Универзитет у Београду

Математички факултет

Слободан Војводић

Одређивање правила придруживања између појављивања карактеристичних ниски у туморским антигенима

Мастер рад

Септембар 2016.

Ментор:	др Ненад Митић, Математички факултет, Универзитет у Београду
Чланови комисије:	др Саша Малков, Математички факултет, Универзитет у Београду
	др Мирјана Павловић, Институт за општу и физичку хемију, Универзитет у Београду
Датум одбране:	_____

Садржај

1. Увод

Имунотерапија има за циљ да активира или сузбије реакцију имуног система на одређене антигене и има примјену у терапији болести. У терапији канцера имуни систем се активира да би уништавао туморске ћелије. Имуни систем препознаје туморске ћелије на основу присуства одговарајућих антигена у њима. Присуство протеинских туморских антигена од стране Т-ћелијског имуног система се препознајеје тако што се препознају карактеристични кратки пептиди – епитопи (ниске молекула аