове шеме енкодирања пептида. Свака од шема је искоришћена за улаз како у модел за бинарну класификацију (SVM) тако и у регресиони модел (SVR). Резултат су три модела за квантитативно и три за квалитативно предвиђање Т - ћелијских епитопа. У оквиру припреме података развијени су и модели за бинарну класификацију засновану на груписању епитопа и неепитопа техником к-средина. Ови модели су дали значајне нове информације које су укључене у прављење резултујућих модела. Шематски приказ овог поглавља је дат на слици 4.1

4.1 Материјал

За потребе прављења нових модела подаци о везујућим и не-везујућим пептидима су узети из IEDB3.1 базе, верзија из Јуна 2015. године. Преузета је комплетна база са свим везујућим и не-везујућим пептидима. Даље истраживање је ограничено само на пептиде дужине 9 (нонамере) јер је познато да су епитопи *MHC* класе I најчешће нонамери. Такође, истраживање је ограничено само на *MHC* класу I, јер за ову класу постоји довољно података за већи број алела, док је за *MHC* класу II довољно података било само за 3 алела. Додатан проблем везан за податке *MHC* класе II је и тај што су доступни подаци



Слика 4.1: *Шематски приказ дијаграма тока нових модела*

о пептидима различите дужине међу којима је већина пептида величине 15, а релативно мали број дужине 9. Нова база података која је послужила за даље истраживање је формирана тако што су из основног скупа избачени следећи подаци:

1. где није било довољно пептида за појединачан алел да би се направио поуздан модел,

2. пептиди за које није постајала и квалитативна и квантитаивна експериментално утврђена мера (да ли се везује / и са којим афинитетом),

3. пептиди означени и као позитивни (везујући) и негативни (не-везујући) а налазе се у истом протеину на истој позицији,

4. пептиди који садрже амино киселине које не припадају основном алфабету $\alpha = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$.

У табели 4.1 су приказани алели за које су направљени одговарајући модели, као и број пептида (везујућих и не-везујућих) који је претходном процедуром установљен и над којим је рађено даље истраживање. Изабрано је 15 алела *MHC* класе I, за које је било највише података, и за сваки алел је направљено по шест модела за предвиђање епитопа три за квантитативно и три за квалитативно предвиђање.

Сви подаци који су прикупљени за прављење (тренирање и тестирање) модела, направљени модели, и програми који су написани за енкодирање пептидне секвенце и припрему података за улаз у класификационе моделе су приложени као додатак тези на пратећем диску. Развијени програми за припрему података и енкодирање пептида су писани у програмским језицима C и Java.

4.2 Припрема података

Најважнији корак у дизајнирању поузданог класификационог модела је избор методологије припреме и представљања података, као и избор атрибута. У оквиру ове тезе су направљени класификациони и регресиони модели засновани на техници подржавајућих вектора, описаној у поглављу 2. Најпогоднији улаз у ове методе је у форми вектора. Проблем који се поставља је

	Број пептида		
Алел	Број позитивних	Број негативних	Укупно
A0101	496	2081	2577
A0201	2872	3043	5915
A0202	883	456	1339
A0203	1133	1171	2304
A0206	1048	647	1695
A0301	1386	2628	4014
A1101	1389	1917	3306
A2402	870	698	1568
A2601	350	2184	2534
A3101	921	2049	2970
A6802	727	1655	2382
B0702	996	1610	2606
B0801	555	1115	1670
B1501	1443	1717	3160
B4403	174	235	409

Табела 4.1: *Алели за које је рађено истраживање и број пептида расположивих за сваки од алела. Сви пептиди су експериментално утврђени, и за њих постоји експериментално утврђена мера о афинитету везивања.*

да сваки пептид буде представљен вектором чије су компоненте добијене применом неке тежинске функције. Како је и описано у поглављу 3.2.3, најконвенционалније представљање пептида, које је коришћено у постојећим методама за предвиђање T ћелијских епитопа је ретко бинарно и BLOSUM енкодирање. У неким од постојећих метода је коришћена и комбинација ове две стратегије. У оквиру тезе су предложене три нове стратегије енкодирања пептида и атрибути (компоненте вектора) су формирани применом предложених стратегија. Применом нових стратегија енкодирања су припремљени подаци за улаз у SVM и SVR моделе. SVM модели су направљени за бинарну класификацију тј. предвиђају да ли је разматрани пептид потенцијални епитоп или неепитоп, док су SVR модели направљени за предвиђање афинитета везивања разматраног пептида.

4.2.1 Δ-TFIDF и Δ-BM25-IDF технике

Рачунање учесталости амино киселина на одређеним позицијама у епитопима и неепитопима има за циљ издвајање информација и правила у појављивању амино киселина код епитопа и неепитопа које би била од значаја за

лакшу класификацију. Све методе засноване на квантитативним матрицама примењују рачунање фреквентности амино киселина. У оквиру ове тезе је, уместо стандардног израчунавања фреквентности, описан нови модификован приступ рачунања фреквентности амино киселина који се успешно користи у класификовању докумената. Техника рачунања фреквентности која се овде користи је Δ - TFIDF. Први пут је уведена 2009. године у истраживању описаном у [67]. Уобичајен поступак у класификовању докумената, када је материјал за тренинг задат у виду речи, је да се речима придружују одговарајуће тежине и тиме добије нови скуп атрибута, који је много погоднији улаз у SVM модел. *Martineau* и *Finin* су у оквиру свог истраживања [67] представили нову технику, названу Δ-TFIDF, која додељује одговарајуће тежине термовима или речима у документу, а тиме преводи атрибуте у векторе тежина. Овај начин придруживања тежина се показао као много ефикаснији од стандардног рачунања фреквентности у класификацији докумената и истраживању мишљена на друштвеним мрежама [48]. Δ-TFIDF мера се веома лако рачуна, а техника је лака за разумевање и имплементацију.

Примери успешне примене ове технике, и побољшања у односу на претходне моделе су детаљно објашњени у [47] где се мера Δ-TFIDF рачуна за појављивања сваке речи, или две узастопне речи у тексту. Алтернатива овом приступу је да се речима придружују вредности 1 или 0 (или се појављују или не) [79] [113], или им се пак придружује вредност IDF4.2 мере као у [54]. Рачунањем Δ-TFIDF тежина за речи или n-граме у позитивном и негативном скупу је знатно поправљена тачност модела направљених у оквиру истраживања описаних у [67].

Проблеми у којима се ова техника успешно користила могу да се представе на сличан начин као и проблем који се разматра у овој тези. У описаним примерима је техника примењена за класификовање речи у два скупа докумената. Класификовање пептида на епитопе и неепитопе може да се поистовети са описаним проблемима, где на основу појављивања појединачних амино киселина или биграма у пептиду треба да се установи да ли је пептид епитоп или неепитоп.

Примена Δ-TFIDF технике на проблем предвиђања Т - ћелијских епитопа

У проблемима класификовања докумената Δ-TFIDF мера се рачуна за термове/речи у документима. Овде је поступак готово исти с тим да се ова мера рачуна за сваку амино киселину у пептиду. Дакле овде терм t, као елемент ове методе, може бити сама амино киселина (eng. *unigram*) или две узастопне амино киселине (eng. *bigram*) из скупа свих амино киселина које се јављају у пептиду. На пример, ако је дат пептид $p = LVIKALLEV$, t може бити било који елемент скупа $\alpha = \{L, V, I, K, A, L, L, E, V, LV, VI, IK, KA, AL, LL, LE, EV\}$. Формулација проблема и одговарајуће дефиниције потребне за примену ове технике рачунања фреквентности су:

Дефиниција 4.1 *$df(t, S)$ представља фреквентност терма t у скупу свих пептида S.*

Дефиниција 4.2 *$idf(t, S) = log_2 \frac{|S|}{df(t,S)}$ представља инверзну фреквентност терма t у скупу свих пептида S, где је $|S|$ кардиналност скупа S.*

Дефиниција 4.3 *$tf(t, Peptide)$ представља број појављивања терма t у самом пептиду $Peptide$*

На основу уведених дефиниција 4.1, 4.2 и 4.3 Δ-TFIDF мера се уводи на следећи начин:

$$\Delta tfidf(t_i, Peptide, S^+, S^-) = tf(t_i, Peptide) * log_2 \frac{|S^+|}{df(t_i, S^+)} * \frac{df(t_i, S^-)}{|S^-|} \quad (4.1)$$

Где t_i представља амино киселину на позицији i у пептиду $Peptide$ из скупа S. S^+ и S^- су подскупови скупа S позитивних (везујућих) пептида (епитопа) и негативних (не-везујућих) пептида (неепитопа). $|S^+|$ и $|S^-|$ представљају кардиналност позитивног и негативног подскупа. Овај начин рачунања фреквентности придружује већу важност термима (амино киселинама и/или биграмима) који немају приближно исту дистрибуцију у позитивном и негативном скупу, а мању важност оним термима чија је дистрибуција у овим подскуповима приближно једнака. Применом Δ-TFIDF технике се пептид представља вектором тежина својих амино киселина. За нонамере, где је терм t амино киселина из пептида, тај вектор је величине 9. Очигледан проблем се јавља у случајевима када је $df(t, S^{\pm}) = 0$ тј. када

постоје термови који се не јављају у оба подскупа од S. Други проблем је проблем нелинеарности фреквентности амино киселина. Постоји неколико предлога решења овог проблема [78]. Проблеми се једноставно превазилазе увођењем фактора изглађивања (eng. *smoothing factor*). Спроведено је више истраживања у којима су упоређивани резултати добијени увођењем различитих фактора изглађивања, и закључак је да $BM25$ фактор даје најбоље резултате [48]. Из тог разлога је и овде уведен $BM25$ фактор изглађивања, и његовим укључивањем се модификује Δ-TFIDF мера, а фреквентност се рачуна на следећи начин:

$$\Delta BM25idf(t_i, Peptide, S^+, S^-) = log\frac{(|S^+| - \Delta tfidf(t_i, S^+) + 0.5) * tfidf(t_i, S^-) + 0.5}{(|S^-| - \Delta tfidf(t_i, S^-) + 0.5) * tfidf(t_i, S^+) + 0.5} \tag{4.2}$$

Уведена мера придружује терму позитивне и негативне вредност на следећи начин:

$$deltaBM25idf(t) = \begin{cases} value < -1, & \text{терм t је много заступљенији у позитивном скупу} \\ value > 1, & \text{терм t је много заступљенији у негативном скупу} \\ -1 \leq value \leq 1 & \text{терм t је приближно једнако заступљен у оба скупа} \end{cases}$$

4.2.2 Енкодирање блок матрицама супституције

Енкодирање супституционим матрицама је описано у одељку 3.2.3. У оквиру ове тезе је првобитно коришћена BLOSUM62 матрица. Свака амино киселина у пептиду је предствљена вектором врсте BLOSUM62 матрице која одговара тој амино киселини, што може да се запише на следећи начин:

$$\phi : \Sigma \to R^n$$

где је $\Sigma = \{a_1, a_2, ..., a_n\}$ коначна азбука свих амино киселина величине $n = 20$, $M \in R^{n \times n}$ је симетрична матрица супституције[1] за коју важи да је $M_{ij} = M_{ji}$. Вредности у врсти ове матрице представљају меру сличности те амино киселине са свим постојећим амино киселинама. Мера је позитивна код супституција које се често јављају и негативна код супституција која се не јављају или се јако ретко јављају. Пептид је применом ове шеме представљен вектором величине $9 \times 20 = 180$ ($\varphi : \Sigma^l \to R^{l \times n}$), где је $l = 9$. Овај вид енкодирања је значајан

[1] не нужно BLOSUM62, може да буде и BLOSUM42, BLOSUM50, PAM, VOGG, итд.

јер обезбеђује додатне информације нпр. чак и када се нека амино киселина не појављује у пептидима у оквиру позитивног и негативног скупа, а могућа је супституција неке друге амино киселине том амино киселином та информација ће бити "ухваћена" у моменту тренирања модела.

4.2.3 Шема 1: комбинација Δ-BM25-IDF и BLOSUM62 енкодирања

Прва нова шема енкодирања коришћена у овој тези је заснована на комбинацији Δ-BM25-IDF мере за унираме и BLOSUM62 стратегије енкодирања. Пептид се прво енкодирана BLOSUM62 шемом, тако што се свака амино киселина у пептиду замењује одговарајућом врстом из BLOSUM62 матрице и тиме репрезентује вектором величине 180. На пример амино киселина L се представља вектором врсте:

$$L : \langle -1, -2, -3, -4, -1, -2, -3, -4, -3, 2, 4, -2, 2, 0, -3, -2, -1, -2, -1, 1, -4, -3, -1, -4 \rangle$$

Ово може, једноставности ради, да се запише као $\phi_i = \phi(a_i)$ за свако $i = 1, ..., 9$, а добијени вектор $\langle \phi_{ij} \rangle, i = 1, ..., 9; j = 1, ...20$. Компоненте добијеног вектора су затим множене Δ-BM25-IDF тежинама добијеним за одговарајућу амино киселину (ону на коју се односи одговарајућа компонента вектора: $\phi_{ij} * \Delta - BM25 - IDF(a_j)$). Резултујући вектор је такође димензије 180. Само BLOSUM енкодирање има следећи недостатак: немогуће је раздвојити случајеве када се једна амино киселина супституише са истом мером у друге две различите амино киселине. Нпр. амино киселина L се супституише у амино киселине A, Y и T са истом мером. Множењем BLOSUM мере са Δ-BM25-IDF тежинама се ти случајеви раздвајају. Може се уочити да у одређеним случајевима може да дође до губитка информација, јер вредности BLOSUM62 супституција могу бити позитивне у случају честих (више вероватних) супституција и негативне у случају ретких (мање вероватних) супституција. Исто важи за вредности Δ-BM25-IDF тежина, на основу једнакости 4.2. Као последица множења са негативним вредностима супституције, случај који је проблематичан је случај ретких супституција, када се одговарајућа амино киселина учестало јавља у једном од два скупа (позитивном или негативном). Да би се избегао могући проблем BLOSUM62 матрица је замењена VOGG [109] матрицом, која је

једна од варијација BLOSUM62 матрице али са свим позитивним вредностима. Множењем са позитивним вредностима сви случајеви остају раздвојени.

4.2.4 SVM и SVR модели засновани на шеми 1

Пептиди из скупа свих пептида придруженог једном алелу су представљени вектором добијеним применом шеме 1. Добијени вектори су погодан улаз за моделе засноване на подржавујућим векторима. Уобичајен поступак у проблемима класификације је да се основни скуп података подели на два подскупа, од којих се један користи за тренирање, а други за тестирање модела (поступак је детаљније описан у глави 2). Овде је основни скуп података (пептида) S, који се састоји од подскупова S^+ и S^- подељен у односу 70% : 30% за тренирање односно тестирање модела. Поступак поделе је извршен на следећи начин:

- 70% пептида је узето и из скупа епитопа S^+ и 70% из скупа неепитопа S^- и од њих је формиран скуп за тренирање модела $TrainSet$.
- Преосталих 30% пептида из оба скупа је искоришћено за тестирање модела, од њих је направљен $TestSet$.
- Број епитопа и неепитопа се у оба скупа значајно разликује, видети табелу 4.1. Да би се избегло преприлагођавање модела једној (бројнијој) класи извршено је балансирање података у оба добијена скупа на следећи начин:
  1. Израчунат је минимум од броја епитопа и неепитопа у одговарајућим скуповима (и скупу за тренинг и тест) $m = min(n_1, n_2)$, где је n_1 број епитопа а n_2 број неепитопа.
  2. На случајан начин је изабрано m пептида из бројнијег скупа, док је мање бројан скуп узет цео, и они су даље коришћени у тренирању и тестирању модела. Све што је преостало од пептида из $TrainSet$-а и $TestSet$-а је искоришћено за другу итерацију тестирања добијених модела.

Описани поступак балансирања епитопа и неепитопа је уобичајена процедура у процесу прављења класификационих модела (видети одељак 2). Ту су и детаљно истакнуте предности оваквог поступка, а једина мана се односи на тренирање

модела над мањим скупом вредних експерименталних података. Како је у овом случају идеја била да се тестирају предложене мере фреквентности као потенцијално добре за овај вид проблема, то незнатан губитак информација није од пресудног значаја.

Избор алгоритама и оптимизација параметара за технику подржавајућих вектора су детаљно објашњени у поглављу 4.3. Резултати класификационог и регресионог модела добијеног применом шеме 1. енкодирања су дати у табели 5.3. У случајевима где је примењена бинарна класификација сваком добијеном вектору, који представља пептид, је придружена ознака (још једна компонента у вектору) $y \in \{1, 0\}$. Ознака 1 је придружена епитопима, док је ознака 0 придружена неепитопима. У случајевима где су прављени регресиони модели сваком вектору је придружен афинитет везивања који је експериментално утврђен и узет је из IEDB базе. Експериментално утврђени афинитет везивања узима вредности изражене у IC_{50} из интервала $[0, 50000]$. Вредности афинитета су за потребе направљених модела скалиране у интервалу $[0, 1]$ према формули:

$$logAffinity = 1 - log_{50k}(Affinity)$$

Скалирана вредност афинитета је била згодна и за поређење са постојећим предикторима јер неки од предиктора дају резултате управо скалиране вредности афинитета. За потребе поређења са предикторима који као излаз приказују предвиђени афинитет изражен у IC_{50} мери, је афинитет израчунат инверзним поступком.

Енкодирање пептида Δ-BM25-IDF техником за униграме и биграме

У претходном моделу нису узете у обзир информације везане за појављивање две узастопне амино киселине, тј. није разматран утицај суседних амино киселина, већ су амино киселине на свакој позицији посматране независно од тога шта јој претходи или следи. Рачунањем Δ-BM25-IDF тежина за биграме у пептиду и њиховим укључивањем у процес енкодирања пептида се укључују и информације о суседним амино киселинама. Нпр. ако је дат пептид $p = LVIKALLEV$ за све биграме из скупа $LV, VI, IK, KA, AL, LL, LE, EV$ се рачунају Δ-BM25-IDF тежине, потпуно аналогно као за униграме. Пептид се у овом случају представља вектором тежина својих амино киселина и свих

узастопних биграма које садржи. Дакле пептид p се представља вектором:

$$w = \langle w_{11}, w_{12}, ..., w_{19}, w_{21}, w_{22}, ..., w_{28} \rangle$$

дужине 17, где су компоненте вектора $w_{1i}, (i = 1, .., 9)$ тежине униграма на позицији i а компоненте $w_{2j}, (j = 1, .., 8)$ тежине биграма на позицији j.

4.2.5 Енкодирање пептида физичко хемијским особинама, ФХ

Како за биграме нису дефинисане матрице супституције, уместо BLOSUM матрице су укључене физичко хемијске особине амино киселина које једноставно могу да се примене и на биграме. Из [103] је преузето 119 физичко хемијских особина амино киселина. По 10 физичко хемијских особина је придружено свакој амино киселини и одговарајућем биграму из пептида. Избор 10 "најбољих" физичко хемијских особина за амино киселине и биграме је описан у поглављу 4.4.1. У оквиру ове шеме се избором по 10 најбољих физичко хемијских особина за све униграме и биграме за појединачне алеле управо истиче специфичност сваког алела и епитопа који се везују за тај алел. Пептид се представља вектором $v = \langle v_{ik}, v_{jk} \rangle, i = 1, ..., 9; j = 1, .., 8; k = 1, ..., 10$, димензије 170 (по 9×10 компоненти за униграме v_{ik} и 8×10 компоненти за биграме v_{jk}).

4.2.6 Шема 2: комбинација Δ-BM25-IDF и ФХ за униграме и биграме

У шеми 2 се комбинују: Δ-BM25-IDF техника примењена на униграме и биграме и енкодирање ФХ особинама за униграме и биграме. Сваки пептид се представља вектором димензије 170, који се добија на следећи начин:

(1) Пептид се преводи у вектор применом претходно описаног ФХ енкодирања за униграме и биграме. Добија се вектор v димензије 170.

(2) Исти пептид се енкодира и шемом Δ-BM25-IDF за униграме и биграме и добија se вектор тежина $w = \langle w_{ij} \rangle, i = 1, 2; j = 1, .., 9 - i + 1$ димензије 17.

(3) Резултујући вектор се добија када се измноже одговарајуће Δ-BM25-IDF тежине са редом физичко хемијским особинама примењеним на ту амино киселину, аналоган поступак следи за биграме. Математички презентовано решење на основу добијених вектора v и w је проширење вектора w тако да се свака компонента вектора понови 10 пута, а затим се изврши скаларни производ добијеног вектора и вектора v. Резултујући вектор је димензије 170.

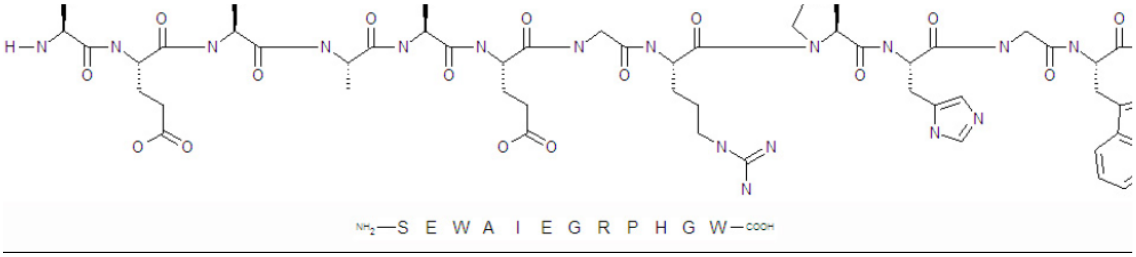
4.2.7 SVM и SVR модели засновани на шеми 2

Скуп свих пептида је преведен шемом 2 у векторску репрезентацију и подаци су припремљени за улаз у SVM и SVR моделе. Скуп свих података је подељен, као у претходном случају на податке за тренирање моделе $TrainSet$ и $TestSet$ у односу 70 : 30 %. Битна разлика у односу на поступак описан у 4.2.4 је што овде није балансиран број епитопа и неепитопа. Како су резултати за моделе добијене у 4.2.4, засноване на шеми 1 енкодирања веома добри, потврђено је да је мера Δ-BM25-IDF добра за представљање пептида. Идеја код прављења ових модела је да се укључе сви доступни експериментални подаци, и да се не губе значајне информације. С тим у вези, је важно напоменути да је за оцену квалитета добијених модела из оваквог скупа података значајно укључити одговарајуће мере за оцену квалитета модела, и да тачност и прецизност самостално нису поуздане мере (видети поглавље 2.1.1, где су детаљно објашњене мере и поступци за оцену квалитета модела). Мере које су укључене за оцену квалитета модела поред тачности и прецизности су: одзив, F-мера, $AROC$ мера и Капа статистика. Однос епитопа и неепитопа у скупу за тренирање и тестирање је задржан (70 % епитопа и 70 % неепитопа је укључено у скуп за тренирање, 30 % епитопа и 30 % неепитопа у скуп за тестирање модела). Оба скупа су небалансирана - имају различит број епитопа и не епитопа. Као и у претходно описаном поступку прављена модела, и овде је за бинарне моделе класификације пептид означен на одговарајући начин (као у поглављу4.2.4), а за регресионе моделе је пептиду придружена компонента која представља афинитет везивања. Резултати добијених модела применом ове стратегије припреме података су дати у табели 5.4 у глави 5.

4.2.8 Енкодирање молекуларним дескрипторима

Молекуларни дескриптори представљају нумеричке вредности придружене свакој амино киселини, а добијени су комбиновањем неколико стотина молекуларних и хемијских особина пептида. Скоро 1000 оваквих особина је разматрано у [86], где је примењена анализа главних компоненти (eng. *Principal component analysis*, PCA) да би се број дескриптора амино киселина редуковао. Показало се да одређене групе главних компоненти добијених овом анализом добро описују пептид. У пракси се најчешће узимају $z3$ и $z5$ дескриптори за математичко представљање пептида. У овој тези су разматрани $z5$ дескриптори. Одговарајући дескриптори су приказани у табели 4.2. Главна разлика између енкодирања физичко хемијским особинама и z дескрипторима је што физичко хемијске особине покривају већи скуп особина амино киселина (*side chain volume*, *pKa*, *isoelectric point*, итд., све особине су приложене у додатку у табели Б.1), док z дескриптори представљају само опис скупа од четири молекуларне особине: запремину (eng. *steric bulk*), електростатичност, хидрофобност и електронски ефекат, $z5$ дескриптори их представљају са 5 карактеристика. Предност представљања пептида z дескрипторима је у томе што се смањује димензионалност проблема. Пептид се репрезентује $z5$ дескрипторима тако што се свака амино киселина у пептиду замени са нумеричким вредностима дескриптора. Са $z5$ дескрипторима се пептид преводи у вектор димензије $9 \times 5 = 45$ нумеричких вредности. Значај представљање пептида z дескрипторима је провераван у QSAR[2] моделима (eng. *Quantitative structure–activity relationship*), где је утврђено да су добра метематичка репрезентација за краће пептида (≤ 15 амино киселина). Са z дескрипторима је могуће нумерички квантификовати структурне варијације унутар пептида. На слици 4.2 је приказана упрошћена структура једног молекула и његових атомских веза, за који је дата и пептидна секвенца која се може једноставно превести у вектор чије компоненте су z дескриптори.

[2] http://www.qsartoolbox.org/about



Слика 4.2: *Графичка репрезентација пептида: (горе) у виду атомске везе, (доле) у виду амино киселинске секвенце.*

АК	С-НАЗИВ	НАЗИВ	Z1	Z2	Z3	Z4	Z5
A	Ala	alanine	0.24	-2.32	0.60	-0.14	1.30
R	Arg	arginine	3.52	2.50	-3.50	1.99	-0.17
N	Asn	asparagine	3.05	1.62	1.04	-1.15	1.61
D	Asp	aspartic acid	3.98	0.93	1.93	-2.46	0.75
C	Cys	cysteine	0.84	-1.67	3.71	0.18	-2.65
Q	Gln	glutamine	1.75	0.50	-1.44	-1.34	0.66
E	Glu	glutamic acid	3.11	0.26	-0.11	-3.04	-0.25
G	Gly	glycine	2.05	-4.06	0.36	-0.82	-0.38
H	His	histidine	2.47	1.95	0.26	3.90	0.09
I	Ile	isoleucine	-3.89	-1.73	-1.71	-0.84	0.26
L	Leu	leucine	-4.28	-1.30	-1.49	-0.72	0.84
K	Lys	lysine	2.29	0.89	-2.49	1.49	0.31
M	Met	methionine	-2.85	-0.22	0.47	1.94	-0.98
F	Phe	phenylalanine	-4.22	1.94	1.06	0.54	-0.62
P	Pro	proline	-1.66	0.27	1.84	0.70	2.00
S	Ser	serine	2.39	-1.07	1.15	-1.39	0.67
T	Thr	threonine	0.75	-2.18	-1.12	-1.46	-0.40
W	Trp	tryptophan	-4.36	3.94	0.59	3.44	-1.59
Y	Tyr	tyrosine	-2.54	2.44	0.43	0.04	-1.47
V	Val	valine	-2.59	-2.64	-1.54	-0.85	-0.02

Табела 4.2: $z5$ *дескриптори за одговарајуће амино киселине*

4.2.9 Шема 3: комбинација енкодирања Δ-BM25-IDF техником, VOGG матрицом и $z5$ - дескрипторима

Шема 3 представља проширење шеме 1 z - дескрипторима. У шеми 3 су комбиноване три стратегије енкодирања: супституционом матрицом VOGG, Δ-BM25-IDF тежинама за унограме и $z5$ - дескрипторима. Свака амино киселина у пептиду је замењена одговарајућом врстом VOGG матрице, и добијен је вектор димензије 180. Свака компонента добијеног вектора је помножена Δ-BM25-IDF тежином амино киселине на коју се односи компонента у вектору. Димензија резултујућег вектора остаје непромењена. Иницијални пептид се енкодира и z - дескрипторима где се добија вектор димензије 45. Да би се уврстиле у модел све информације ова два вектора се спајају и добија се вектор

димензије 225.

4.2.10 SVM и SVR модели засновани на шеми 3

Сви пептиди придружени једном алелу су енкодирани шемом 3 чиме је сваки пептид представљен вектором димензије 225, који представља погодан улаз за SVM и SVR моделе. Сваки вектор је проширен за још једну компоненту: нумеричку вредност 0 или 1 за потребе прављења модела бинарне класификације, или вредност афинитета пептида из интервала $[0, 1]$ за потребе прављења регресионих модела. За прављење модела, скуп свих пептида придружених једном алелу је подељен у подскуп за учење и подскуп за тестирање као у 4.2.7. Поступак је поновљен за све алеле укључене у истраживање, приказане у табели 4.1, и за све алеле су направљени засебни модели. Резултати добијених модела применом ове стратегије припреме података су дати у табели 5.5 у глави 5.

4.3 Експерименти - тренирање модела

За сваку од три описане шеме представљања пептида су направљена по два модела за сваки од 15 алела. Један модел је направљен за бинарну класификацију, тј. предвиђање да ли је улазни пептид епитоп или неепитоп. Други модел је регресиони и направљен је за предвиђање афинитета везивања улазног пептида. Афинитет везивања пептида је за потребе регресионих модела логаритамски скалиран на интервал $[0, 1]$. Провера тачности моделе је урађена 10-унакрсном провером на скупу података за тренирање модела. Циљ унакрсне провере је оптимизација параметара модела у циљу добијања модела са најбољим перформансама на скупу података за тренирање. Коначна провера тачности добијених модела је рађена на скупу података за тестирање. За прављене модела заснованих на техници подржавајућих вектора је коришћен Weka[3] алат. Алгоритам који је у основи овог алата за технику подржавајућих вектора је алгоритам секвенцијалне минималне оптимизације (eng. *Sequential minimal optimization*) [87][88]. У изведеним експериментима се показало да је најбоља кернел функција за проблем бинарне класификације полиномијална,

[3] http://www.cs.waikato.ac.nz/ml/weka/

а за регресиони проблем радијално засноване функције, *RBF*. Унакрсном провером су оптимизовани параметри C и *exponent* за полиномијални кернел односно C и *gamma* за *RBF* кернел. За оптимизацију параметара је коришћена метода похлепног претраживања (eng. *Grid search*), а мере које се максимизују у моделу су Капа статистика и F-мера. Са утврђеним оптималним параметрима модели су примењени на тестни скуп података. У глави 5 су приказани резултати свих добијених модела применом унакрсне провере, као и применом модела са оптимизованом параметрима на тестном скупу података. Резултати за различите моделе су међусобно упоређивани, такође су приказани и резултати поређења са резултатима других постојећих метода за предвиђање Т - ћелијских епитопа.

4.4 Бинарна класификација заснована на кластеровању

4.4.1 Поступак израчунавања "најбољих" физичко хемијских особина АК

Физичко хемијске особине амино киселина (АК) за представљање пептида у постојећим методама за предвиђање Т - ћелијских епитопа су коришћење на један од два начина. Оба приступа су заснована на примени технике анализе главних компоненти (eng. *Principal Component Analysis*, PCA) или фактор анализе (eng. *Factor Analysys*, FA). Технику анализе главних компоненти је први пут увео Карл Пирсон 1901. године, за две променљиве. Знатно касније појавом рачунара је уследила широка употреба ове технике. Анализа главних компоненти представља једну од најједноставнијих мултиваријантних техника. Примењује се када је велики број променљивих у скупу редудантан, тј. када се више променљивих односи на исту димензију а не дају никакву нову информацију која већ није обухваћена неком другом променљивом. Геометријски гледано, то значи да на простору од k димензија имамо p променљивих при чему је $k < p$. Очекује се да ће k највећих компоненти бити довољно да објасни цео скуп података. Циљ анализе је да се узме p променљивих $(X_1, X_2, ..., X_p)$ и да се пронађе њихова комбинација и израчунају нове вредности $(Z_1, Z_2, ..., Z_p)$ које међусобно нису у корелацији. Променљиве Z заправо пред-

стављају главне компоненте и за њихове варијансе важи да су у опадајућем поретку ($Var(Z_1) \geq Var(Z_2) \geq ... \geq Var(Z_p)$). Детаљно објашњење технике се може наћи у [83][38][49]. Техником анализе главних компоненти се редукује број посматраних физичко хемијских особина, где се потом уместо свих физичко хемијских особина узима у обзир неколико компоненти или фактора, који представљају линеарну комбинацију иницијалног скупа физичко хемијских особина. Добијене компоненте су међусобно ортогоналне.

1. У првом приступу се све изабране компоненте примењују на сваку амино киселину у пептиду и добијене нумеричке вредности представљају векторску репрезентацију пептида. Добијени вектор је димензије $9 \times n$, где је n број изабраних главних компоненти. Пример где је коришћен овај начин представљања пептида је POPI предиктор [105].

2. У другом случају се добијене главне компоненте, применом описане технике анализе главних компоненти, примењују на цео пептид. Овај поступак подразумева да се сумирају нумеричке вредности које се добијају применом компоненти за сваку од амино киселина из пептидне секвенце, и применом једне компоненте на пептид се добија једна нумеричка вредност. На овај начин се пептид представља вектором димензије n, где n представља број изабраних главних компоненти. Пример овог начина представљања пептида је коришћен у [39].

Мана описаних приступа је што се избором неколико компоненти сви алели унификују, односно не води се рачуна о специфичностима алела и чињеници да се за појединачне алеле везују епитопи који се разликују по физичко хемијским карактеристикама. У оквиру ове тезе се полази од претпоставке да се епитопи који се везују за специфичне алеле разликују по физичко хемијским особинама, и да би одређивање специфичних физичко хемијских особина карактеристичних за сваку пептид - *MHC* молекул везу могло значајно да побољша перформансе предвиђања Т - ћелијских епитопа. У циљу идентификовања најзначајних физичко хемијских особина пептида који се везују за одређене *MHC* молекуле су направљени бинарни класификациони модели засновани на техници груписања k - срединама. Прво су разматране све расположиве физичко хемијске особине, тачније 119 физичко хемијских особина (23 особине су везане за електронска

својства, 37 за стерна својства, 54 за хидрофобна својства и 5 за водоничне везе амино киселина) које су преузете из [103]. Свака од ових особина је дата као нумеричка вредност за сваку од основних 20 амино киселина. На сваки пептид из скупа свих пептида везаних за један алел је примењена техника Δ-BM25-IDF за додељивање тежина амино киселинама и биграмима у пептиду (поступак је описан у поглављу 4.2.4). Добијени вектор w димензије 17 се затим множи са вредностима изабране физичко хемијске особине за одговарајућу амино киселину. Поступак је приказан кроз пример за једну физичко хемијску особину. Нека је физичко хемијска особина која се разматра f_1:

$$[w_{f_1}] = w \cdot f_1 = \langle w_{11} \cdot f_1(AK_1), w_{12} \cdot f_1(AK_2), .., w_{19} \cdot f_1(AK_9), \\ w_{21} \cdot \frac{f_1(AK_1) + f_1(AK_2)}{2}, w_{22} \cdot \frac{f_1(AK_2 + f_1(AK_3)}{2}, .., w_{28} \cdot \frac{f_1(AK_8) + f_1(AK_9)}{2} \rangle \tag{4.3}$$

Компоненте вектора $[w_{f_1}]$ су затим усредњене са вредностима компоненти својих суседа (претходника и следбеника), нпр. за унираме:

$$\langle v_{1i} \rangle = \left\langle \frac{w_{1i-1} \cdot f_1(AK_{i-1}) + w_{1i} \cdot f_1(AK_i) + w_{1i+1} \cdot f_1(AK_{i+1})}{3} \right\rangle, i = 2, .., 8.$$

За $i = 1$ подразумевано је да је вредност претходника 0, такође за $i = 9$ вредност следбеника је 0. Аналогно је урађено за све узастоне биграме $(B_j), i = 1, ..., 8$ у пептиду:

$$\langle v_{2j} \rangle = \left\langle \frac{w_{2j-1} \cdot f_1(B_{j-1}) + w_{2j} \cdot f_1(B_j) + w_{2j+1} \cdot f_1(B_{j+1})}{3} \right\rangle, j = 2, .., 7.$$

И у случају биграма вредности претходника и следбеника, $j = 1$ и $j = 8$, су постављени на 0. Где се физичко хемијска мера $f_i(B_j)$ за биграме рачуна као $f_i(B_j) = \frac{f_i(AK_j) + f_i(AK_{j+1})}{2}, i = 1, .., 119, j = 2, .., 7$. На овај начин се сваки пептид представља вектором чије су компоненте комбинација претходно добијена два вектора $\langle v \rangle = \langle v_{1i}, v_{2j} \rangle$. За сваку физичко хемијску особину $f_k, k = 1, .., 119$ и сваки алел је извршен исти поступак, и направљен је по један класификациони модел заснован на техници груписања k - срединама. Улаз у класификатор је овако добијени вектор v. Поступак прављења бинарног класификационог модела је следећи:

- Скуп свих пептида који су придружени једном алелу је подељен на подскупове за тренирање и тестирање модела (исто као у 4.2.7).

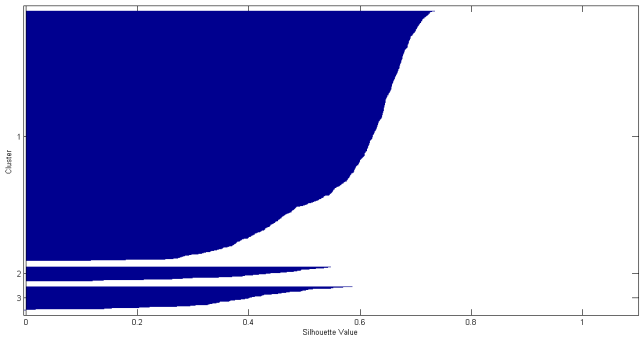
- На скупу података за тренирање модела је примењено груписање техником k - средина, са $k = 3$ кластера, одвојено за подскуп епитопа и подскуп неепитопа. Као мера растојања је коришћено Еуклидско растојање. Еуклидско растојање између тачака $P = (p_1, p_2, .., p_n)$ и $Q = (q_1, q_2, .., q_n)$ у Еуклидском n-простору се дефинише као: $\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + .. + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$. У оквиру истраживања су разматране и друге мере растојања, нпр. косинусно растојање је дало готово исте мере оцена квалитета коначних модела. Број кластера је установљен емпиријски. Већи број кластера је резултовао бољом прецизношћу коначног модела, али се одзив у тим случајевима смањио. За оцену квалитета кластера је коришћена *Silhouette* мера, која се добија на следећи начин:

  - Нека је за сваку тачку скупа i, $a(i)$ просечна вредност одстојања од свих других тачака у оквиру истог кластера. $a(i)$ се интерпретира као мера која указује колико добро се тачка i уклапа у кластер (што је мања вредност ове мере то је кластеровање боље).
  - Нека је са $b(i)$ означено најмање просечно одстојање тачке i од сваког другог кластера коме не припада i. Тада се *Silhouette* мера израчунава:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

  Очигледно је да су вредности ове мере из интервала $[-1, 1]$. Кластеровање је добро за вредности ове мере блиске 1 тј. када је $a(i) \ll b(i)$. Што су веће вредности $b(i)$ кластеровање је лошије. Ако већина кластера има високе вредности *Silhouette* мере кластеровање је прикладно решење за одговарајући проблем. Пример графичке репрезентације ове мере за позитивне кластере алела HLA-A11:01 је дат на слици 4.3

  За све разматране алеле, и направљене кластере на позитивном и негативном скупу пептида се показало да је *Silhouette* мера веома добра (преко 95% инстанци има придружену меру већу од 0.5).



Слика 4.3: *Пример графичке репрезентације* Silhouette *мере за оцену квалитета кластера. Пример је дат за позитивне кластере алела HLA-A11:01.*

Прављење бинарних класификационих модела на основу добијених кластера је описано у следећим корацима:

(1) Сваки пептид је представљен изабраном физичко хемијском особином и Δ-BM25-IDF тежинама, и сваком пептиду из скупа за тренирање модела је придружена једна од шест група (кластера).

(2) Класификовање засновано на кластеровању, овде подразумева да ако је разматрани пептид ближи једном од три центроида позитивног скупа онда се сматра епитопом, ако је ближи неком од центроида из негативног скупа онда се сматра неепитопом.

(3) Кораци (1) и (2) се понављају за сваку физичко хемијску особину.

(4) За сваки модел (добијен у кораку (3)) се рачунају мере: прецизност, одзив, капа статистика и тачност. Све мере су провераване и на тестном скупу података.

(5) Бира се 10 модела са најбољим карактеристикама, и физичко хемијске особине које су коришћене у тим моделима се сматрају најбољим, у смислу да најбоље раздвајају епитопе од неепитопа. Табела са израчунатим најбољим физичко хемијским особинама је приложена уз тезу у додатку Б.1.

(6) Направљен је консензусни модел од модела добијених у кораку (5). Консензусни модел комбинује резултате k од 10 изабраних модела. Ако k

од 10 добијених модела за разматрани пептид предвиђа да јесте епитоп, онда је излаз из консензус модела 1 (тј. разматрани пептид јесте епитоп) у супротном је 0 (разматрани пептид није епитоп). Утврђивање броја модела k који учествују у консензусном моделу се добија тако што се узима минимална вредност k, таква да се максимизује тачност укупног модела.

Резултати добијених модела су приказани у табели 5.1. У овој тези су добијени модели послужили само за издвајање најбољих физичко хемијских особина и селекцију атрибута за моделе засноване на машинском учењу SVM и SVR. Иако је првобитна намена ових модела била само за издвајање најбољих физичко хемијских особина, показало се да су добијени модели довољно добре тачности и упоредиви са постојећим методама за предвиђање T - ћелијских епитопа. Резултати поређена са постојећим предикторима су приказани у табели 5.1. Модели су такође тестирани на још једном независном скупу података прикупљеном из MHCBN3.1 базе. Листа од првих 20 физичко хемијских особина које су се показала као најбоље за највећи број алела је дата у табели 5.2, а комплетна листа свих издвојених најбољих физичко хемијских особина за сваки појединачни алел је приложена у додатку Б.1.

Поглавље 5

Резултати примене предложених модела

У оквиру ове главе биће приказани резултати свих новоразвијених модела. Прво су приказани резултати нових класификационих модела заснованих на кластеровању k - срединама (поглавље 5.1), затим су приказани резлтати класификационих (SVM) и регресионих (SVR) модела заснованих на шемама 1, 2 и 3 редом (поглавље 5.2) и на крају резултати поређења развијених модела са постојећим методама за предвиђање Т - ћелијских епитопа (поглавље 5.3).

5.1 Резултати добијени кластер анализом и класификацијом заснованом на кластеровању

Модели бинарне класификације засноване на кластеровању су, као што је објашњено у претходној глави 4.4.1 где је и детаљно објашњен поступак прављења ових модела, направљени као помоћни модели за издвајање најбољих физичко хемијских особина и припрему улаза за SVM и SVR моделе засноване на шеми 2 представљања пептида. Показало се да ови модели самостално могу са довољно добром тачношћу да предвиде да ли је пептид епитоп или не.

У табели 5.1 су приказани резултати класификационих модела заснованих на кластеровању. Оцена квалитета модела је дата у мерама: тачност, прецизност, одзив, F - мера и капа статистика, које су израчунате на тренинг скупу података, коришћеном за груписање техником k-средина, и на тестном

(независном) скупу података. Да се епитопи заиста групишу око израчунатих центроида добијених техником k-средина потврђују резултати из табеле 5.1. Тачност модела на скупу података за тренирање је од 94 до 100%. Тачност ових модела је близу или изнад 80% на тестном (независном) скупу података, што потврђује да израчунати центроиди добро генерализују све епитопе и неепитопе за одговарајући алел. Модели су тестирани и на скупу података из MHCBN базе, који нема преклапања са подацима из IEDB базе. Добијени резултати тестирања модела су приказани у табели A.1 у додатку. Иако су добијени модели послужили само за избор физичко хемијских особина које најбоље раздвајају епитопе од неепитопа за појединачне алеле, довољно су добре тачности и упоредиви су са постојећим предикторима за предвиђање Т - ћелијских епитопа. У оквиру тезе је урађено поређене ових модела са два предиктора: NetMHCPan и MHCPred. Детаљни упоредни резултати тестирања модела са свим израчунатим мерама за оцену квалитета модела су приказани у табели A.2 у додатку, а графички упордени приказ добијених резултата је дат на слици[1] 5.1. Може се приметити да је тачност класификационих модела врло блиска тачности предиктора NetMHCPan, који важи за тренутно најтачнији предиктор. Ипак, треба напоменути да скуп података који је послужио за тестирање модела није независан за NetMHCPan предиктор и да су сви подаци из IEDB базе коришћени у тренирању модела овог предиктора. Тачност MHCPred предиктора је знатно лошија од тачности направљеног бинарног класификационог модела заснованог на кластеровању.

Све изабране физичко хемијске особине (по 10 најбољих за унограме и биграме за сваки алел) које су учествовале у прављену модела су приложене у додатку у табели Б.1. У табели 5.2 је дата листа са првих 20 физичко хемијских особина по важности у раздвајању епитопа од неептипа, добијена применом ових модела. Важност се мери по броју алела за које су се те физичко хемијске особине показале као одлучујуће у раздвајању.

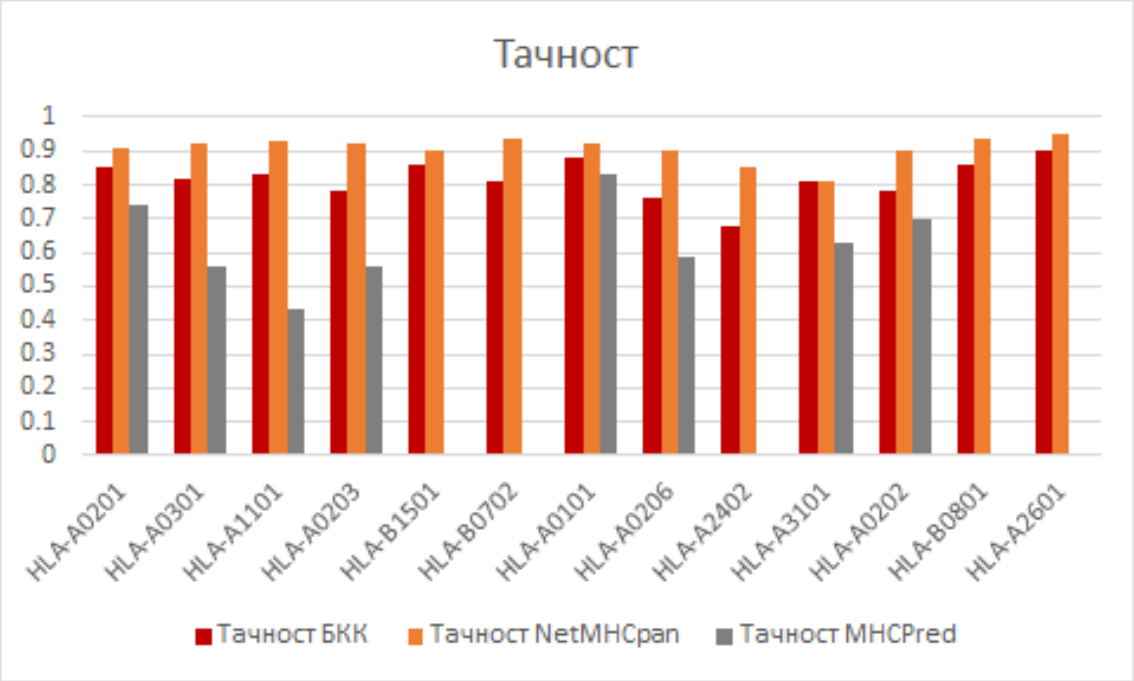
Корелација издвојених физичко хемијских особина и алела је приказана на слици 5.2. У графички приказ је укључено првих 6 физичко хемијских особина ради лакше анализе и тумачења. Може се видети да су за неке

[1]MHCPred предиктор не подржава предвиђања за све алеле разматране у раду, где нису приказани резултати за овај предиктор (сиви стубићи) значи да ти алели нису у скупу подржаних алела за предвиђање.

Алел	k	Скуп	Прецизност	Одзив	Капа	Тачност
HLA-A0201	8	Тренинг	0.93	0.95	0.88	0.94
		Тест	0.87	0.82	0.71	0.85
HLA-A0301	10	Тренинг	0.88	0.98	0.89	0.95
		Тест	0.76	0.70	0.60	0.82
HLA-A1101	10	Тренинг	0.93	0.97	0.90	0.95
		Тест	0.86	0.71	0.64	0.83
HLA-A0203	9	Тренинг	0.95	0.98	0.93	0.97
		Тест	0.80	0.75	0.57	0.78
HLA-B1501	10	Тренинг	0.95	0.98	0.94	0.97
		Тест	0.90	0.79	0.72	0.86
HLA-B0702	10	Тренинг	0.95	0.99	0.95	0.98
		Тест	0.86	0.60	0.60	0.81
HLA-A0101	10	Тренинг	0.93	0.997	0.96	0.97
		Тест	0.78	0.51	0.55	0.88
HLA-A0206	5	Тренинг	0.98	0.99	0.96	0.98
		Тест	0.79	0.84	0.48	0.76
HLA-A2402	5	Тренинг	0.98	0.98	0.95	0.98
		Тест	0.71	0.71	0.35	0.68
HLA-A3101	8	Тренинг	0.89	0.97	0.90	0.95
		Тест	0.72	0.61	0.53	0.81
HLA-A0202	3	Тренинг	0.99	0.995	0.97	0.99
		Тест	0.80	0.90	0.48	0.78
HLA-B0801	10	Тренинг	0.98	1	0.98	0.99
		Тест	0.84	0.74	0.68	0.86
HLA-A2601	10	Тренинг	0.87	1	0.92	0.98
		Тест	0.74	0.42	0.48	0.90
HLA-DRB10101	3	Тренинг	1	0.94	0.82	0.95
		Тест	0.88	0.92	0.19	0.82
HLA-DRB10401	3	Тренинг	1	0.99	0.99	0.996
		Тест	0.77	0.93	0.30	0.76

Табела 5.1: *Резултати бинарне класификације засноване на моделима добијеним техником груписања k - срединама над подацима IEDB базе.*

од различитих алела карактеристичне исте физичко хемијске особине. У случајевима где алели припадају истом супертипу то је и очекивано, и слаже се тврђењем да супертипови алела деле заједничке карактеристике које се могу искористити за предвиђање епитопа који се везују за специфичан алел а за које не постоје утврђени експериментални подаци [41][75]. Предиктори који су засновани на овом тврђењу су користили експерименталне податке, добијене за један алел, за предвиђање епитопа који се везују за други алел [41] како би повећали број експерименталних података за тренирање модела. Иако су резултати таквих модела знатно поправљени, оно што је мана у овом приступу је што није вођено рачуна које специфичне информације су заједничке за



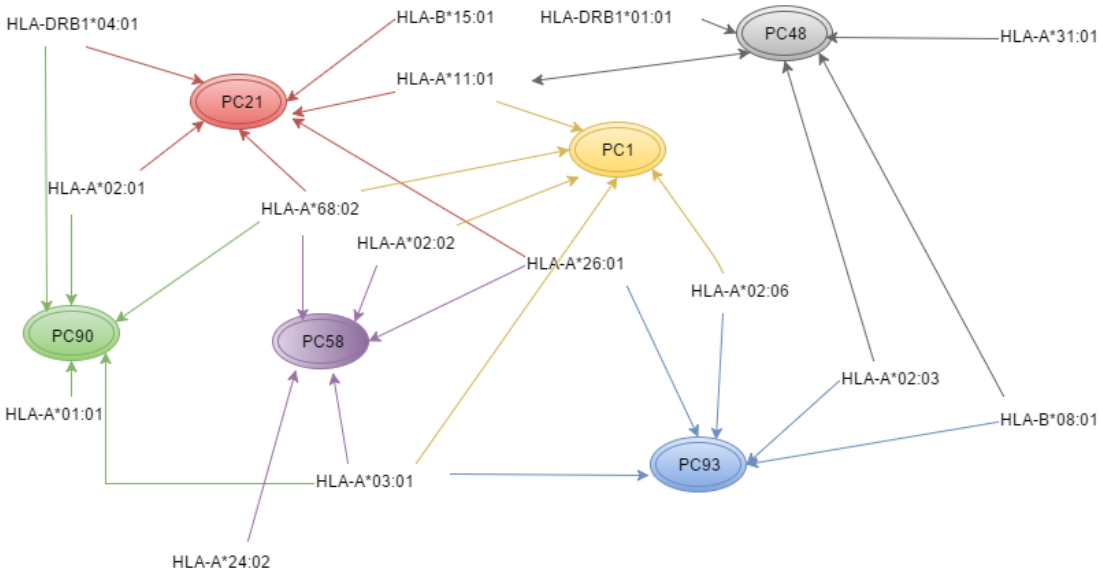
Слика 5.1: *Резултати бинарне класификације засноване на моделима добијеним техником k - средина (БКК) и предиктора NetMHCPan и MHCPred, изражени у мери тачност, на истом тестном скупу.*

конкретан супертип. Добијеним резултатима о најбољим физичко хемијским особинама је, међутим, утврђено да постоји и корелација између алела који не припадају истом супертипу (пример су алели HLA-A*02:03 и HLA-B*08:01 који припадају различитим супертиповима). Односно да специфични алели који припадају различитим супертиповима деле неке заједничке информације. Ови резултати се слажу са резултатима истраживања спроведеним у [34], где је утврђено да груба подела алела по супертиповима није добра, и да алели који припадају различитим супертиповима деле значајне информације. *Heckerman* и његови сарадници су сугерисали да би перформансе постојећих предиктора за специфичан алел могле знатно да се побољшају утврђивањем:

(1) карактеристика везујућих пептида које су заједничке за супертипове и укључивањем тих карактеристика као атрибута у постојеће моделе.

(2) карактеристика које су заједничке за све везујуће пептиде, независно од алела и супертипа коме припадају, и њиховим уврштавањем као атрибута у предикторе направљене за специфичан алел.

Треба напоменути да груписање алела по супертиповима није стриктно,



Слика 5.2: *Приказ шест најзначајних физичко хемијских особина и њихова корелација са алелима. Ознаке на графику одговарају физичко хемијским особинама: PC1 - alpha-NH chemical shifts; PC48 – Bulkiness; PC58 - Flexibility parameter for no rigid neighbor's; PC90 - Partition coefficient; PC21- pK-C; PC93 Average gain ratio in surrounding hydrophobicity (ФХ су преузете из [103])*

и постоји неколико различитих подела које су углавном оријентисана ка структурним сличностима *MHC* молекула [98].

Резултати добијени кластер анализом и класификацијом заснованом на груписању за избор најбољих физичко хемијских особина за специфичне алеле су укључени у направљене SVM и SVR моделе засноване на шеми 2 предтсвљања пептида. Детаљно су објашњени и приказани у поглављу 5.2.

ФХ	Алел	Алел	Алел	Алел	Алел	Алел	Алел
50	HLA-A*02:01	HLA-A*01:01	HLA-A*02:02	HLA-B*44:03			
53	HLA-A*02:01	HLA-B*15:01	HLA-A*01:01	HLA-A*26:01			
49	HLA-A*02:01	HLA-A*02:06	HLA-A*31:01	HLA-A*02:02	HLA-B*44:03		
6	HLA-A*02:01	HLA-A*03:01	HLA-A*02:06	HLA-B*44:03	HLA-DRB1*04:01		
89	HLA-A*02:01	HLA-A*03:01	HLA-A*68:02	HLA-A*01:01	HLA-DRB1*04:01		
20	HLA-A*02:01	HLA-A*11:01	HLA-A*68:02	HLA-B*15:01	HLA-B*44:03	HLA-A*26:01	HLA-DRB1*04:01
92	HLA-A*03:01	HLA-A*02:03	HLA-A*02:06	HLA-B*08:01	HLA-A*26:01		
17	HLA-A*03:01	HLA-A*02:03	HLA-A*01:01	HLA-B*08:01			
0	HLA-A*03:01	HLA-A*11:01	HLA-A*68:02	HLA-A*02:06	HLA-A*02:02		
57	HLA-A*03:01	HLA-A*68:02	HLA-B*07:02	HLA-A*24:02	HLA-A*02:02	HLA-A*26:01	
51	HLA-A*03:01	HLA-A*68:02	HLA-B*08:01	HLA-B*44:03			
112	HLA-A*11:01	HLA-A*68:02	HLA-B*15:01	HLA-A*26:01			
47	HLA-A*11:01	HLA-A*02:03	HLA-A*31:01	HLA-B*08:01	HLA-DRB1*01:01		
45	HLA-A*02:03	HLA-B*15:01	HLA-A*24:02	HLA-B*44:03	HLA-DRB1*01:01		
37	HLA-A*02:03	HLA-B*15:01	HLA-B*07:02	HLA-A*01:01	HLA-A*24:02		
13	HLA-A*02:03	HLA-A*68:02	HLA-B*15:01	HLA-B*07:02	HLA-A*02:02		
14	HLA-A*68:02	HLA-B*15:01	HLA-B*44:03	HLA-DRB1*04:01			
91	HLA-B*15:01	HLA-B*07:02	HLA-A*31:01	HLA-B*08:01			
24	HLA-B*07:02	HLA-A*02:06	HLA-A*24:02	HLA-A*31:01	HLA-A*26:01		
33	HLA-A*02:06	HLA-B*08:01	HLA-B*44:03	HLA-DRB1*01:01			

Табела 5.2: *Првих 20 физичко хемијских особина, по броју алела за које су најважније у раздвајању епитопа од неепитопа; ФХ је редни број физичко хемијске особине из табеле [103]. Нумерисање почиње од 0, тако да редни број 0 у ствари представља прву физичко хемијску особину, итд.*

5.2 Резултати SVM и SVR модела

У оквиру овог дела су приказани резултати направљених модела заснованих на техници подржавајућих вектора.

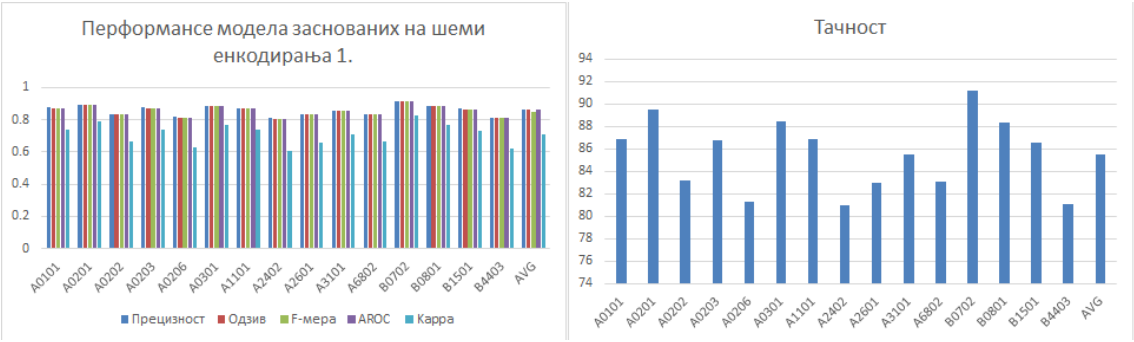
Резултати свих добијених класификационих и регресионих модела где се као припрема података за улаз у модел користи прва шема енкодирања пептида, и пептид се представља као вектор димензије 180 су дати у табели 5.3. Резултати су приказани за сваки направљени модел (за сваки алел су направљени засебни модели). Упоредо су дати резултати за моделе бинарне класификације и регресионе моделе. Мере у којима су приказане карактеристике бинарних модела су: тачност, прецизност, одзив, F-мера, $AROC$ и капа статистика. За регресионе моделе је као мера оцене квалитета модела коришћен Пирсонов коефицијент корелације предвиђених афинитета везивања пептида и стварне, експериментално узете мере из IEDB базе. Приказани су резултати унакрсне провере (10-уп) и резултати добијени применом модела на тестном скупу података (eng. *blind test*).

	SVM							SVR	
	Тачност		Карактеристике на тестном скупу					КК	
Алел	10-уп	Тест скуп	Прецизност	Одзив	F-мера	$AROC$	Капа	10-уп	Тест скуп
A0101	87.17	86.91	0.87	0.87	0.87	0.87	0.74	0.80	0.78
A0201	88.92	89.48	0.90	0.90	0.90	0.90	0.79	0.79	0.80
A0202	79.62	83.21	0.83	0.83	0.83	0.83	0.66	0.79	0.75
A0203	86.75	86.76	0.88	0.87	0.87	0.87	0.74	0.79	0.81
A0206	79.75	81.28	0.82	0.83	0.81	0.81	0.63	0.67	0.72
A0301	86.86	88.49	0.90	0.90	0.90	0.89	0.77	0.76	0.78
A1101	85.85	86.93	0.87	0.87	0.87	0.87	0.74	0.76	0.77
A2402	80.22	81.02	0.81	0.80	0.80	0.80	0.61	0.64	0.63
A2601	87.70	83.01	0.83	0.83	0.83	0.83	0.66	0.77	0.68
A3101	85.09	85.56	0.86	0.86	0.86	0.86	0.71	0.77	0.79
A6802	80.50	83.11	0.83	0.83	0.83	0.83	0.66	0.70	0.71
B0702	91.53	91.24	0.91	0.91	0.91	0.91	0.83	0.79	0.76
B0801	89.43	88.32	0.88	0.88	0.88	0.88	0.77	0.74	0.73
B1501	81.33	86.61	0.87	0.87	0.87	0.87	0.73	0.70	0.69
B4403	80.57	81.13	0.81	0.81	0.81	0.81	0.62	0.71	0.72
AVG	84.75	85.54	0.86	0.86	0.85	0.86	0.71	0.75	0.74

Табела 5.3: *Резултати класификационих модела заснованих на шеми 1 представљана пептида. 10 - унакрсна провера је рађена на скупу података за тренирање, остале карактеристике су дате на тестном независном скупу података. КК представља коефицијент корелације за регресионе моделе.*

Тачност бинарних класификационих модела на тестном скупу података

иде од 81 до 91.3%, а просечна вредност мере тачности за све моделе је око 85%. Коефицијент корелације за регресионе моделе је у интервалу 0.63-0.81, а просечна вредност коефицијента корелације за све моделе је око 0.75. Резултати потврђују да је уведена мера за рачунање учесталости амино киселина веома ефикасна и добар избор за представљање пептида у комбинацији са BLOSUM (VOGG) матрицом. Резултати регресионих модела показају да постоји веома висок степен корелације предвиђених афинитета са стварним вредностима афинитета везивања пептида за одговарајуће молекуле *MHC* класа. Графички приказ карактеристика бинарних модела по алелима на тестном скупу је дат на слици 5.3.



Слика 5.3: *Резултати класификационих модела заснованих на шеми 1 енкодирања пептида, добијени применом на тестни скуп података. Лево: све укључене мере за оцену квалитета класификационих модела по алелима; Десно: тачност класификационих модела за сваки алел*

Са слике 5.3 (лево) се јасно може видети да су све мере за оцену квалитета свих направљених модела веома добре. Слика 5.3 (десно) је приказ мере тачности направљених модела за сваки од 15 алела који су укључени у истраживање. Како је у овом моделу балансиран број епитопа и неепитопа сама мера тачност је довољна за оцену квалитета модела и она је изнад 80% за све моделе.

Резултати добијени шемом 2 енкодирања пептида су дати у табели 5.4. Овај модел је заснован на представљану улазних података (пептида) у виду вектора димензије 170. Компоненте вектора се добијају избором и применом 10 најбољих физичко хемијских особина, добијених применом поступка описаном у 4.2.6, по алелу, за униграме и биграме и рачунању учесталости униграма и биграма у пептиду $\Delta - BM25$ техником.

Тачност бинарних модела иде од 81 до 95%, а просечна вредност тачности свих модела је око 87%. Коефицијент корелације на тестном скупу података иде од 0.71 до 0.85, а просечна вредност коефицијента корелације је око 0.8.

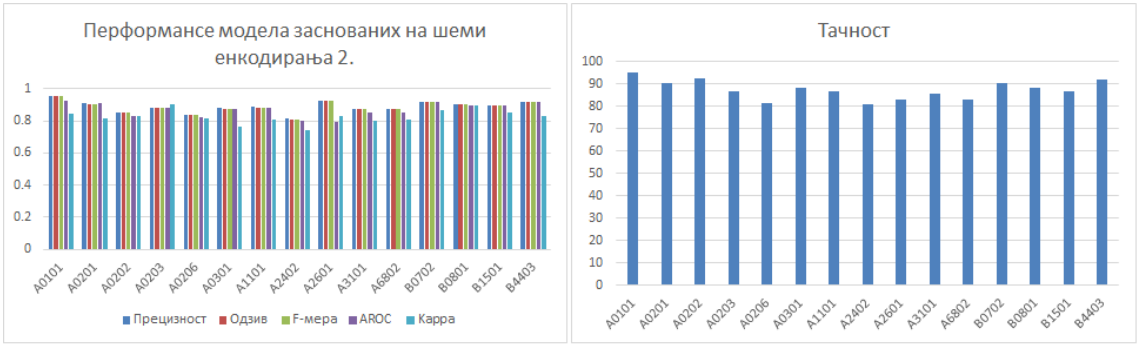
	SVM							SVR	
	Тачност		Карактеристике на тестном скупу					КК	
Алел	10-уп	Тест скуп	Прецизност	Одзив	*F*-мера	*AROC*	Капа	10-уп	Тест скуп
A0101	95.39	95.47	0.96	0.96	0.96	0.92	0.85	0.85	0.85
A0201	91.42	90.58	0.91	0.91	0.91	0.91	0.81	0.84	0.85
A0202	93.81	92.53	0.85	0.85	0.85	0.83	0.83	0.81	0.84
A0203	86.75	87.05	0.88	0.88	0.88	0.88	0.91	0.84	0.85
A0206	79.75	81.28	0.84	0.84	0.84	0.82	0.81	0.82	0.82
A0301	86.86	88.48	0.88	0.87	0.87	0.87	0.77	0.76	0.78
A1101	85.85	86.93	0.89	0.88	0.88	0.88	0.81	0.82	0.80
A2402	80.22	81.02	0.81	0.81	0.81	0.80	0.74	0.75	0.72
A2601	86.88	83.01	0.92	0.93	0.92	0.80	0.83	0.77	0.68
A3101	85.09	85.55	0.87	0.87	0.87	0.85	0.80	0.77	0.79
A6802	83.56	83.10	0.88	0.88	0.88	0.85	0.81	0.70	0.71
B0702	91.53	90.30	0.92	0.91	0.92	0.92	0.86	0.79	0.85
B0801	89.43	88.32	0.91	0.90	0.90	0.89	0.90	0.84	0.84
B1501	89.10	86.60	0.90	0.89	0.89	0.90	0.85	0.80	0.80
B4403	93.35	91.87	0.92	0.92	0.92	0.92	0.83	0.85	0.84
AVG	87.93	87.47	0.89	0.87	0.89	0.87	0.83	0.80	0.80

Табела 5.4: *Резултати класификационих модела заснованих на шеми 2 представљања пептида. 10-УК представља резултате 10 - унакрсне провере за класификационе и регресионе моделе на скупу података за тренирање. КК - се односи на коефицијент корелације добијен за регресионе моделе. Карактеристике на тестном скупу представљају резултате тестирања класификационих и регресионих модела на независном скупу података.*

Према просечним вредностима укључених мера за оцену квалитета модела се модели засновани на овој шеми представљања података показују боље него на првој шеми представљања улазних података. Регресиони модели засновани на другој шеми енкодирања су знатно бољи него у првом случају, што потврђује да изабране физичко хемијске особине имају велики утицај на афинитет везивања пептида. На слици 5.4 су приказане карактеристике бинарних класификационих модела заснованих на шеми 2 представљања пептида.

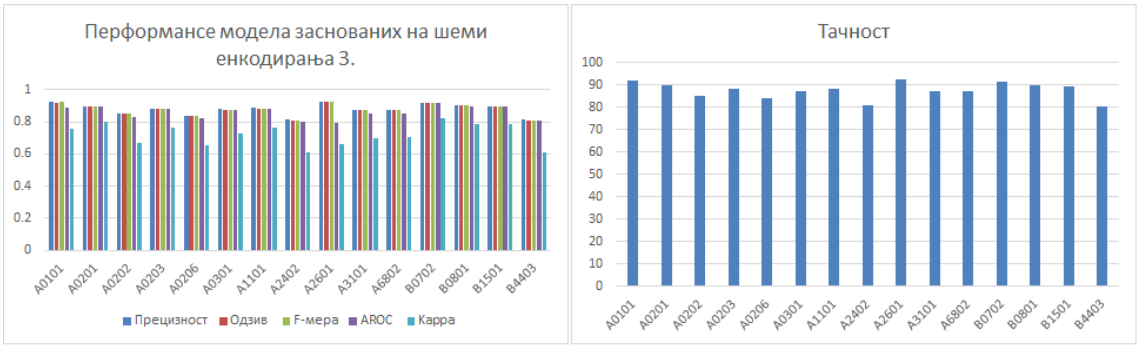
У табели 5.5 су приказани резултати добијени шемом 3 енкодирања података, где је улазни пептид представљен вектором димензије 225. Компоненте вектора којим се представља пептид су комбинација $\Delta - BM25$ тежина амино киселина, компоненти VOGG матрице и изабраних z дескриптора. Шема 3 енкодирања пептида заправо представља проширење шеме 1 са z дескрипторима.

Тачност добијених бинарних модела иде од 80 до 92.11% а просечна тачност модела је око 88%, док се коефицијент корелације креће од 0.68 до 0.83 а



Слика 5.4: *Карактеристике модела заснованих на шеми 2 енкодирања пептида, добијене применом на тестном скупу података. Лево: мере за оцену перформанси модела по алелима; Десно: тачност модела за сваки алел*

просечна вредност коефицијента корелације износи 0.78. На слици 5.5 је дат графички приказ карактеристика модела заснованих на шеми 3.



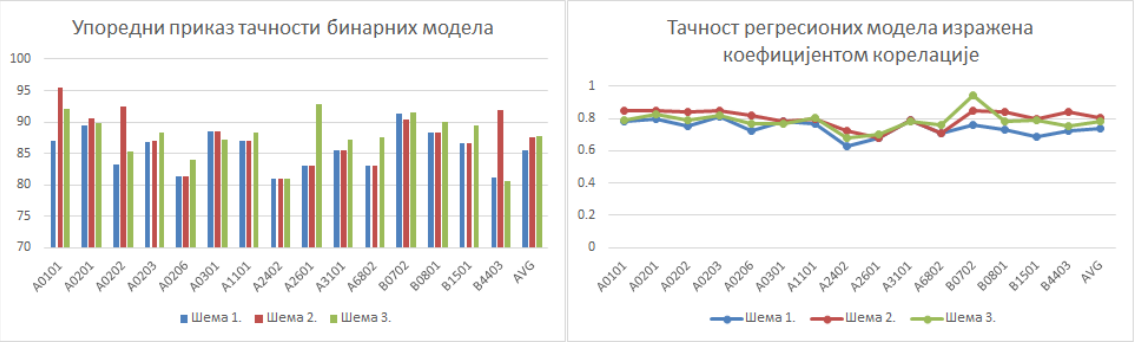
Слика 5.5: *Карактеристике модела заснованих на шеми 3 енкодирања пептида, добијене применом на тестном скупу података. Лево: мере за оцену перформанси модела по алелима; Десно: тачност модела за сваки алел*

Из приложених таблица са резултатима се може видети да су све мере за оцену квалитета модела одличне и да су сви модели добро условљени. Упоредни приказ резултата нових модела, добијених на тестном скупу података је приказан на слици 5.6. Лево је приказана упоредна тачност бинарних модела, а на десно коефицијент корелације за сва три регресиона модела. Приказани резултати на слици 5.6 су добијени тестирањем модела на независном скупу података који нема преклапања са скупом података за тренирање модела.

Може се приметити да је бинарни модел заснован на шеми 2 представљања улазних података готово у свим случајевима бољи од друга два модела (заснована на шеми 1 и шеми 3), док се сви регресиони модели у великој мери поклапају и тешко је проценити која шема најбоље генерализује на тестном скупу података. Упоредним приказом само мере тачност за бинарне моделе (слика 5.7) се може закључити да за неке алеле постоји знатно одступање у

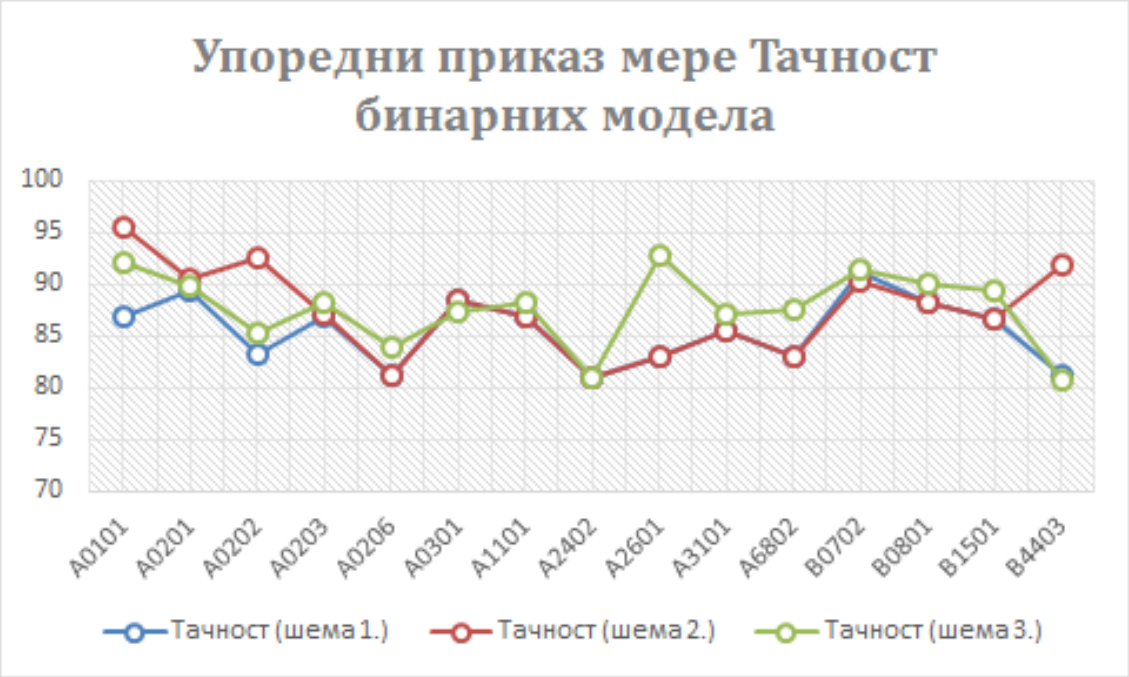
	SVM							SVR	
	Тачност		Карактеристике на тестном скупу					КК	
Алел	10-уп	Тест скуп	Прецизност	Одзив	*F*-мера	*AROC*	Капа	10-уп	Тест скуп
A0101	89.52	92.11	0.92	0.92	0.92	0.90	0.75	0.75	0.79
A0201	89.86	89.85	0.90	0.90	0.90	0.90	0.80	0.82	0.83
A0202	86.87	85.32	0.85	0.85	0.85	0.83	0.70	0.78	0.79
A0203	89.20	88.29	0.88	0.88	0.88	0.88	0.77	0.81	0.82
A0206	84.47	83.92	0.84	0.84	0.84	0.82	0.65	0.76	0.77
A0301	88.57	87.21	0.88	0.87	0.87	0.87	0.72	0.79	0.77
A1101	87.98	88.31	0.89	0.88	0.88	0.88	0.76	0.81	0.81
A2402	82.95	80.89	0.81	0.81	0.81	0.80	0.61	0.68	0.68
A2601	92.88	92.78	0.92	0.93	0.92	0.80	0.66	0.70	0.70
A3101	87.96	87.11	0.87	0.87	0.87	0.85	0.70	0.81	0.78
A6802	88.29	87.57	0.88	0.88	0.88	0.85	0.70	0.80	0.76
B0702	90.41	91.43	0.92	0.91	0.92	0.92	0.82	0.79	0.94
B0801	90.71	90.04	0.90	0.90	0.90	0.89	0.78	0.80	0.78
B1501	90.05	89.36	0.90	0.89	0.89	0.90	0.79	0.77	0.79
B4403	82.81	80.65	0.81	0.81	0.81	0.81	0.61	0.77	0.75
AVG	88.17	87.66	0.88	0.88	0.88	0.86	0.72	0.78	0.78

Табела 5.5: *Резултати класификационих модела заснованих на шеми 3 представљана пептида. 10-УК представља резултате добијене 10 - унакрсном провером на скупу података за тренирање за класификационе и регресионе моделе. Резултати регресионих модела су изражени коефицијентом корелације КК. Карактеристике на тестном скупу представљају резултате тестирања модела на независном скупу података.*



Слика 5.6: *Упоредни приказ карактеристика свих модела заснованих на шеми 1, шеми 2 и шеми 3 Лево: тачност бинарних модела класификације, Десно: тачност регресионих модела изражена коефицијентом корелације.*

тачности ова три модела. Односно да би комбинација добијених модела за предвиђање Т - ћелијских епитопа могла да резултује моделом са још бољим карактеристикама и тачношћу предвиђања Т - ћелијских епитопа.



Слика 5.7: *Упоредни приказ тачности бинарних модела заснованих на шеми 1, шеми 2 и шеми 3*

5.3 Поређење дефинисаних модела са постојећим методама за предвиђање Т - ћелијских епитопа

Поређење различитих система (предиктора, алата, модела) за предвиђање Т - ћелијских епитопа је доста компликовано због природе рада постојећих алата. Већина алата је доступна само преко веб сервера који могу прихватити као улаз само комплетну протеинску секвенцу, а као излаз се добијају пептиди за које је предвиђено да су потенцијални епитопи. Излазним резултатима су придружене нестандардне мере афинитета везивања за *MHC* молекуле, које се даље не могу упоређивати без превођења. Други вид излаза су сви пептиди из протеинске секвенце који се издвајају клизно као узастопни нонамери. Дакле већина постојећих предиктора не прима као улаз пептид(е), а једини експериментални подаци су доступни само у виду пептида различитих дужина. Изузетак је NetMHC(pan) предиктор који као улаз може да прими и саме пептиде и да оцену афинитета везивања у мери IC_{50} као и скалирану вредност $1 - log_{50k}(Aff_{IC_{50}})$. Како су и NetMHC(pan) и SMM^{PMBEC} предиктори тренирани над подацима из IEDB базе, и подаци, на основу којих су модели који су у основи ових предиктора направљени, су јавно доступни то је прво поређење

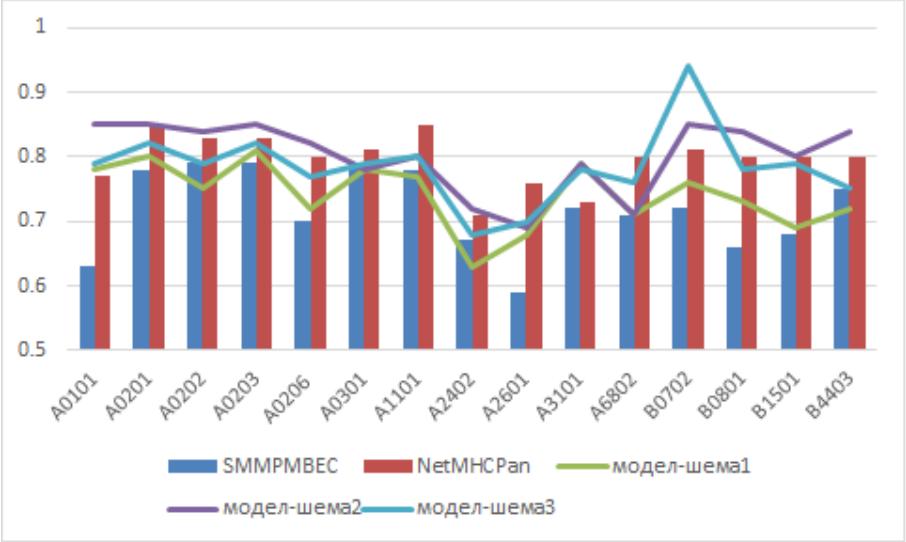
урађено управо са овим предикторима. Ова два предиктора су оцењивана у многим истраживањима и упоређивана са другим постојећим алатима и показали су се као најтачнији и најпоузданији [52][118]. Оба предиктора се редовно побољшавају тренирањем на новим подацима из IEDB базе, па треба нагласити да добијене оцене карактеристика ова два предиктора нису оцењене на непознатом скупу, већ скупу података који је вероватно учествовао у прављену самих модела. Табела 5.6 садржи упоредни приказ резултата ова два предиктора и резултата добијених новим моделима направљеним у оквиру ове тезе. У табели су подебљани резултати нових модела који су бољи од резултата оба предиктора NetMHC(pan) и SMM^{PMBEC}. Подвучени су резултати добијени новим моделима који су бољи од резултата бар једног од ова два предиктора. Закључак је да готово у свим случајевима се један од нових модела показао боље

Алел	SMMPMBEC	NetMHCPan	Шема 1	Шема 2	Шема 3
A0101	0.66- 0.63	0.82 -0.77	0.80-0.78	**0.85-0.85**	0.75-0.79
A0201	0.80-0.78	0.869 -0.85	0.79-0.80	0.84-0.85	0.82-0.82
A0202	0.79 -0.80	0.83-0.84	0.79-0.75	0.81-0.84	0.78-0.79
A0203	0.83-0.79	0.86-0.88	0.79-0.81	0.84-0.85	0.81-0.82
A0206	0.75 -0.70	0.80- 0.74	0.67-0.72	**0.82-0.82**	0.76-0.77
A0301	0.79- 0.79	0.81- 0.83	0.76-0.78	0.76-0.78	0.79-0.77
A1101	0.78- 0.78	0.86-0.85	0.76-0.77	0.82-0.80	0.81-0.80
A2402	0.67- 0.74	0.71-0.75	0.64-0.63	0.75-0.72	0.68-0.68
A2601	0.62-0.59	0.78-0.76	0.77-0.68	0.77-0.68	0.70-0.70
A3101	0.77- 0.72	0.83-0.73	0.77-0.79	0.77-0.79	0.81-0.78
A6802	0.71-0.81	0.80-0.88	0.70-0.71	0.70-0.71	0.80-0.76
B0702	0.72- 0.78	0.84- 0.81	0.79-0.76	0.79-0.85	0.79-0.94
B0801	0.66- 0.81	0.76-0.89	0.74-0.73	**0.84-0.84**	**0.80-0.78**
B1501	0.68- 0.74	0.74- 0.809	0.70-0.69	**0.80-0.80**	0.77-0.79
B4403	0.75-0.81	0.80-0.83	0.71-0.72	**0.85-0.84**	0.77-0.75

Табела 5.6: *Резултати предиктора SMMPMBEC, NetMHCpan и нових регресионих модела. Резултати су добијени 10-fcv (лева вредност у табели) и на тест скупу (десна вредност у табели)*

од бар једног а негде и оба постојећа предиктора. Резултати су упоређивани на основу добијеног Пирсоновог коефицијента корелације између предвиђених и стварних вредности афинитета везивања пептида. Графички приказ поређења резултата тестирања постојећих модела и нових модела је дат на слици 5.8 где су упоређивани само на тестном скупу података.

У оквиру IEDB организације се ради упоређивање предиктора који припадају IEDB алатима. Добијени резултати се налазе на локацији `http://tools.immuneepitope.org/analyze/html/mhc_binding.html`. Алати укључују неколико различитих предиктора ANN, ARB, SMM^{PMBEC}, PickPocket и



Слика 5.8: *Упоредни приказ резултата регресионих модела заснованих на шеми 1, шеми 2 и шеми 3 са постојећим предикторима: NetMHCpan (црвени стубићи) и SMM^{PMBEC} (плави стубићи).*

консензус методу (добијену комбинацијом набројаних предиктора)[2]. При сваком ажурирању IEDB базе ради се провера тачности наведених предиктора, тако што се прво примене на нове унете експерименталне податке и упореде се међусобно резултати добијени различитим предикторима. Ти примери су у тези послужили као додатан прави "слепи" тест, за поређене више предиктора[3]. У табели 5.7 су приказани резултати осам различитих предиктора као и резултати једног од нових класификационих и једног регресионог модела (модели заснован на шеми 3 представљања пептида). Последње три колоне у табели представљају резултате новог модела изражене у мерама афинитета IC_{50}, скалираној вредности афинитета $1 - log_{50k}(Aff)$ као и бинарна одлука (јесте епитоп - вредност 1, није епитоп - вредност 0). Све три мере су укључене ради лакшег поређења са резултатима других предиктора. Друга колона у табели представља експериментално утврђену вредност, односно да ли се пептид везује за молекул *MHC* класе HLA-B07:02 (1 - везује се, 0 не везује се).

Резултати новог бинарног модела на приказаном скупу пептида су тачност од 100%, у смислу да је сваки пептид за који је експериментално утврђено да се везује за молекул класе HLA-B07:02 и модел дефинисан у раду предвидео да се везује. Тврђење важи и за пептиде за које је експериментално утврђено да се не везују за молекул класе HLA-B07:02, модел је такође предвидео да се ти пептиди

[2]Сви предиктори су детаљно описани у глави 3.2

[3]Резултати приказани у табели 5.6 представљају резултате поређена на великом тестном скупу IEDB базе. Овде су додати резултати поређена на малом тестном скупу који је у моменту спровођења експеримената једино био доступан.

не везују. У прилог томе говори и предвиђени афинитет везивања, резултат регресионог модела, који је за епитопе испод границе (прага везивања), а за неепитопе изнад подразумеване границе. Подразумевана граница за афинитет везивања код епитопа је испод $500 IC_{50}$, док у случају да је вредност афинитета везивања пептида изнад $500 IC_{50}$ пептид је неепитоп[4]. Сви предвиђени епитопи добијени бинарним моделом имају предвиђен афинитет везивања, добијен као резултат регресионих модела, у складу са наведеним границама. Експериментални подаци су били доступни и за алел HLA-A02:01. Резултати постојећих и нових модела су дати у табели 5.8. И у случају овог алела, експериментални подаци су дати у виду бинарне одлуке (1 - јесте епитоп, 0 - није епитоп).

Резултати добијени за алел HLA-A02:01 показују да је бинарни модел погрешио само у једном случају. Регресиони модел такође предвиђа висок афинитет везивања за тај пептид, што значи да је сагласан у предвиђању пептида као неепитопа. У табели је слог са погрешно класификованим пептидом уоквирен. У осталим случајевима модел тачно класификује епитопе и неепитопе. Анализом резултата нових предиктора (последње три колоне у табели) и резултата других предиктора за пептид који модел погрешно класификује се види да су сви предиктори сагласни око класификовања тог пептида као неепитопа. Сви предиктори предвиђају веома висок афинитет везивања што значи да тај пептид не би био препознат као епитоп ни у случају других постојећих предиктора. Највиши коефицијент корелације предвиђених афинитета везивања за пептиде у табели 5.8 имају модел, дефинисан у раду, и NetMHCPan предиктор, и он износи 0.96.

[4]Афинитет везивања пептида се најчешће представља мером IC_{50} која је изражена у јединици nM. Јаче везивање пептида подрзумева ниже вредности мере IC_{50}, и обрнуто, слабије везивање пептида подрзумева јако високе вредности мере IC_{50}. У радовима се често јако везивање пептида за молекуле *MHC* класа објашњава са јаким афинитетом везивања пептида, што може довести до погрешног разумевања ове мере јер то управо значи да мера IC_{50} има јако мале вредности. Детаљније објашњење се може наћи у [95]

Пептид	ЕУМ	NetMHCpan	SMM	ANN	ARB	SMMPMBEC	IEDB Consensus	NetMHCcons	PickPocket	Афинитет	1-log50k(Aff)	БМ
HPRQEQIAL	1	4.63	8.37	9	0.79	6.86	0.2	6.71	30.54	6.237442307	0.618	1
AAGIGILTV	0	18382.9	5886.95	15199	3553.64	7114.76	13	16763.86	5382.48	1335.659212	0.122	0
NPATPASKL	1	55.26	96.36	24	10.34	113.02	0.6	37.11	79.14	15.81677592	0.532	1
TPRVTGGAM	1	3.5	2.79	8	0.94	2.76	0.1	5.35	15.61	0.354633281	0.883	1
SPSLRILAI	1	13.72	31.4	15	15.5	31.13	0.3	14.48	103.72	24.91579232	0.49	1
LPQKKSNAL	1	6.59	18.36	9	2.56	13.59	0.2	7.98	71.02	5.845379628	0.624	1
FPALRFVEV	1	32.34	88.69	28	23.17	81.31	0.6	30.21	119.38	33.01018227	0.464	1
APAGVREVM	1	14.34	15.84	11	2.6	16.68	0.3	12.78	57.2	15.6465644	0.533	1
SPASSRTDL	1	6	10.54	9	2.05	9.17	0.2	7.44	51.89	2.988660484	0.686	1
KPQQKGLRL	1	10.82	43.94	12	17.84	34.21	0.3	11.66	165.16	42.7979732	0.44	1
NLVPMVATV	0	22795.96	25697.45	17074	4340.25	25012.67	23	19717.81	5382.48	438.2296453	0.225	0
LPQQPPLSL	1	9.87	23.65	22	1.36	22.19	0.2	14.79	63.05	12.19951827	0.556	1

Табела 5.7: *Резултати предиктора из IEDB алата за алел HLA-B07:02, и поређење са новим моделом заснованим на шеми 3 енкодирања пептида. ЕУМ представља експериментално утврђено мерење; БМ представља резултат предвиђања бинарног модела; Афинитет је резултат предвиђања регресионог модела.*

Пептид	ЕУМ	NetMHCpan	SMM	ANN	ARB	SMMPMBEC	IEDB Consensus	NetMHCcons	PickPocket	Афинитет	1-log50k (Aff)	БМ
AAGIGILTV	1	3448.53	1429.26	2498	672.28	1460.26	8.9	2936.57	912.75	1062.03	0.356	1
RLLEAIIRL	1	4.17	19.82	7	16.76	19.12	0.4	5.59	10.35	8.705506	0.8	1
FLSSANEHL	1	14.94	60.83	19	9.45	63.89	0.8	17.03	52.46	14.32012	0.754	1
SLQEKVAKA	1	250.12	230.73	73	59.3	242.9	3	135.2	90.11	316.1201	0.468	1
RPRAPTEEL	1	33787.75	190594.3	27646	243846.73	201571.9	56	30560.95	50000	8208.088	0.167	0
AMLERQFTV	1	3.24	7.52	5	1.4	7.7	0.3	4.04	5.35	7.087835	0.819	1
VLQNVAFSV	1	10.56	18.24	10	1.51	18.43	0.5	10.58	11.29	4.499423	0.861	1
LPDGGVRLL	0	26255.57	14458.06	18835	8337.44	15290.79	28	22209.93	9654.43	15208.39	0.11	0
ALAPAPAEV	1	9.96	24.38	19	5.13	25.43	0.8	13.86	24.87	51.33636	0.636	1
TPRVTGGAM	0	40472.64	6809418	31035	1000000	7053911.39	95	35367.39	50000	45854.11	0.008	0
SLAAYIPRL	1	4.28	15.78	4	1.07	16.27	0.4	4.55	16.31	5.013564	0.851	1
MMYKDILLL	1	8.52	13.19	9	3.83	13.44	0.4	8.85	11.53	5.586454	0.841	1
NLVPMVATV	1	21.29	66.39	29	6.9	63.3	1	25.14	16.66	2.647935	0.91	1

Табела 5.8: *Резултати предиктора из IEDB алата за алел HLA-A02:01, и новог модела заснованог на шеми 3 репрезентовања пептида. ЕУМ представља експериментално утврђено мерење; БМ предствља резултат предвиђања бинарног модела; Афинитет је резултат предвиђања регресионог модела.*

Поглавље 6

Неуређена структура протеина и Т - ћелијски епитопи

У оквиру овог дела тезе су изложени сумарни резултати исцрпног истраживања везе Т - ћелијских епитопа, структуре протеина и хидропатије протеинских региона. Циљ истраживања је био да се установе значајна правила и обрасци код Т - ћелијских епитопа који би послужили за лакше идентификовање епитопа и прављење нових модела за предвиђање. Ово истраживање је претходило истраживању спроведеном у првом делу тезе, а резултати који су овде добијени су били мотивација за нове шеме енкодирања пептида приказане у глави 4. Детаљна анализа је спроведена над великим скупом података добијеним применом постојећих алата за предвиђање: Т - ћелијских епитопа, уређених/неуређених структура протеина и хидропатије протеинских региона. Развијен је нови софтверски систем који је омогућио припрему и обраду добијених података. Примењена је и техника правила придруживања како би се издвојила интересантна правила везана за протеинску секвенцу, структуру, хидропатију и Т - ћелијске епитопе.

6.1 Структура протеина

Редослед амино киселина у протеину одређује просторну структуру протеина, а од просторне структуре протеина директно зависи функција протеина. Постоје четири нивоа структуре протеина:

1. **Примарна структура** представља редослед амино киселина у полипептидном ланцу (секвенца узастопних амино киселина).

2. **Секундарна структура** представља локалну просторну организацију

(конформацију) атома полипептидне кичме која је дефинисана водоничним везама између амино и карбоксилне групе у секвенци амино киселина у полипептиду (при чему се природа везе бочних остатака амино киселина и њихове конформације не узима у обзир). Торзиони углови φ и ψ Рамахандрановог дијаграма (eng. *Ramachandran φ and ψ dihedral torsion angles*) између α - C атома и C атома у COOH групи и N атома у NH_2 групи одређује секундарну структуру протеина.

3. **Терцијарна структура** је тродимензионална структура читавог полипептидног ланца.

4. **Кватернарна структура** је просторни распоред више полипептида (подјединица) које чине протеин.

Примарну структура протеина чине његова јединствена амино киселинска секвенца и распоред дисулфидних мостова. Број и распоред амино киселина варира од протеина до протеина. Директна информација о распореду је садржана у генима. И најмања промена у примарној структури може значајно да утиче на укупну структуру и функционисање протеина. Секундарна структура је конформација полипептидног ланца заснована на водоничним везама. Основни облици који се подразумевају под секундарном структуром су α - хеликс, β - набрана структура, и β завој. Секундарна структура протеина није непроменљива, те су могуће конформационе промене везане за функционисање протеина, промене у околини. Терцијарна структура одређује распоред подјединица и заснована је на низу различитих интеракција. Реч је о интеракцијама између удаљених делова полипептидног ланца примарне структуре. Кватернарна структура је просторни распоред полипептида у протеинима који имају више подјединица.

6.1.1 Уређена и неуређена структура протеина

Многи протеински региони или неки цели протеини немају дефинисану 3Д структуру, као што показују експериментални подаци добијени у *in vitro* условима. Они показују различите конформационе изомере у којима се позиције атома и торзионих углова полипептидне кичме мењају у току времена. Постојећи називи ових протеина обухватају више израза као што су "урођена неуређена, неувијена, денатурисана" структура, али су ипак најчешће у употреби "суштински неуређени, неувијени, неструктуирани" протеини (eng. *intrinsically disordered, unfolded, unstructured proteins*). У оквиру ове тезе

се користе искључиво термини уређени, односно, неуређени протеини. Они могу бити потпуно уређени или неуређени или се састоје од уређених и неуређених региона. Неуређени региони се експериментално идентификују на основу 3Д структуре протеина. Традиционално, идентификовање 3Д структуре се изводи експерименталним методама, од који су најзначајније: Дифракциона кристалографија X - зрацима и Нуклеарно магнетна резонантна спектроскопија (NMR). Експерименталне методе су временски веома захтевне и имају низ других ограничења. До данас је развијено преко 20 биофизичких и биохемијских метода за одређивање неуређених делова протеина. Такође је развијен велики број програма за предвиђање неуређене структуре. Програми за предвиђање неуређених структура се деле у две основне групе:

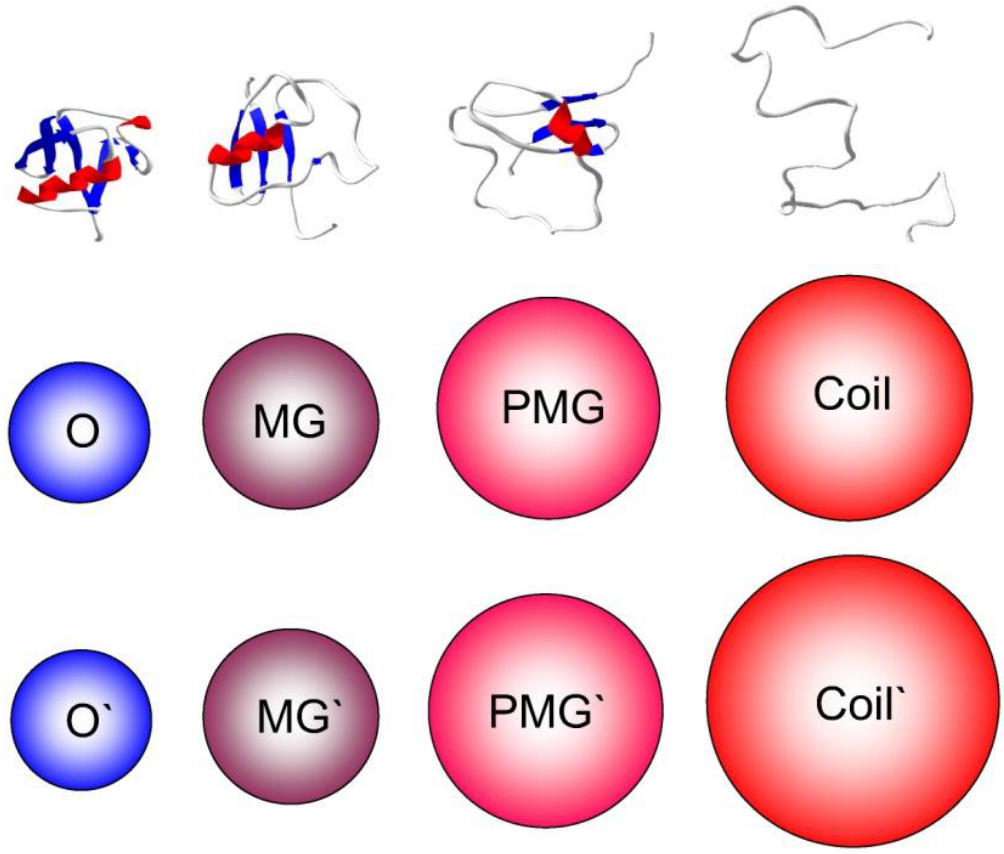
1. програме засноване на физичко хемијским особинама амино киселина у протеину (неки програми из ове групе су: PONDR, FoldUnFold, PreLINK, IUPred, GlobProt, FoldIndex),

2. програме засноване на методама "поравнања" (eng. *alignement*) хомологих протеинских секвенци (програми из ове групе су RONN, DISOPRED).

Новији програми за предвиђање неуређених региона углавном комбинују оба приступа. На основу експерименталних података и предвиђања неуређени региони се могу поделити на 3-5 група: (а) кратке: 1–3, 4–15, 16–30, (б) дуге: 30–100 и 100-200 и (ц) веома дуге: >200 амино киселина [84]. Разлике у облику и структури неуређених протеина су такође велике. Најнеуређенија структура је насумично клупко (eng. *random coil*), које одговара највише развијеном стању глобуларних протеина, пре-топљива глобула (eng. *pre-molten globule*) је издужена, делимично структуирана форма, топљива глобула (eng. *molten globule*) је компактна неуређена структура која може садржати значајне делове уређене структуре. Последње стање је уређена (eng. *order*) структура. Наведене структуре су приказане на слици 6.1

Било које од ових стања може бити природно стање, тј. стање које је битно за биолошку функцију. Неки неуређени протеини могу да прелазе из неуређеног у уређено стање и обрнуто после интеракције са другим макромолекулима или после промена у биохемијским процесима, док други остају у неуређеном облику у току обављања своје функције. Најпознатија база експериментално утврђених неуређених протеина је DisProt[1] база, са преко 600 неуређених протеина, који садрже неуређене регионе различитих дужина.

[1] http://www.disprot.org/



Слика 6.1: *Илустрација различитих структура протеина: O - одговара уређеној стрктури; MG (molten globule) је компактана неуређена структура са значајним деловима уређене структуре; PMG (pre-molten globule) је издужена делимично структуирана форма; Coil - комплетно неуређена структура. Први ред на слици представља просторну 3Д структуру полипептида дужине 100 АК. Други ред представља релативну хидродинамичку запремину одговарајућих структура исте дужине. Трећи ред представља релативну хидродинамичку запремину ове четири конформације за полипептид величине 500 АК. Кругови у другом и трећем реду означавају пораст релативне хидродинамичке запремине у односу на уређену структуру [107].*

Један од циљева ове тезе је и успостављање везе између Т - ћелијских епитопа и уређених/неуређених делова протеина у циљу што бољег разумевања Т - ћелијских епитопа и издвајања информација које би могле да се искористе у новим моделима за предвиђање епитопа.

6.2 Однос Т - ћелијских епитопа и уређених / неуређених делова протеина

Истраживање везано за утврђивање односа Т - ћелијских епитопа са уређеним и неуређеним структурама протеина је спроведено над 619 протеина. Већина протеина (465) је преузета из DisProt базе верзије 5.0, у којој се налазе протеини који припадају различитим организмима са експериментално утврђеним уређеним и неуређеним регионама. Преостали протеини су преузети да се задовоље следећи критеријуми:

- да се балансира број еукариотских и прокариотских протеина, као и однос уређених и неуређених региона. Већина преузетих протеина из DisProt базе је из групе еукариотских протеина са 70% неуређеном структуром. Із NCBI[2] базе су преузети прокариотски протеини (115). Из PDB[3] базе је преузето 15 протеина са преко 90% експериментално утврђене уређене структуре,
- да би се формирала и једна група тумор - асоцираних антигена (eng. *tumor-associated antigens*, TAA), укључено је и 19 протеина из групе канцер - тестис антигена [71].

Протеини су груписани у 3 таксономске категорије (223 бактеријска протеина, 376 еукариотских протеина и 20 еукариотски вируси). Неуређени региони у протеинима су одређени коришћењем 7 различитих предиктора наведених у табели 6.1, а Т - ћелијски епитопи су предвиђени коришћењем програма NetMHCpan и NetMHCIIpan (који су описани у поглављу 3.2). Резултати везани за однос предвиђених Т - ћелијских епитопа и уређених и неуређених структура у протеину су детаљније приказани у [71][82][72][73], док су овде приказани сумарни резултати који су били мотивација за избор и прављење нових шема за представљање пептида за улаз у нове моделе.

Утврђено је да се предвиђени Т - ћелијски епитопи у протеинским регионима (уређеним/неуређеним) појављују у једном од следећих облика:

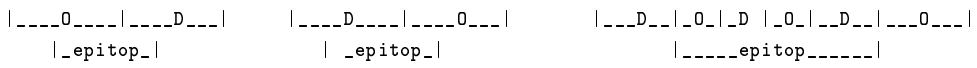
- О епитопи - епитопи који комплетно припадају уређеним регионима протеина.
- D епитопи - епитопи који комплетно припадају неуређеним регионима протеина.

[2] http://www.ncbi.nlm.nih.gov
[3] http://www.pdb.org

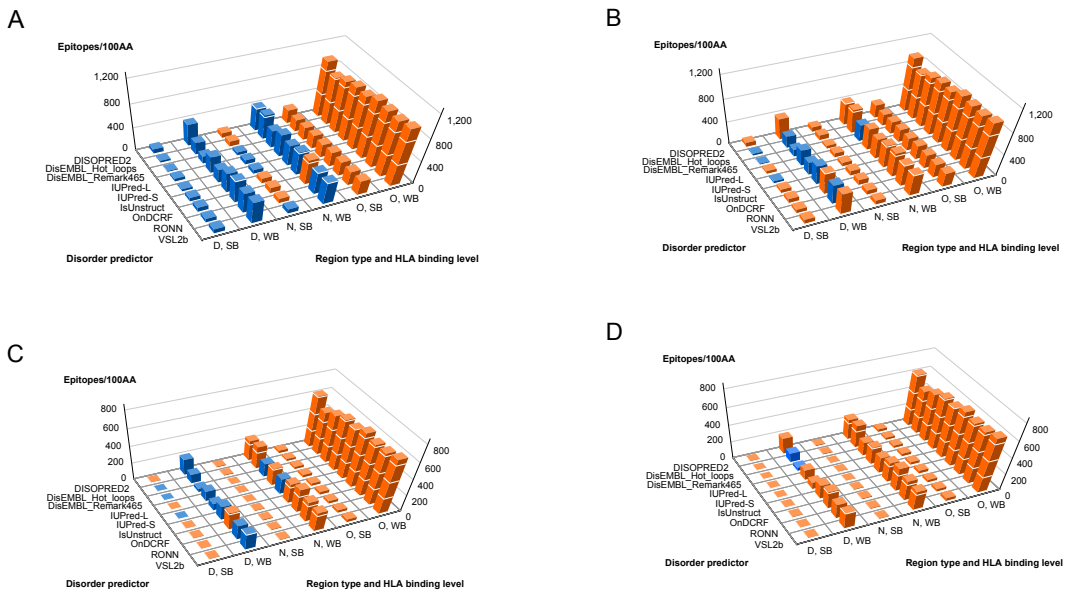
Предиктор	Локација
VSL2b	http://www.ist.temple.edu/disprot/predictorVSL2.php
IsUnstruct V2.02	http://bioinfo.protres.ru/IsUnstruct/
IUPred 1.0 - long disorder (IUPred -Long) - short disorder (IUPred-Short)	http://iupred.enzim.hu/
RONN 3.1	http://www.strubi.ox.ac.uk/RONN
DisEMBL - Hot-loops - Remark465	http://dis.embl.de/
OnDCRF 1.0	http://babel.ucmp.umu.se/ond-crf/
DISOPRED 2.4.3	http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1

Табела 6.1: *Предиктори неуређених региона коришћени у истраживању*

- N епитопи - епитопи који се делимично налазе у уређеним а делимично у неуређеним регионима, тј. на прелазима између ове две структуре.



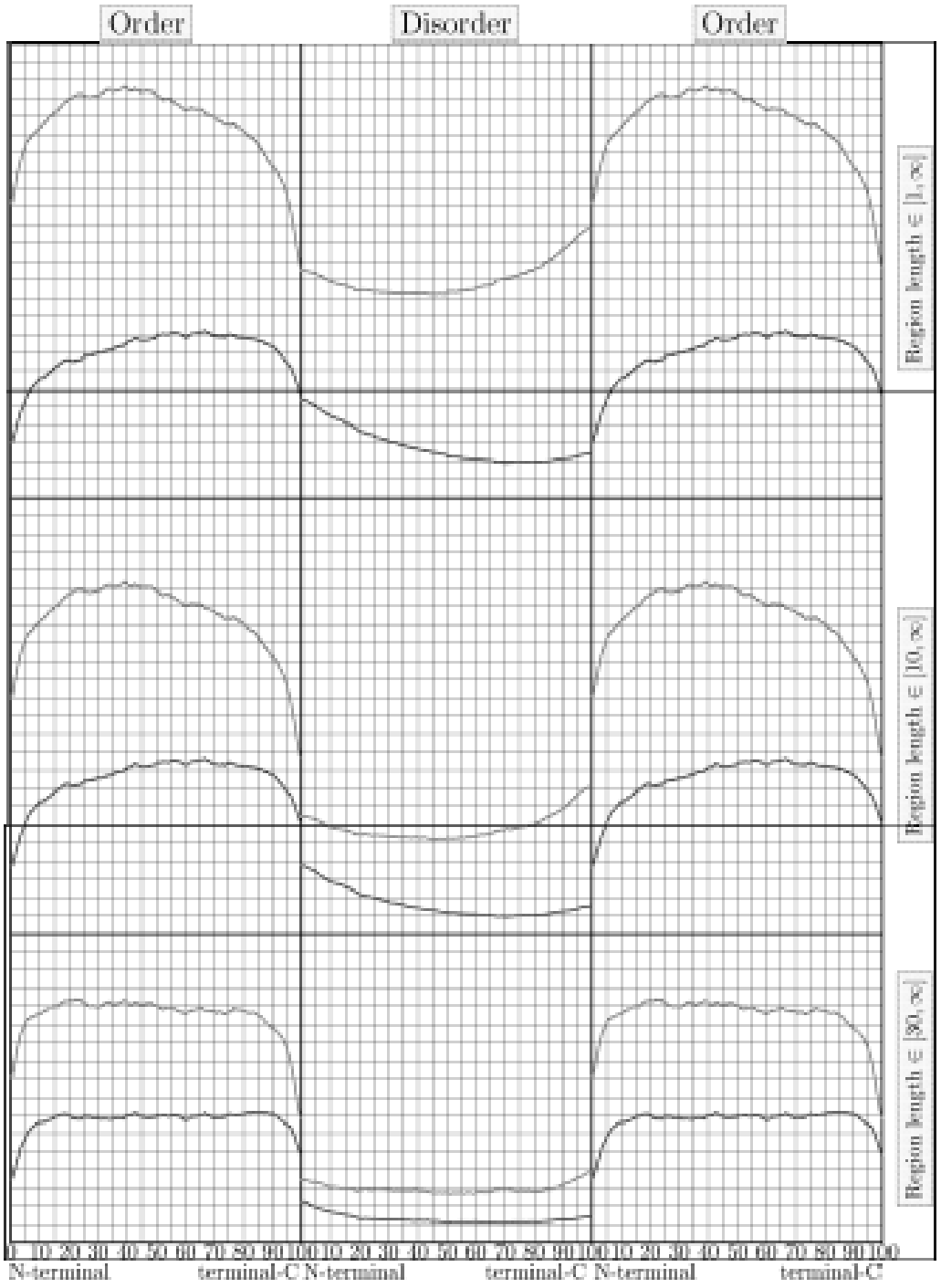
Анализа појављивања епитопа у протеинским регионима је извршена према броју појављивања епитопа у одговарајућим регионима и према нормализованој мери на 100 амино киселина. Утврђено је да су епитопи у уређеним структурама увек бројнији него епитопи у неуређеним структурама. Укупан број епитопа у уређеним структурама је од 2.83 до 24.75 пута већи него у неуређеним структурама када су у питању епитопи који се везују за молекуле *MHC* класе I. Тај однос за *MHC* класу II је знатно већи и иде од 3.41 до 44.65 пута у корист уређених региона. Резултати варирају у зависности од изабраног предиктора за предвиђање неуређених региона. Такође је утврђено да је однос јаких и слабих епитопа нешто већи у уређеним регионима него у неуређеним. У уређеним регионима тај однос варира од 0.25 до 0.26 за *MHC* класу I, и од 0.06 до 0.07 за *MHC* класу II. Уколико се уместо броја епитопа, анализира учесталост епитопа добијају се слични резултати. Добијени резултати потврђују да су неуређени региони слабији кандидати за T - ћелијске епитопе. Дистрибуција епитопа у предвиђеним уређеним и неуређеним структурама протеина је приказана на слици 6.2



Слика 6.2: *Припадност епитопа различитим структурама протеина, добијених предикторима из табеле 6.1*

Независно од избора предиктора за неуређене регионе, јасно се види да су епитопи у уређеним регионима знатно бројнији. Добијени резултати су у складу са истраживањем спроведеним у [18], у којем су анализирани T - ћелијски епитопи који се везују за *MHC* молекуле класе II, нуклеарних системских антигена. *Carl* и сарадници су установили да је учесталост T - ћелијских епитопа у неуређеним структурама анализираних протеина знатно мања него у уређеним структурама. На слици 6.3 је приказана покривеност различитих протеинских региона епитопима. Да би се израчунала покривеност епитопима уређених и неуређених региона различитих дужина, сви региону су прво скалирани на димензију највећег региона (799 амино киселина), а затим нормализовани на дужину 100. Релативна позиција епитопа у процесу скалирања/нормализације, остаје непромењена унутар региона. Над тако добијеним новим интервалима је рачуната покривеност епитопима. Покривеност епитопима је приказана за регионе различите дужине, у оригиналу: дужине ≥ 1, дужине ≥ 10 и дужине ≥ 30. Утврђено је да на тренд криве која представља покривеност епитопима не утиче избор предиктора за предвиђање неуређене структуре нити афинитет везивања епитопа (док год је афинитет преко $500 IC_{50}$).

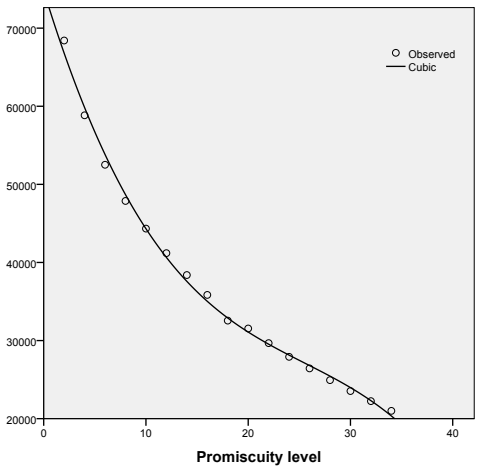
Са слике 6.3 се јасно види да су епитопи у уређеним регионима бројнији, и да је тенденција опадања учесталости ка прелазу из уређених у неуређене делове протеина (на N и C терминалима региона). Такође, крива која представља

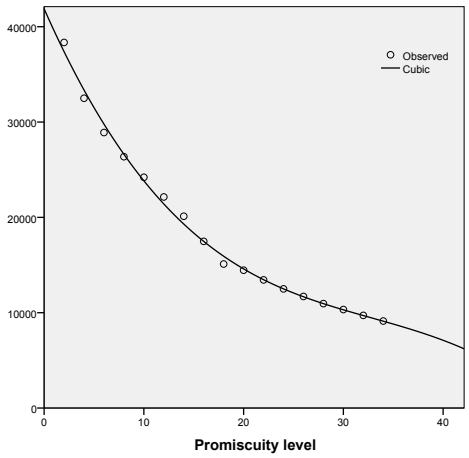


Слика 6.3: *Дистрибуција епитопа у различитим регионима протеина (уређеним и неуређеним, на графику означеним као* Order *и* Disorder*): x - оса представља позицију у одговарајућем региону, y - оса представља број епитопа на одговарајућој позицији. Дистрибуција је рачуната за све епитопе МНС класе I и II.*

покривеност је конкавна у уређеним регионима и достиже максималне вредности око централног дела неуређеног региона. Насупрот томе, је тенденција раста броја епитопа ка прелазима из неуређених у уређене делове протеина, за *MHC* I класу више него за *MHC* II класу. Крива којом је представљена покривеност епитопима у неуређеним регионима је конвексна и достиже минимилне вредности око централног дела неуређеног региона. Тиме је потврђено да су уређени региони не само бољи кандидати за епитопе, већ су и богатији промискуитетним епитопима (епитопи који се везују за већи број алела), док неуређени региони оскудевају са епитопима а посебно промискуитетним епитопима.

Утврђено је да се промискуитетни епитопи, у зависности од броја алела за које се везују, могу апроксимирати полиномом трећег степена (слика 6.4), што се може искористити за предвиђање броја епитопа који ће се везивати са одређеним нивоом промискуитетности. Тврђење важи за обе HLA класе.





Model Summary and Parameter Estimates

Dependent Variable: Number of HLA-I epitopes

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Cubic	,997	1437,664	3	13	,000	74813,838	-4284,156	142,063	-1,858

Dependent Variable: Number of HLA-II epitopes

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Cubic	,996	1167,375	3	13	,000	41887,176	-2373,809	63,316	-,643

The independent variable is Promiscuity level.

Слика 6.4: *Апроксимација број промискуитетних епитопа у зависности од нивоа промискуитетности за обе HLA класе.*

6.2.1 Хидрофобна и хидрофилна својства T - ћелијских епитопа у уређеним и неуређеним структурама протеина

У оквиру тезе исцрпно је истраживана хидропатија (eng. *hydropathy*) епитопа, вероватноћа да је епитоп хидрофобан/хидрофилан, интезитет хидропатије епитопа који се везују за сваки од *MHC* молекула, у уређеним и неуређеним структурама. Хидропатија епитопа је рачуната према две најпопуларније (најчешће коришћене) скале: *Kyte–Doolittle* (KD) и *Hopp–Woods* (HW)[4] [58][37]. Примењене су две методе за обе скале, где се хидропатија епитопа рачуна:

- као просечна вредност мере хидропатије свих појединачних амино киселина које улазе у састав пептида ($AvgH$), и
- као број амино киселина које су хидрофобне/хидрофилне у пептиду (MAA скр. од eng. *majority of amino acids*).

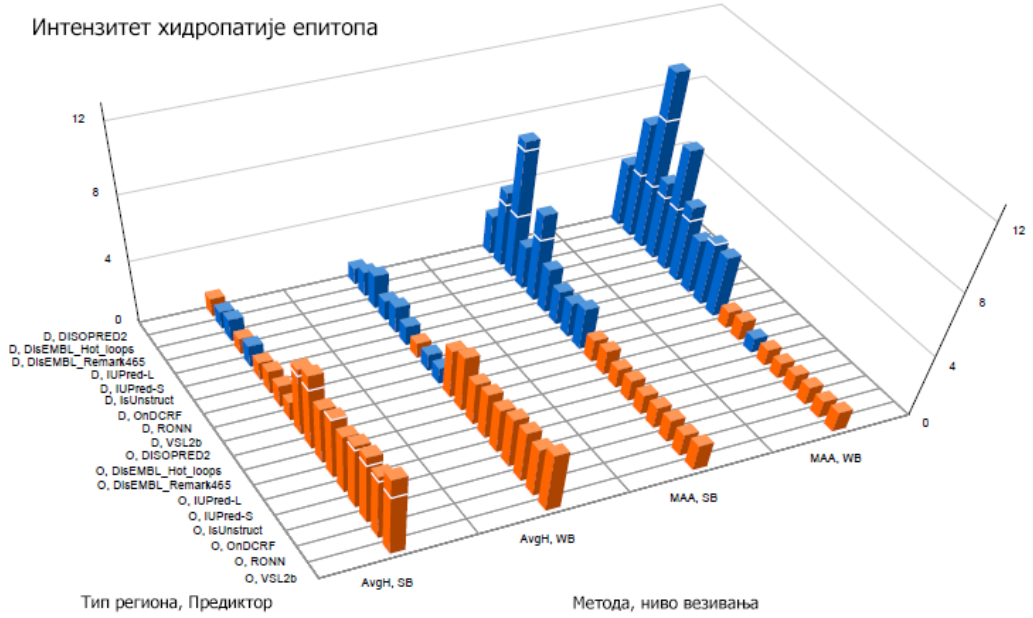
Применом прве методе је подразумевано да је пептид хидрофобан ако је $(AvgH) \geq 0$, у супротном је пептид хидрофилан према KD скали, према HW скали је хидрофобан за $(AvgH) \leq 0$[5]. Применом методе MAA пептид се сматра хидрофобним ако су ходрофобне амино киселине бројније у пептиду, у супротном се сматра хидрофилним. На слици 6.5 је приказана дистрибуције епитопа у уређеним и неуређеним регионима у протеину, и приказан је однос хидрофобних/хидрофилних епитопа по тим регионима.

Сумарни резултати указују да су епитопи у уређеним регионима углавном хидрофобни, док су епитопи у неуређеним регионима углавном хидрофилни. Такође је утврђено да су епитопи са јачим афинитетом везивања хидрофобнији од епитопа са слабијим афинитетом везивања без обзира у ком се региону налазе, што указује на то да хидрофобност има директан утицај на афинитет везивања епитопа. Посматрањем хидропатије по супертиповима, приказано је на слици 6.6, се може закључити да за појединачан супертип не важи једно опште правило.

Добијени резултати, везани за хидрофобност епитопа по супертиповима и уређеним/неуређеним регионима, потврђују да су епитопи учесталији у уређеним регионима. Међутим, може се видети са слике 6.6 да резултати указују на то да и у оквиру самих супертипова постоје епитопи са различитим карактеристикама (и хидрофобни и хидрофилни епитопи). Постоје супертипови где су епитопи претежно хидрофобни (нпр. A2), и такође

[4] http://gcat.davidson.edu/rakarnik/kyte-doolittle-background.htm

[5] KD и HW скале дају различите мере за хидрофобност пептида, видети [37][58]

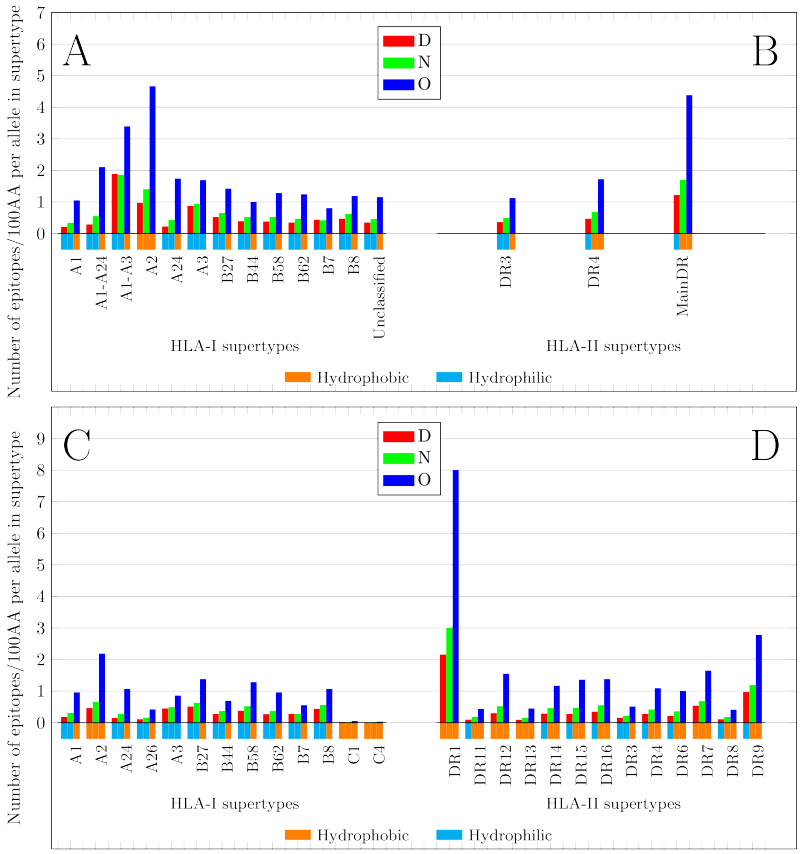


Слика 6.5: *Хидрофобност и диструбуција епитопа у уређеним и неуређеним структурама протеина, добијеним применом различитих предиктора. На графику су наранџастом бојом означени претежно хидрофобни епитопи, док су плавом бојом означени претежно хидрофилни.*

постоје супертипови где су епитопи претежно хидрофилни (нпр. A3), као и они где хидропатија епитопа директно зависи од структурног региона у коме се налази епитоп (нпр. A1-A2, A1-A3, B27). Закључује се да би додатан избор физичко хемијских особина, поред хидропатије, епитопа могао прецизније да одреди понашање епитопа у оквиру супертипа.

6.3 Правила придруживања

Првобитно је техника правила придруживања примењена на све предвиђене епитопе како би се установила правила о амино киселинама које се јављају на одређеној позицији у епитопу или које се заједно јављају. Посматран је скуп свих предвиђених епитопа за појединачан алел E. Епитоп је представљен као скуп ставки $e = \{e_1, e_2, ..., e_9\}$ тј. амино киселина које улазе у састав епитопа. Сваки епитоп представља једну трансакцију. За издвајање правила је коришћен Априори алгоритам детаљно описан у глави 2. Добијена правила су у форми $X_i \rightarrow Y_j\,[s, c]$, где су X и Y амино киселине i и j су позиције у пептиду а $s\%$ и $c\%$ представљају ниво подрешке и поверења, респективно. Нпр. правило $V_9 \rightarrow L_2\,[20.2, 0.68]$ значи да се амино киселина V јавља на другој позицији у 20.2% пептида и да се у 68% тих случајева амино киселина L јавља на позицији



Слика 6.6: *Учесталост епитопа по супертиповима алела класа HLA-I и HLA-II и хидропатија епитопа према KD скали и AvgH методи за две различите класификације по супертиповима.*

9. Правила добијена на овај начин одговарају већ познатим информацијама о сидро позицијама и нису приказана овде. У циљу добијања нових информација техника правила придруживања је примењена на све уређене и неуређене регионе у протеину, где се као једна трансакција посматра састав региона (заступљеност амино киселине у региону), учесталост амино киселина у региону, заступљеност епитопа обе *MHC* класе у региону и мера хидропатије епитопа у одговарајућем региону. Неки од издвојених резултата добијених правилима придруживања су и следећа тврђења:

- Уколико је процентуална заступљеност амино киселине A у неком неуређеном региону већа од просечне заступљености те амино киселине у неуређеним регионима онда тај регион највероватније садржи епитопе *MHC* класе I.

- Уколико је процентуална заступљеност амино киселине W већа од просечне заступљености те амино киселине, регион вероватно садржи

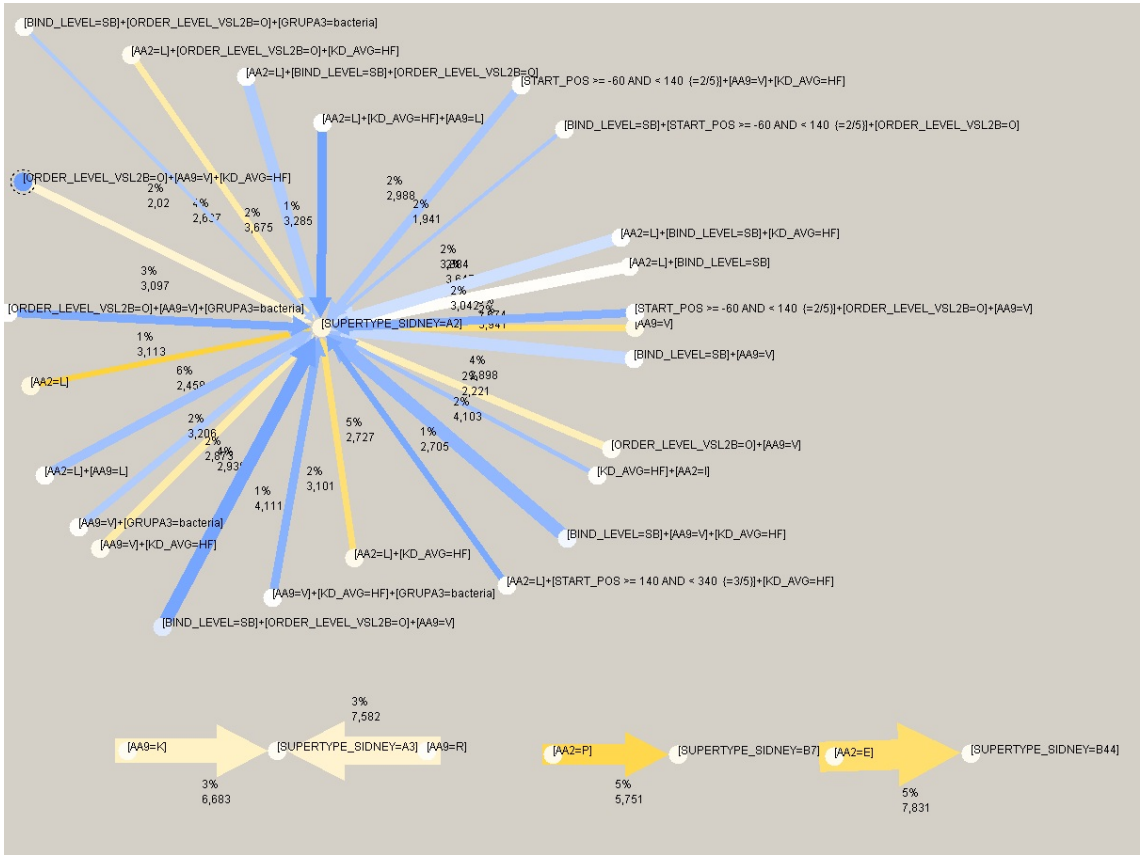
епитопе *MHC* класе I (исто правило је добијено и за уређене и за неуређене регионе).

- већа учесталост амино киселине L у неком делу региона него што је просечна учесталосм у том региону указује на присуство епитопа обе *MHC* класе у том региону.

- Ако неки од региона има мању заступљеност амино киселине N од просечне, онда је вероватно да се у томе региону налазе епитопи *MHC* класе I, ако је пак учесталост већа од просечне онда регион највероватније садржи епитопе *MHC* класе II.

Нека од правила везана за супертипове алела и заступљеност амино киселина на одговарајућој позицији у епитопу су графички приказана на слици 6.7. Добијена правила су у складу са познатим подацима везаним за сидро позиције у епитопима (друга и девета позиција у епитопу, видети 3.2.1).

Сви добијени резултати указују да учесталост амино киселина у пептиду, као и његова припадност уређеном/неуређеном региону протеина, а тиме и одређена физичко хемијска својста (у конкретном случају хидропатија пептида), могу да укажу на то да ли је пептид потенцијални епитоп или не. Управо ови резултати су били мотивација за нове шеме енкодирања и избор атрибута за представљање пептида у новим направљеним моделима, описаним у глави 4, добар.

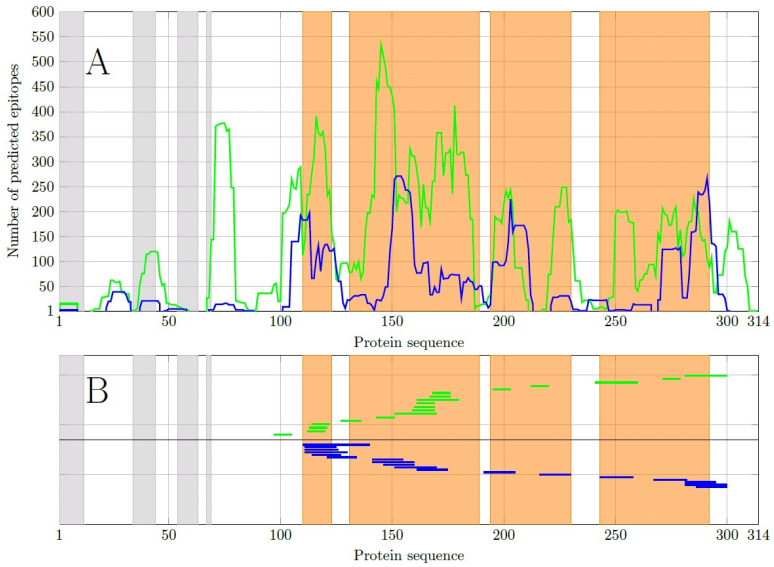


Слика 6.7: *Правила придруживањима у супертиповима алела, заснована на аминокиселинском садржају епитопа на одговарајућој позицији.*

6.4 Експериментална потврда о припадности T - ћелијских епитопа различитим структурама протеина

Сви описани резултати о припадности T - ћелијских епитопа различитим регионима протеина су изведени на основу предвиђених *MHC* везујућих пептида. У оквиру овог поглавља су анализирани прикупљени експериментално утврђени епитопи и њихове везе са уређеним и неуређеним регионима у протеину. Група канцер антигена (eng. *cancer testis antigena*, CTA) која се јавља у различитим туморима, али не и здравим ћелијама је анализирана ради експерименталне потврде понашања T - ћелијских епитопа у уређеним и неуређеним регионима. Већина епитопа из 19 антигена је лоцирана баш у консензусу уређених региона различитих предиктора, или евентуално њиховој близини (видети слику 6.8).

Протеин MAGE-A3 је заступљен у 76% метастазираних меланома, преко 50% у канцеру плућа, и многим другим малигнитетима. Интезивно је



Слика 6.8: *Покривеност епитопима хуманог MAGE-A3 протеина (Uniprot Acc. No. P43357). А) Епитопи предвиђени NetMHCpan (HLA класа I)(зелена боја) и NetMHCIIpan предиктором (HLA класа II) (плава); Предвиђања су рађена за све алеле. В) Експериментално утврђени епитопи HLA класе I (зелена) и HLA класе II (плава)*

проучаван за имунотерапију тумора. У резултатима добијеним програмима за предвиђање епитопа, готово сви епитопи обе *HLA* I и *HLA* II класе припадају консензусу уређених региона (према свим предикторима) слика 6.8(А). Сви експериментално утврђени епитопи се налазе у уређеним регионима добијеним као консензус из свих предиктора, слика 6.8(В). Урађено је пуно експерименталних студија над протеином MAGE-A3. Једна од њих је [8], где су истраживачи вакцинисали 18 пацијената са канцером плућа, за вакцину су користили MAGE-A3 комплетан антиген рекомбинован са упалним адјувансима. У оквиру истраживања и рада на овој тези је истраживање са експерименталним подацима проширено, не само на експерименталне епитопе већ и експериментално утврђену секундарну структуру. Сви добијени резултати су публиковани у [82]. Добијеним резултатима је потврђено да се експериментални Т - ћелијски епитопи обе класе налазе у уређеним или на прелазним регионима. Анализирани су експериментални подаци за тумор асоциране антигене и нуклеарне системске антигене.

У оквиру овог дела истраживања је направљен нови софтверски систем, који је омогућио припрему података и детаљну анализу. Нови софтверски систем има широку примену у предвиђању различитих карактеристика протеина, обради, складиштењу и визуелизацији добијених резултата. Кратак опис развијеног система се налази у додатку B.1, а детаљан опис система је публикован у [45]. Пример употребе систем је дат у [44].

Поглавље 7

Закључак и даљи рад

У оквиру ове тезе разматран је проблем идентификовања Т - ћелијских епитопа, тренутно један од најактуелнијих проблема у области имуноинформатике. Описане су постојеће рачунарске методе за проналажење Т - ћелијских епитопа, и представљени су нови направљени модели. Представљени модели су упоређени са тренутно најбољим системима и методама за предвиђање Т - ћелијских епитопа, где су се показали као упоредиви, а у неким случајевима и бољи у предвиђању Т - ћелијских епитопа. Нови модели би могли да буду веома корисна допуна постојећим системима. Направљени су класификациони и регресиони модели засновани на техници подржавајућих вектора, SVM и SVR. Да би се техника подржавајућих вектора применила на конкретан проблем (предвиђање Т - ћелијских епитопа), било је потребно прилагодити улаз у векторском облику, који је најпогоднији улаз у SVM и SVR модела. Сви расположиви и прикупљени подаци о Т - ћелијским епитопима постоје у виду пептидне секвенце. У оквиру тезе, у сврху припреме података за погодан улаз у моделе, развијене су три нове стратегије енкодирања пептидне секвенце. Мотивација за нове стратегије/шеме енкодирања су резултати добијени исцрпним истраживањем везе Т - ћелијских епитопа и уређених и неуређених региона протеина, испитивањем мере хидропатије епитопа у уређени и неуређеним регионима, као и истраживање и анализа методологије рада постојећих метода, и узимањем у обзир свих недостатака у постојећим методама.

Применом прве нове шеме представљана пептида је добијен модел који је упоредив са два тренутно најбоља предиктора за Т - ћелијске епитопе. У оквиру шеме је примењена нова техника рачунања учесталости амино киселина по позицијама у пептидној секвенци $\Delta - BM25 - IDF$ техника. Техника се већ дуги низ година успешно користи у проблемима класификовања документа

и проблемима истраживања мишљена на друштвеним мрежама, али никада пре није коришћена за ову врсту проблема. Оно што је нарочито важно истаћи је да су нови направљени модели знатно мање димензионалности него постојећи модели, и да се нова техника веома једноставно примењује и лака је за разумевање. Нови направљени модели засновани на овој шеми су из тог разлога при извршавању знатно бржи и једноставнији од постојећих модела. Пептидна секвенца нонамера се представља вектором димензије 180, док је димензија у постојећим моделима 360 или више.

Друга направљена шема енкодирања пептида се такође ослања на $\Delta - BM25 - IDF$ технику рачунања фреквентности, али уз то комбинује резултате добијене истраживањем утицаја физичко хемијског састава пептида на афинитет везивања пептида, односно да ли ће бити препознат као епитоп или неепитоп. Добијени резултати потврђују да физичко хемијски састав амино киселина које улазе у састав пептида имају директан утицај на афинитет везивања. Кластер анализом, и класификационим моделима заснованим на кластеровању су издвојене физичко хемијске особине амино киселина које имају највећи утицај на раздвајање епитопа од неепитопа. Међу издвојеним особинама се налази оне повезане са хидропатијом, поларност, шаржа, волумен итд. Резултати који су мање били очекивани су да неки алели који не припадају истом супертипу чак ни истој *MHC* класи, деле заједничке физичко хемијске особине које су одлучујуће у раздвајању епитопа од неепитопа. Управо ти резултати наговештавају могуће будуће правце у побољшању перформанси постојећих модела. Направљени класификациони модели засновани на кластер анализи, за издвајање најзначајних физичко хемијских особина амино киселина у пептиду, су се такође показали као упоредиви по карактеристикама са постојећим предикторима.

Трећа развијена шема представља проширење прве шеме енкодирања са 5 молекуларних дескриптора амино киселина.

Модели направљени коришћењем све три шеме су дали одличне резултате, и у великом броју случајева боље од тренутно најбољих предиктора за овај проблем. Међусобним упоређивањем нових модела, је показано да су перформансе модела врло блиске али и да постоје одступања односно да је за неке алеле боља једна шема енкодирања пептида, за неке друга док је за неке трећа. То указује да би комбинација три нова модела могла да резултује знатно бољим перформансама.

У фази припреме су самостални предиктори који у основи имају нове направљене моделе, и планирана је њихова интеграција у описани софтверски

систем развијен током рада на овој тези. Тиме ће бити омогућено директо поређење свих предиктора за T - ћелијске епитопе, укључујући и нове засноване на представљеним моделима, као и њихово комбиновање. Такође, у плану је и прављене модела за алеле који нису укључени у ово истраживање а за које су прикупљени експериментални подаци.

А Резултати класификационих модела заснованих на кластеровању

У табели А.1 су приказани додатни резултати тестирања бинарних класификационих модела заснованих на кластеровању над подацима MHCBN базе која нема преклапања са прикупљеним подацима из IEDB базе.

		Тренинг скуп				MHCBN тест скуп		
Алел	бр.	Прецизност	Одзив	Капа	Тачност	Прецизност	Одзив	Тачност
HLA-A0201	8	0.93	0.95	0.88	0.94	0.98	0.83	0.83
HLA-A0301	10	0.88	0.98	0.89	0.95	0.99	0.76	0.76
HLA-A1101	10	0.93	0.97	0.9	0.95	0.98	0.88	0.87
HLA-A0203	9	0.95	0.98	0.93	0.97	0.99	0.9	0.9
HLA-B1501	10	0.95	0.98	0.94	0.97	0.99	0.82	0.82
HLA-B0702	10	0.95	0.99	0.95	0.98	0.99	0.72	0.72
HLA-A0101	10	0.93	0.997	0.96	0.97	1	0.43	0.81
HLA-A0206	5	0.98	0.99	0.96	0.98	0.99	0.87	0.87
HLA-A2402	5	0.98	0.98	0.95	0.98	0.97	0.74	0.7
HLA-A3101	8	0.89	0.97	0.9	0.95	1	0.47	0.47
HLA-A0202	3	0.99	0.995	0.97	0.99	1	0.87	0.88
HLA-B0801	10	0.98	1	0.98	0.99	Нема података		
HLA-A2601	10	0.87	1	0.92	0.98	Нема података		
HLA-DRB10101	3	1	0.94	0.82	0.95	0.8	0.93	0.76
HLA-DRB10401	3	1	0.99	0.99	0.996	0.7	0.74	0.65

Табела А.1: *Резултати класификационих модела заснованих на кластеровању применом над подацима MHCBN базе3.1, бр. представља број модела добијених техником k - средина који је укључен у консенсузни модел. Сви сакупљени пептиди су нонамери и не постоје преклапања са подацима прикупљеним из IEDB базе3.1*

У табели А.2 су приказани упоредни резултати тестирања класификационих модела заснованих на кластеровању са резултатима два предиктора:

	БКК						NetMHCpan				MHCPred			
Алел	Бр.м	тест скуп	Прецизност	Одзив	Капа	Тачност	Прецизност	Одзив	Капа	Тачност	Прецизност	Одзив	Капа	Тачност
HLA-A0201	8	тест	0.87	0.82	0.71	0.85	0.97	0.83	0.82	0.91	0.73	0.76	0.48	0.74
HLA-A0301	10	тест	0.76	0.7	0.6	0.82	0.9	0.86	0.83	0.92	0.43	0.86	0.21	0.56
HLA-A1101	10	тест	0.86	0.71	0.64	0.83	0.97	0.86	0.85	0.93	0.42	1	0.0087	0.43
HLA-A0203	9	тест	0.8	0.75	0.57	0.78	0.96	0.88	0.85	0.92	0.53	0.87	0.13	0.56
HLA-B1501	10	тест	0.9	0.79	0.72	0.86	0.99	0.8	0.81	0.9	нема података			
HLA-B0702	10	тест	0.86	0.6	0.6	0.81	0.92	0.92	0.88	0.94	нема података			
HLA-A0101	10	тест	0.78	0.51	0.55	0.88	0.72	0.97	0.77	0.92	0.55	0.71	0.51	0.83
HLA-A0206	5	тест	0.79	0.84	0.48	0.76	0.94	0.89	0.79	0.9	0.62	0.82	0.04	0.59
HLA-A2402	5	тест	0.71	0.71	0.35	0.68	0.88	0.85	0.7	0.85	нема података			
HLA-A3101	8	тест	0.72	0.61	0.53	0.81	0.72	0.61	0.53	0.81	0.41	0.41	0.14	0.63
HLA-A0202	3	тест	0.8	0.9	0.48	0.78	0.86	0.82	0.77	0.9	0.77	0.78	0.33	0.7
HLA-B0801	10	тест	0.84	0.74	0.68	0.86	0.97	0.87	0.88	0.94	нема података			
HLA-A2601	10	тест	0.74	0.42	0.48	0.9	0.78	0.9	0.81	0.95	нема података			

Табела А.2: *Упоредни резултати тестирања предиктора NetMHCpan, MHCPred, и модела бинарне класификације заснован на кластеровању (ББК) над подацима IEDB базе из тестног скупа података који нема преклапања са скупом података за тренирање модела. Бр.м. представља број појединачних модела груписања добијених техником кластеровања k - срединама укључених у консензусни модел и ББК.*

Б "Најбоље" ФХ особине по алелима за унираме и бираме

Табела Б.1: *Издвојене најбоље физичко хемијске особине по алелима*

Алел	За унираме	За бираме
A0101	Information measure for extended without H-bond	Free energy of solution in water, kcal/mole
	Free energy change of e(i) to a(Rh)	Polarizability parameter
	Free energy of solution in water, kcal/mole	Residue accessible surface area in tripeptide
	Refractivity	Normalized van der Waals volume
	Normalized van der Waals volume	Residue Side-Chain Volume
	Flexibility parameter for two rigid neighbors	Side-chain volume
	Number of hydrogen bond acceptors	STERIMOL maximum width of the side-chain
	Partial specific volume	van der Waals parameter e
	Residue accessible surface area in tripeptide	Relative mutability
	van der Waals parameter Ralpha-NH chemical shifts	Refractivity
A0201	Surrounding hydrophobicity in folded form	HPLC parameter
	Direction of hydrophobic moment	Mean polarity
	Hydropathy index	Polarity
	HPLC parameter	Partition coefficient

	Partition coefficient	Solvation free energy
	Side-chain hydropathy, corrected for solvation	Electron-ion interaction potential
	Average gain in surrounding hydrophobicity	Average flexibility indices
	Mean polarity	Hydropathy index
	pK-C	Hydrophobic parameter p
	Unfolding Gibbs energy in water, pHNet charge.alpha-NH chemical shifts	Side-chain hydropathy, corrected for solvation
A0202	HPLC parameter	Partition coefficient
	Normalized hydrophobicity scales for alpha-proteins	HPLC parameter
	Transfer free energy to lipophilic phase	Average gain in surrounding hydrophobicity
	Unfolding Gibbs energy in water, pH9.alpha-NH chemical shifts	Retention coefficient in NaClOPositive charge
	Optimal matching hydrophobicity	Polarity
	Mean polarity	Normalized hydrophobicity scales for beta-proteins
	Hydration free energy	Transfer free energy, CHP/water
	Volume	Mean polarity
	Transfer energy, organic solvent/water	Retention coefficient in HFBA
	Normalized hydrophobicity scales for beta-proteins	Hydrophobic parameter p
A0203	Hydropathy index	Polarity
	Unfolding Gibbs energy in water, pH9.alpha-NH chemical shifts	Hydropathy index
	Transfer free energy to lipophilic phase	Average gain in surrounding hydrophobicity
	Surrounding hydrophobicity in folded form	Surrounding hydrophobicity in folded form

	Partition coefficient	Mean polarity
	Hydration potential	Average gain ratio in surrounding hydrophobicity
	Side-chain hydropathy, corrected for solvation	Optimal matching hydrophobicity
	Transfer free energy from chx to wat	Partition coefficient
	Consensus normalized hydrophobicity scale	Transfer free energy from chx to wat
	Surrounding hydrophobicity in turn	HPLC parameter
A0206	Solvation free energy	Negative charge
	Transfer free energy to lipophilic phase	STERIMOL minimum width of the side-chain
	Normalized hydrophobicity scales for alpha + beta-proteins	The number of atoms in the side-chain labeled
	Optimal matching hydrophobicity	Net charge
	HPLC parameter	Polarity
	Hydropathy index	Isoelectric point
	Mean polarity	Isoelectric point
	Side-chain hydropathy, corrected for solvation	Graph shape index
	Electron-ion interaction potential	Free energy change of e(i) to a(Rh)
	Average flexibility indices	Electron-ion interaction potential
A0301	Energy transfer from out to in(9Negative charge%buried)	Positive charge
	Hydration number	Polar requirement
	Net charge	Net charge
	Consensus normalized hydrophobicity scale	Polarity
	Hydrogen bond acceptor factors	Hydrophobicity
	Average gain ratio in surrounding hydrophobicity	Energy transfer from out to in(9Negative charge%buried)
	Polarity	Hydration number
	Hydrophobicity	Fraction of site occupied by water

	Surrounding hydrophobicity in folded form	Average gain ratio in surrounding hydrophobicity
	Atom-based hydrophobic moment	Hydrophobic parameter
A1101	Hydrophobicity	Positive charge
	Consensus normalized hydrophobicity scale	Hydrophilicity value
	Electron-ion interaction potential	Hydrophobicity
	Average flexibility indices	Polarity
	Energy transfer from out to in(9Negative charge%buried)	Energy transfer from out to in(9Negative charge%buried)
	Hydrogen bond acceptor factors	Consensus normalized hydrophobicity scale
	Average gain ratio in surrounding hydrophobicity	Hydrophobic parameter
	Atom-based hydrophobic moment	Polar requirement
	Polar requirement	Average gain ratio in surrounding hydrophobicity
	Surrounding hydrophobicity in folded form	Hydrogen bond acceptor factors
A2402	Retention coefficient in HPLC, pHA parameter of charge transfer donor capability.1	van der Waals parameter e
	Direction of hydrophobic moment	Transfer energy, organic solvent/water
	Flexibility parameter for two rigid neighbors	Solvation free energy
	Retention coefficient in TFA	Optimized propensity to form reverse turn
	Mean polarity	Average volume of buried residue
	Side-chain hydropathy, corrected for solvation	Hydrophobic parameter
	Retention coefficient in HFBA	Normalized van der Waals volume

	Free energy change of e(i) to a(Rh)	Side-chain hydropathy, corrected for solvation
	Transfer energy, organic solvent/water	Hydrophilicity value
	Transfer free energy from oct to wat	Polarizability parameter
A2601	Negative charge	Negative charge
	Optimal matching hydrophobicity	STERIMOL minimum width of the side-chain
	STERIMOL minimum width of the side-chain	The number of atoms in the side-chain labeled
	Net charge	Net charge
	Normalized hydrophobicity scales for beta-proteins	Polarity
	The number of atoms in the side-chain labeled	Isoelectric point
	Partial specific volume	a-CH chemical shifts
	Dependence of partition coefficient on ionic strength	Graph shape index
	Isoelectric point	Free energy change of e(i) to a(Rh)
	Normalized hydrophobicity scales for alpha-proteins	Electron-ion interaction potential
A3101	Principal property value z3	Polarity
	Number of hydrogen bond donors	Atom-based hydrophobic moment
	Consensus normalized hydrophobicity scale	Consensus normalized hydrophobicity scale
	Side-chain hydropathy, uncorrected for solvation	Hydrophobicity
	Solvation free energy	Positive charge
	Transfer free energy from oct to wat	Net charge
	Hydrogen bond acceptor factors	Activation Gibbs energy of unfolding, pH9.alpha-NH chemical shifts

	Polarity	Activation Gibbs energy of unfolding, pHNet charge.alpha-NH chemical shifts
	Hydration potential	Energy transfer from out to in(9Negative charge%buried)
A6802	The number of bonds in the longest chain	Isoelectric point
	Isoelectric point	Energy transfer from out to in(9Negative charge%buried)
	1Positive charge	Hydrogen bond acceptor factors
	The number of atoms in the side-chain labeled	Polarity
	Distance between Ca and centroid of side-chain	STERIMOL minimum width of the side-chain
	Hydrogen bond acceptor factors	Atom-based hydrophobic moment
	Energy transfer from out to in(9Negative charge%buried)	Hydrophobicity
	STERIMOL minimum width of the side-chain	Consensus normalized hydrophobicity scale
	Radius of gyration of side-chain	Hydropathy index
	STERIMOL maximum width of the side-chain	Average gain ratio in surrounding hydrophobicity
B0702	Smoothed upsilon steric parameter	pK-N
	Principal property value z3	alpha-NH chemical shifts
	Free energy of solution in water, kcal/mole	Accessible surface area
	Transfer free energy, CHP/water	Transfer free energy, CHP/water
	The number of bonds in the longest chain	Information measure for extended without H-bond
	Hydrogen bond acceptor factors	Smoothed upsilon steric parameter
	Side-chain angle theta(AAR)	Free energy of solution in water, kcal/mole

	Optimized propensity to form reverse turn	Side-chain angle theta(AAR)
	Side-chain hydropathy, uncorrected for solvation	pK-a(RCOOH)
	Volume	The number of atoms in the side-chain labeled
B0801	Positive charge	Polarity
	Net charge	Net charge
	Polarity	Positive charge
	Atom-based hydrophobic moment	The number of bonds in the longest chain
	Side-chain hydropathy, corrected for solvation	Residue Side-Chain Volume
	Mean polarityNet charge	Atom-based hydrophobic moment
	Distance between Ca and centroid of side-chain	Distance between Ca and centroid of side-chain
	The number of bonds in the longest chain	A parameter of charge transfer donor capability
	Energy transfer from out to in(9Negative charge%buried)	Energy transfer from out to in(9Negative charge%buried)
	Principal property value z3	Hydrogen bond acceptor factors
B1501	Optimal matching hydrophobicity	Optimal matching hydrophobicity
	Surrounding hydrophobicity in folded form	Hydrophilicity value
	Surrounding hydrophobicity in beta-sheet	Information measure for extended without H-bond
	Average gain in surrounding hydrophobicity	Principal property value z3
	Normalized hydrophobicity scales for alpha-proteins	Polar requirement
	van der Waals parameter e	Electron-ion interaction potential
	Transfer free energy, CHP/water	Average flexibility indices
	Free energy change of e(i) to a(Rh)	Polarity

	Hydrophobic parameter Transfer energy, organic solvent/water	Isoelectric point Net charge
B4403	Dependence of partition coefficient on ionic strength Retention coefficient in HPLC, pH-Net charge.Positive charge Normalized hydrophobicity scales for beta-proteins Average volume of buried residue Accessible surface area Isoelectric point Hydrophilicity value Residue volume Size Fraction of site occupied by water	Dependence of partition coefficient on ionic strength Retention coefficient in HPLC, pH-Net charge.Positive charge Transfer free energy from oct to wat Hydrophobic parameter Transfer energy, organic solvent/water Retention coefficient in HPLC, pHA parameter of charge transfer donor capability.1 Accessible surface area Normalized van der Waals volume Polarizability parameter Size

B EpDis-MassPred систем

Систем је настао интеграцијом алата који су направљени за потребе припреме података за истраживања везана за T - ћелијске епитопе, неуређене регионе у протеину и одређивање хидропатије протеинских региона.

Истраживачи који се баве идентификовањем T - ћелијских епитопа, уређеним и неуређеним регионима у протеину су суочени са великим бројем проблема: чест случај који се јавља у оваквим истраживањима је потреба за комбиновањем резултата везаних за две или више карактеристика протеина, као што је случај описан у претходном поглављу где се комбинују резултати везани за T - ћелијске епитопе, неуређене/уређене регионе у протеину и хидропатију протеинског региона. Ради добијања квалитетних резултата применом техника истраживања података потребно је применити неку методу на велики број кандидатских протеина. Укључивање више протеина у истраживање додатно усложњава добијање резултата, успорава процес приказа и анализе резултата и захтева начин чувања добијених резултата који омогућава њихово брзо дохватање ради замене протеина који учествује у анализи. Додатни проблем код коришћења различитих предиктора представља неусаглашеност начина представљања излазних резултата што их чини непогодним за даљу анализу или поређење чак и у случају појединачних протеина. Неки од предиктора су временски захтевни; уз то ако је приступ предиктору омогућен само преко веб апликације то га чини неодговарајућим за масовну примену на велики број протеина. Са друге стране ни предиктори који су јавно расположиви у облику самосталних апликација нису увек погодни за масовну примену на више стотина протеина, а уз то неретко при инсталацији захтевају одређене програмерске вештине које шири круг истраживача не поседује. Сви наведени разлози су били мотивација за развој новог система EpDis-MassPred [45]. Систем чине два алата EpDis (EPitopes in DISorder) који је првобитно настао као алат за припрему података у оквиру рада на магистарској тези [46], и MassPred алата који је развијен као самосталан алат за потребе припреме података и

истраживања различитих карактеристика протеина[1]. Оба алата су у задње две године доста унапређена подршком за нове постојеће методе и предикторе различитих карактеристика протеина. За потребе истраживања у оквиру ове тезе направљен је јединствени систем који интегрише ова два алата. Систем као целина омогућује једноставан приступ и коришћење свих предиктора, обраду улазних података, приказ и чување резултата на униформан начин, као и припрему података за даљу анализу и обраду. Оба алата могу да се користе самостално, или као интегрисан систем, у том случају нуде виши ниво функционалности. Алати не фаворизују појединачне предикторе, већ омогућавају коришћење више различитих предиктора на униформан начин. У оквиру новог система је направљен интерфејс за једноставан приступ ка:

(1) осам различитих метода за предвиђање неуређених региона у протеину,

(2) седам метода за идентификовање Т - ћелијских епитопа,

(3) две имплементиране методе за рачунање хидропатије региона у протеину, и

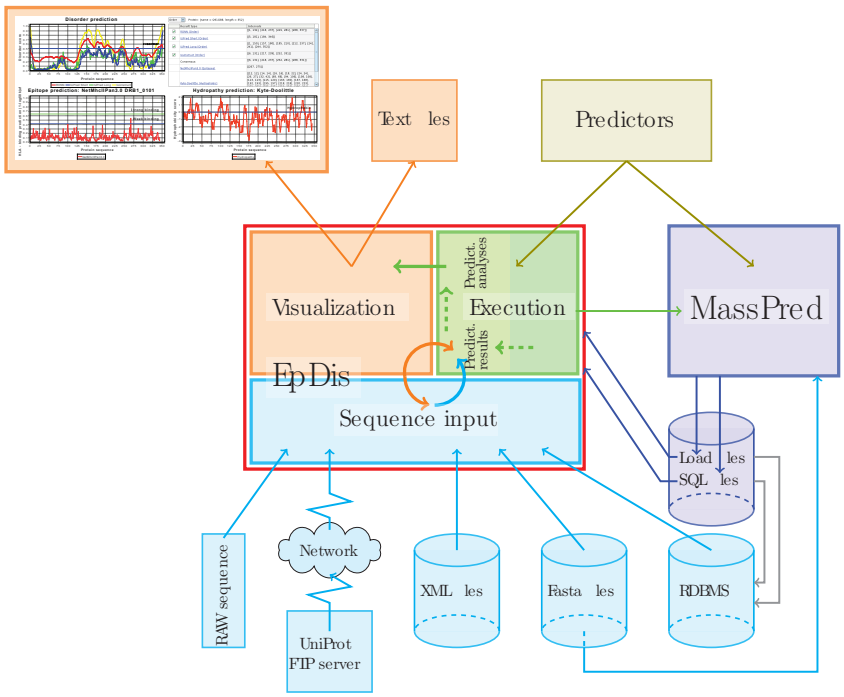
(4) једној методи за рачунање неуређених-везујућих (eng. *disorder-binding*) региона у протеину.

За приказ резултата кориснику је на располагању графички кориснички интерфејс који омогућава паралелан приказ резултата добијених неком од подржаних метода. Паралелним приказом различитих карактеристика се може установити корелација између њих нпр. веза Т - ћелијских епитопа са неуређеним односно уређеним деловима протеина, мера хидропатије епитопа као и одређеног региона у протеину. Такође у приказу резултата је омогућено укључивање и експерименталних података о неуређеним/уређеним структурама и Т - ћелијским епитопима. Укључивање експерименталних података омогућава проверу квалитета резултата појединачних предиктора. Развијени систем омогућава чување/читање израчунатих вредности (предвиђања предиктора) у релационој бази података или датотеци. Систем је отвореног кода, и јавно је доступан на локацији: `http://bioinfo.matf.bg.ac.rs/home/downloads.wafl?cat=Software`. Код се дистрибуира под MIT лиценцом отвореног кода [102]. EpDis је комплетно написан у програмском језику Java, а MassPred у програмском језику C и колекцији *bash* скрипти.

[1] Аутор MassPred алата је колега Горан Лазић mailto:chupcko@alas.matf.bg, настао је као резултат рада у оквиру Биоинформатичке истраживачке групе `http://bioinfo.matf.bg.ac.rs/home/`

Архитектура система

Систем као целина омогућује (полу)аутоматизовану инсталацију изабраног скупа предиктора, примену одабраног скупа предиктора на појединачан протеин или на произвољан скуп протеина (масовна примена), чување добијених резултата у различитим форматима погодним за даљу обраду, као и визуелни приказ резултата. Компоненте система могу да се позивају преко GUI (eng. *Graphical user interface*) или преко командног интерфејса. Организација система је приказана на слици B.1.



Слика B.1: *Приказ архитектуре EpDis-MassPred система*

Обе компоненте система се могу релативно једноставно проширити додавањем нових функционалности и укључивањем нових предиктора. Списак предиктора укључених у тренутну верзију (V1.4) је детаљно приказан у оквиру документације [44]. Главне функционалности алата су:

(1) MassPred је скуп алата који омогућавају једноставну инсталацију свих подржаних предиктора и метода, примену подржаних предиктора на скуп протеина, и примену филтера како би се добио излаз у жељеном формату. Сви предиктори за које постоји подршка се прво морају скинути са одговарајућих локација, затим се аутоматски инсталирају помоћу скрипти које

омогућава MassPred алат. Алат такође омогућава полу-аутоматизовану инсталацију свих предиктора, у смислу свих подешавања окружења за исправан рад предиктора. MassPred подржава примену предиктора над скупом протеина у оквиру једне или више датотека или директоријума. Садржаји датотека или директоријума се издвајају и прави се засебан процес за сваки пар (протеин, предиктор). Процеси се могу симултано извршавати. Након завршетка рада, MassPred прави колекцију датотека са резултатима у TSV формату, који је погодан за учитавање у релациону базу података (за детаљнији опис алата видети [44]).

(2) EpDis омогућава симултани приступ различитим предикторима, као и комбиновање и поређене резултата добијених њиховом применом над истим улазним подацима. Поред приступа предикторима, алат нуди могућност укључивања експерименталних података, који се поред резултата предикције цртају у визуелном приказу резултата и приказују се интервали преклапања резултата предикције и експерименталних података. Алат чине међусобно независне компоненте (видети слику B.1) које обезбеђују:

(а) интерфејс за унос података. Улазни подаци могу бити протеини или експериментални подаци (о секундарној структури, Т - ћелијским епитопима, MHC везујућим пептидима). Унос протеинских секвенци је омогућен на неколико начина: из датотеке (FASTA формата) са једним или више протеина; из табела релационе базе где је протеин задат својом амино киселинском секвенцом и јединственим идентификатором; преко GUI-а уносом амино киселинске секвенце протеина; директно из UniProt[2] базе преко веб сервиса. Унос експерименталних података се заснива на додавању xml датотека према унапред дефинисаној xsd шеми.

(б) обраду протеинских секвенци одређивањем Т - ћелијских епитопа различите дужине, позивањем неког од подржаних предиктора.

(ц) обраду протеинских секвенци одређивањем уређених/неуређених региона позивом неког од подржаних програма за предикцију структуре протеина.

(д) израчунавање индекса хидропатије за протеинске регионе различите дужине.

[2] http://www.uniprot.org/

(д) обраду протеинске секвенце одређивањем везујућих неуређених региона, позивом предиктора за везујуће неуређене регионе.

Главне карактеристике EpDis алата су:

- омогућава аутоматизовано извршавање свих метода за предвиђање, серијски или паралелно. У циљу бржег и ефикаснијег извршавања метода алат нуди две предефинисане стратегије за кеширање резултата. Избор стратегије кеширања се подешава у конфигурационим датотекама. Локално кеширање се користи када компонента за рад са базом није омогућена, у супротном се користи кеширање у бази података. Погодност кеширања података је посебно исплатива и значајна када због дужине протеина или саме методе која се извршава, је време извршавања предиктора јако дуго. У наредним захтевима за идентичном обрадом се добијају резултати из кеш меморије.

- омогућава чување резултата, добијених позивом неког од предиктора, директно у релациону базу података, коришћењем MassPred модула направљеним за потребе интегрисања алата и употребе MassPred алата кроз графички кориснички интерфејс. Табеле релационе базе података су дизајниране тако да омогућавају брзу претрагу извршавањем SQL упита, поређење добијених резултата или примену различитих техника истраживања података.

- све подржане методе могу бити примењене на великом броју протеина позивом MassPred алата, коме се приступа кроз MassPred компоненту EpDis алата. MassPred компонента обезбеђује прослеђивање улазних података MassPred алату, као и брз унос добијених резултата из MassPred алата коришћењем LOAD програма.

- модуларност, конфигурабилност и проширивост. Укључивање и употреба нових предиктора је врло једноставна. Довољно је додати имплементацију нове метода, која ће након ажурирања одговарајуће конфигурационе датотеке, бити препозната приликом покретања апликације.

- омогућава визуелизацију резултата подржаних предиктра, и њихово експортовање у различитим графичким форматима као што су: PNG, PDF, SVG и EPS, у високом квалитету. Графици су прилагодљиви па се тако боје и анотације могу изменити како пре тако и у току самог рада.

- омогућава издвајање резултата предиктора у текстуалне датотеке у изворним форматима дефинисаним од стране самог предиктора, као и у форму DML (INSERT, UPDATE, DELETE) наредби за релациону базу података.

Детаљнији опис са примерима употребе EpDis-MassPred система, са исцрпном анализом и теоријским тумачењем добијених резултата се могу пронаћи у [45] и [44].

Литература

[1] A. K. Abbas and A. H. Lichtman. *Cellular and molecular immunology*. Fifth ed., Pub. Saunders, 2003.

[2] C. Aggarwal and P. Yu. "Outlier detection for high dimensional data". In: *SIGMOD '01 Proceedings of the 2001 ACM SIGMOD international conference on Management of data*. Vol. 30. 2001, pp. 37 –46.

[3] C. C. Aggarwal. *Data Classification: Algorithms and Applications*. Data Mining and Knowledge Discovery Series. Chapman and Hall//CRC, 2014, p. 707.

[4] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules". In: *In Proceedings of 20th International Conference on Very Large Data Bases*. 1994, pp. 487–499.

[5] Y. Altuvia, J. A. Berzofsky, R. Rosenfeld, and H. Margalit. "Sequence features that correlate with MHC restriction". In: *Molecular Immunology* 31 (1994), pp. 1–19.

[6] S. Amnon. "Introduction to Machine Learning". In: Fall, 2008. Chap. The Kernel Trick.

[7] M. H. Andersen, L. Tan, I. Søndergaard, and et al. "Poor correspondence between predicted and experimental binding of peptides to class I MHC molecules". In: *HLA Immune Response Genetics* 556 (2001), 519–531.

[8] D. Atanackovic, N. K. Altorki, Y. Cao, and et al. "Booster vaccination of cancer patients with MAGE-A3 protein reveals long-term immunological memory or tolerance depending on priming". In: *Proc Natl Acad Sci U S A* 1055 (2008), pp. 1650–1655.

[9] M. Bhasin and G. P. S. Raghava. "A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes". In: *Journal of Biosciences* 32 (2006), pp. 31–42.

[10] M. Bhasin, H. Singh, and G. Raghava. "MHCBN: a comprehensive database of MHC binding and non-binding peptides." In: *Bioinformatics* 195 (2003), pp. 665–6.

[11] M. C. Bishop. *Pattern Recognition and Machine Learning*. Springer, Information Science and Statistics, 2006.

[12] P. Bjoern, B. Huynh-Hoa, F. Sune, and et al. "A Community Resource Benchmarking Predictions of Peptide Binding to MHC-I Molecules". In: *PLOS Computational Biology* 26 (2006), e65.

[13] B. E. Boser, I. M. Guyon, and V. N. Vapnik. "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory – COLT '92. p. 144. doi:10.1145/130385.130401. ISBN 089791497X.* (1992).

[14] P. S. Bradley, U. M. Fayyad, and C. A. Reina. *Scaling EM (Expectation Maximization) clustering to large databases*. Tech. rep. MSR-TR-98-35. Microsoft Research, 1998.

[15] S. Brin, R. Motwani, J. Ullman, and S. Tsur. "Dynamic itemset counting and implication rules for market basket data". In: *In Proceedings of ACM-SIGMOD International Conference on Management of Data, Tucson*. Vol. 26. ACM Press, Arizona, 1997, pp. 255–264.

[16] V. Brusić and D. R. Flower. "Bioinformatics tools for identifying T-cell epitopes". In: *Drug Discovery Today: BIOSILICO* 21 (2004), pp. 18–23.

[17] V. Brusic, G. Rudy, and L. Harrison. "Prediction of MHC binding peptide by using artificial neural networks". In: *In Stonier, R.J., Yu, X.S. (Eds.). Complex Systems: Mechanism of Adaptation, Amsterdam, IOS Press* (1994), 253–60.

[18] P. Carl, B. Temple, and P. Cohen. "Most nuclear systemic autoantigens are extremely disordered proteins: implications for the etiology of systemic autoimmunity". In: *Arthritis Research and Therapy* 7 (2005), pp. 1360–74.

[19] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine Learning* 203 (1995), pp. 273–297.

[20] J. D'Amaro, J. G. Houbiers, J. W. Drijfhout, and et al. "A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs". In: *Human Immunology* 431 (1995), pp. 13–18.

[21] L. D. Denise, S. Hoffman, S. Southwood, and A. Sette. "Degenerate Cytotoxic T Cell Epitopes from P. falciparum Restricted by Multiple HLA-A and HLA-B Supertype Alleles". In: *Immunity* 7 (1997), pp. 97–112.

[22] P. Dönnes and A. Elofsson. "Prediction of MHC class I binding peptides, using SVMHC". In: *BMC Bioinformatics* 325 (2002).

[23] P. Dönnes and A. Elofsson. "Prediction of MHC class I binding peptides, using SVMHC". In: *BMC Bioinformatics* 325 (2002), pp. 25–35.

[24] A. I. Doytchinova, P. Guan, and R. F. Darren. "EpiJen: a server for multistep T cell epitope prediction". In: *BMC Bioinformatics* 7131 (2006).

[25] I. Doytchinova, M. Blythe, and D. Flower. "Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201". In: *Journal of Proteomics Research* 13 (2002), pp. 263–72.

[26] J. S. Goodswen, J. P. Kennedy, and E. T. J. "Enhancing In Silico Protein-Based Vaccine Discovery for Eukaryotic Pathogens Using Predicted Peptide-MHC Binding and Peptide Conservation Scores". In: *PLOS One* 912 (2014), e115745.

[27] K. Gouda and M. J. Zaki. "Eficiently mining maximal frequent itemsets". In: *In 1st IEEE Int'l Conf. on Data Mining*. 2001, pp. 163 –170.

[28] A. S. D. Groot, A. B. N. Chinai, and et al. "From genome to vaccine: in silico predictions, ex vivo verification". In: *Vaccine* 191 (2001), pp. 4385–4395.

[29] Q. Gu, L. Zhu, and Z. Cai. "Evaluation Measures of the Classification Performance of Imbalanced Data Sets". In: vol. 51. Communications in Computer and Information Science. http://dx.doi.org/10.1007/978-3-642-04962-0_55: Springer, 2009. Chap. Computational Intelligence and Intelligent Systems, pp. 461–471.

[30] P. Guan, I. Doytchinova, C. Zygouri, and D. Flower. "MHCPred: A server for quantitative prediction of peptide-MHC binding." In: *BMC Bioinformatics* 3113 (2003), 3621–3624.

[31] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA, 2000.

[32] J. Hanley and B. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve". In: *Diagnostic Radiology* 143(1) (1982), pp. 29–36.

[33] D. Hawkins. *Identification of outliers*. Monographs on Applied Probability and Statistics. Springer Netherlands, 1980.

[34] D. Heckerman, C. Kadie, and J. Listgarten. "Leveraging Information Across HLA Alleles/Supertypes Improves Epitope Prediction". In: *Journal of Computational Biology* 146 (2006), pp. 736–46.

[35] S. Y. Ho, F. Chia, L. H. Chen, and M. H. Huang. "Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications". In: *IEEE Transactions on Systems* 341 (2006).

[36] T. K. Ho. "Random Decision Forests". In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal* (1995), 278–282.

[37] Hopp and W. K.R. "Prediction of protein antigenic determinants from amino acid sequences". In: *Proceedings of the National Academy of Sciences,PNAS* 786 (1981), 3824–3828.

[38] H. Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24 (1933), 417–441 and 498–520.

[39] J. H. Huang, H. L. Xie, J. Yan, and et al. "Using random forest to classify T-cell epitopes based on amino acid properties and molecular features". In: *Anal Chim Acta* 804 (2013), 70–75.

[40] S. C. J., K. N., S. T., and K. S. "Use of BONSAI decision trees for the identification of potential MHC class I peptide epitope motifs". In: *Pacific Symposium on Biocomputing* 4 (1999), 182–189.

[41] L. Jacob and J. P. Vert. "Efficient peptide-MHC-I binding prediction for alleles with few known binders". In: *Bioinformatics* 243 (2008), pp. 358–66.

[42] D. R. Jandrlić. "SVM and SVR-based MHC-binding prediction using a mathematical representation of peptide sequences". In: *In review proccess* (2016).

[43] D. R. Jandrlić. "The rule based classsification models for MHC binding prediction and identification of the most relevant physicochemical properties for the individual allele". In: *University thought - Publication in Natural Sciences* (2016).

[44] D. R. Jandrlić, G. M. Lazić, N. S. Mitić, and M. Pavlović. *Dokumentacija EpDis-MassPred sistema*. Tech. rep. http://bioinfo.matf.bg.ac.rs/home/downloads.wafl?cat=Software&project=EpDis: Matematički fakultet, Univezitet u Beogradu, 2016.

[45] D. R. Jandrlić, G. M. Lazić, N. S. Mitić, and M. D. Pavlović. "Software tools for simultaneous data visualization and T cell epitopes and disorder prediction in proteins." In: *Journal of Biomedical Informatics* 60 (2016), pp. 120–131.

[46] D. R. Jandrlić. *Primena tehnika istraživanja podataka na uspostavljanje korelacije izmedju neuredjenih i antigenih regiona proteina.* Magistarski rad. 2010.

[47] T. Joachims. "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization". In: *InInternational Conference on Machine Learning (ICML)* (1997).

[48] T. Joachims. "Text categorization with Support Vector Machines: Learning with many relevant features". In: *Springer, Lecture Notes in Computer Science* (2005), pp. 137–142.

[49] I. Jolliffe. *Principal Component Analysis.* 2nd ed. Springer Series in Statistics. Springer, NY,XXIX, 2002.

[50] G. Jung, B. Fleckenstein, V. der F. Mülbe, and et al. "From combinatorial libraries to MHC ligand motifs, T-cell superagonists and antagonists". In: *Journal of the International Association of Biological Standardization* 29 (2001), pp. 179–181.

[51] M. Kanehisa, S. Goto, S. Kawashima, and et al. "The KEGG resource for deciphering the genome." In: *Nucleic Acids Research* 32 (2004), D277–D280.

[52] E. Karosiene, C. Lundegaard, O. Lund, and M. Nielsen. "NetMHCcons: a consensus method for the major histocompatibility complex class I predictions". In: *Immunogenetics* 643 (2012), pp. 177–86.

[53] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley and Sons, New York, 1990.

[54] S. Kim, P. Pantel, T. Chklovskiand, and M. Pennacchiotti. "Automatically assessing review helpfulness". In: *In Proc. of EMNLP, 423–430.* (2006), pp. 423–430.

[55] M. Klemettinen, H. Mannila, P. Ronkainen, and et al. "Finding interesting rules from large sets of discovered association rules". In: *In Proceedings of 3rd International Conference on Information and Knowledge Management, pages 401-408, Gaithersburg, Maryland, ACM Press.* 1994, pp. 401–407.

[56] E. M. Knorr and R. T. Ng. “Algorithms for mining distance-based outliers in large datasets”. In: *In 24th Intl. Conf. Very Large Databases*. 1998, pp. 392–403.

[57] H. W. Kuhn and A. W. Tucker. “Nonlinear programming”. In: *Proceedings of 2nd Berkeley Symposium. Berkeley: University of California Press*. 1951, 481–492.

[58] J. Kyte and R. Doolittle. “A simple method for displaying the hydropathic character of a protein”. In: *Journam of Molecular Biology* 1571 (1982), pp. 105–132.

[59] M. V. Larsen, C. Lundegaard, K. Lamberth, and et al. “An integrative approach to CTL epitope prediction: A combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions”. In: *European Journal of Immunology* 358 (2005), 2295–2303.

[60] H. H. Lin, S. Ray, S. Tongchusak, and et al. “Evaluation of MHC class I peptide binding prediction servers: Applications for vaccine research”. In: *BMC Immunology* 98 (2007).

[61] C. Lundegaard, K. Lamberth, M. Harndahl, and et al. “NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11”. In: *Nucleic Acids Research* 36 (2008), 519–12.

[62] C. Lundegaard, I. Hoof, O. Lund, and M. Nielsen. “State of the art and challenges in sequence based T-cell epitope prediction”. In: *Immunome Research* 62 (2010).

[63] J. MacQueen. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* 1 (1967), pp. 281–297.

[64] H. Mamitsuka. “Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models”. In: *Proteins* 33 (1998), 460–474.

[65] M. B. Markus, H. Kriegel, T. N. Raymond, and J. Sander. “LOF: identifying density-based local outliers”. In: *Mining and Knowledge Discovery*. Vol. 29. n Int. Conf. on Management. 2000, pp. 93–104.

[66] M. B. Markus, H. Kriegel, T. N. Raymond, and J. Sander. “OPTICS-OF: Identifying local outliers”. In: *Mining and Knowledge Discovery*. Vol. 1704. Lecture Notes in Computer Science. In Int’l Conf. on Principles of Data. 1999. Chap. Principles of Data Mining and Knowledge Discovery, pp. 262–270.

[67] J. Martineau and T. Finin. "Delta TFIDF: An Improved Feature Space for Sentiment Analysis". In: *In Proceedings of the Third AAAI International Conference on Weblogs and Social Media, San Jose, CA, May. AAAI Press.* (2009).

[68] G. E. Meister, C. G. Roberts, J. A. Berzofsky, and A. S. D. Groot. "Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences". In: *Vaccine* 136 (1995), pp. 581–591.

[69] J. Mercer. "Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations". In: *Philosophical transactions of the royal society A, mathematical, physical and engineering sciences* 209 (1909), pp. 441–458.

[70] T. Milledge, G. Zheng, and G. Narasimhan. "An Application Of Association Rule Mining to Hla-A* 0201 Epitope Prediction". In: *ICBA* (2004).

[71] N. S. Mitić, M. D. Pavlović, and D. R. Jandrlić. "Epitope distribution in ordered and disordered protein regions - Part A. T-cell epitope frequency, affinity and hydropathy". In: *Journal of Immunological Methods* 406 (2014), pp. 83–103.

[72] N. S. Mitić, M. D. Pavlović, and D. R. Jandrlić. "T- cell Epitope Frequency in Ordered and Disordered Protein Regions". In: *DMBI.2012, International Meeting on Data Mining in Bioinformatics.* 2012.

[73] N. S. Mitić, M. D. Pavlović, D. R. Jandrlić, and S. N. Malkov. "Determining correlation of T-cell epitope location and order/disorder protein structure". In: *Theoretical Approaches to BioInformation Systems, TABIS.* 2013.

[74] L. Nanni. "Machine learning algorithms for T-cell epitopes prediction". In: *Neurocomputing* 69 (2006), pp. 866–868.

[75] C. Naugler. "Origins and relatedness of human leukocyte antigen class I allele supertypes". In: *Humman Immunology* 719 (2010), pp. 837–842.

[76] M. Nielsen, C. Lundegaard, P. Worning, and et al. "Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach". In: *Bioinformatics* 209 (2004), pp. 1388–97.

[77] M. Nielsen, C. Lundegaard, P. Worning, and et al. "Reliable prediction of T-cell epitopes using neural networks with novel sequence representations". In: *Protein Science* 125 (2003), pp. 1007–1017.

[78] G. Paltoglou and M. Thelwall. "A study of Information Retrieval weighting schemes for sentiment analysis". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. for Computational Linguistics. Uppsala, Sweden, 11-16 July 2010, 1386–1395.

[79] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up? sentiment classification using machine learning techniques." In: *In Proc. of EMNLP 2002* (2002), pp. 79–86.

[80] K. C. Parker, M. A. Bednarek, and J. E. Coligan. "Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains". In: *The Journal of Immunology* 1521 (1994), pp. 163–75.

[81] A. Patronov and I. Doytchinova. "T-cell epitope vaccine design by immunoinformatics". In: *Open Biology* 31 (2012).

[82] M. D. Pavlović, D. R. Jandrlić, and 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". In: *Journal of Immunological Methods* 407 (2014), pp. 90–107.

[83] K. Pearson. "On Lines and Planes of Closest Fit to Systems of Points in Space". In: *Philosophical Magazine* 211 (1901), 559–572.

[84] K. Peng, P. Radivojac, S. Vučetić, and et al. "Length-dependent prediction of protein intrinsic disorder". In: *BMC Bioinformatics* 7208 (2006).

[85] B. Peters, H. H. Bui, S. Frankild, and et al. "A community resource benchmarking predictions of peptide binding to MHC-I molecules". In: *PLoS Comput Biology* 26 (2006), e65.

[86] G. Pingping, I. A. Doytchinova, A. Walshe, and et al. "Analysis of Peptide-Protein Binding Using Amino Acid Descriptors: Prediction and Experimental Verification for Human Histocompatibility Complex HLA-A*0201". In: *Journal of Medical Chemistry* 4823 (2005), pp. 7418–25.

[87] C. J. Platt. *A Fast Algorithm for Training Support Vector Machines*. Tech. rep. Microsoft Research, 1998.

[88] J. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization in Advances in Kernel Methods – Support Vector Learning , B. Scholkopf, C. Burges, A. Smola, eds.* MIT Press, 1998.

[89] A. W. Purcell and J. Gorman. "Immunoproteomics, Mollecular and cellular proteomics". In: *Mol Cell Proteomics.* 33 (2004), pp. 193–208.

[90] L. R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proc IEEE* 77 (1989), pp. 257–86.

[91] M. Rajapakse, B. Schmidt, L. Feng, and V. Brusic. "Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms". In: *BMC Bioinformatics* 8459 (2007).

[92] S. Ramaswamy, R. Rastogi, and K. Shim. "Efficient algorithms for mining outliers from large data sets". In: *In Int'l Conference on Management of Data*. Vol. 29. 2000, pp. 427–438.

[93] H. G. Rammensee, J. Bachmann, N. P. N. Emmerich, and et al. "SYFPEITHI: database for MHC ligands and peptide motifs". In: *Immunogenetics* 50 (1999), pp. 213–213.

[94] A. P. Rechea, J. P. Gluttinga, and L. E. Reinherz. "Prediction of MHC class I binding peptides using profile motifs". In: *Human Immunology* 639 (2002), pp. 701–709.

[95] K. Roomp, I. Antes, and T. Lengauer. "Predicting MHC class I epitopes in large datasets". In: *BMC Bioinformatics* 1190 (2010), pp. 1471–2105.

[96] M. Sathiamurthy, H. Hickman, J. Cavett, and et al. "Population of the HLA Ligand Database". In: *Tissue Antigens* 61 (2003), 12–19.

[97] C. Schönbach, L. Y. J. Koh, R. D. Flower, and et al. "FIMM, a database of functional molecular immunology: update 2002". In: *Nucleic Acids Research* 301 (2002), 226–229.

[98] J. Sidney, P. Bjoern, F. Nicole, and et al. "HLA class I supertypes: a revised and updated classification". In: *BMC Immunology* 91 (2008).

[99] J. A. Smola and B. Scholkopf. *A Tutorial on Support Vector Regression*. Tech. rep. NC2-TR-1998-030. NeuroCOLT 2, 1998.

[100] S. A. Smola and K. Muller. "Nonlinear component analysis as a kernel eigenvalue problem". In: *Neural Computation* 10(5) (1998), 1299–1319.

[101] P.-N. Tan and et al. *Introduction to data mining*. Pearson Education India, 2007.

[102] M. I. of Technology. *MIT License, Permissive free software license*. Massachusetts Institute of Technology. https://opensource.org/licenses/MIT.

[103] F. Tian, L. Yang, F. Lv, and et al. "In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach". In: *Amino Acids* 36 (2009), p. 535.

[104] J. C. Tong, T. Tan, and S. Ranganathan. "Methods and protocols for prediction of immunogenic epitopes". In: *Briefinga in Bioinformatics* 8 (2006), pp. 96–108.

[105] C. W. Tung and S. Y. Ho. "POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties". In: *Bioinformatics* 238 (2006), pp. 942–949.

[106] K. Udaka, K. Wiesmüller, S. Kienle, and et al. "Tolerance to Amino Acid Variations in Peptides Binding to the Major Histocompatibility Complex Class I Protein H-2Kb". In: *The Journal of Biological Chemistry* 27041 (1995), 24130–24134.

[107] V. N. Uversky and A. K. Dunker. "Understanding protein non-folding". In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 18046 (2010), 1231–1264.

[108] V. Vapnik and A. Lerner. "A note on one class of perceptrons". In: *Automation and Remote Control* 143 (1963), 774–780.

[109] G. Vogt, T. Etzold, and P. Argos. "An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited." In: *Journal of Molecular Biology* 249 (1995), pp. 816–831.

[110] G. Wahba. "Soft and Hard Classification by Reproducing Kernel Hilbert Space Methods". In: *Proceedings of the National Academy of Sciences* 9926 (2002), 16524–16530.

[111] G. Wahba. "Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV". In: *Advances in Kernel Methods*. Ed. by B. Schölkopf, C. J. C. Burges, and A. J. Smola. http://dl.acm.org/citation.cfm?id=299094.299099: MIT Press, Cambridge MA, USA, 1999, pp. 69–88.

[112] J. Wan, W. Liu, Q. Xu, and et al. "SVRMHC prediction server for MHC-binding peptides". In: *BMC Bioinformatics* 7463 (2006).

[113] C. Whitelaw, N. Garg, and S. Argamon. "Using appraisal groups for sentiment analysis". In: *Proc. of the 14th ACM international conf. on Information and knowledge management* (2005), 625–631.

[114] I. H. Witten, E. Frank, and A. H. Mark. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, 2000.

[115] G. G. Yardımcı, A. Kucukural, and Y. Saygin. "Modified Association Rule Mining Approach for the MHC-Peptide Binding Problem". In: Lecture Notes in Computer Science (2006), pp. 165–173.

[116] M. J. Zaki. "Scalable algorithms for association mining". In: *IEEE Transactions on Knowledge and Data Engineering* 123 (2000), pp. 372–390.

[117] M. J. Zaki and C. Hsiao. "ChARM: An eficient algorithm for closed itemset mining". In: *In 2nd SIAM International Conference on Data Mining*. 2002, pp. 457–473.

[118] H. Zhang, C. Lundegaard, and M. Nielsen. "Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods". In: *Bioinformatics* 251 (2009), pp. 83–89.

[119] S. Zhu, K. Udaka, J. Sidney, and et al. "Improving MHC binding peptide prediction by incorporating binding data of auxiliary MHC molecules". In: *Bioinformatics* 2213 (2006), pp. 1648–1655.

Биографија аутора

Мр Даворка Јандрлић рођена је 12. фебруара 1981. године у Дубровнику. Основну школу и гимназију завршила је у Херцег Новом. Школске 1999/2000. године уписала је Математички факултет у Београду (смер Рачунарство и информатика), и дипломирала школске 2003/2004. године са просечном оценом 9,10. Школске 2004/2005. године је уписала магистарске студије на Математичком факултету, смер Рачунарство и информатика. Последипломске магистарске студије је завршила 2010. године одбраном магистарске тезе под насловом "Примена техника истраживања података на одређивање корелације између неуређених и антигених региона протеина" под руководством проф. др Ненада Митића. Од 2004. је запослена на Машинском факултету Универзитета у Београду као асистент приправник, а од 2010. као асистент на катедри за математику. До сада је држала вежбе из следећих предмета: Програмирање, Рачунарски алати, C-C++, Веб пројектовање у машинству и Објектно оријентисано програмирање и Јава. Основне области интересовања су јој базе података, развој и примена техника истраживања података и биоинформатика. Објавила је већи број научних радова и учествовала на неколико међународних и домаћих конференција. Учествовала је на неколико међународних и домаћих конференција.

Прилог 1.

Изјава о ауторству

Потписани-а Даворка Јандрлић

број индекса _____

Изјављујем

да је докторска дисертација под насловом

Примена правила придруживања и метода подржавајућих вектора за предвиђање Т - ћелијских епитопа

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 01.04.2016

Јандрлић

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Даворка Јандрлић

Број индекса _____

Студијски програм Рачунарство и информатика

Наслов рада Примена правила придруживања и метода подржавајућих вектора за предвиђање Т - ћелијских епитопа

Ментор проф др. Ненад Митић

Потписани/а Даворка Јандрлић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 01.04.2016

ДЈандрлић

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Примена правила придруживања и метода подржавајућих вектора за предвиђање Т - ћелијских епитопа

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 01.04.2016

Јандрлић

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.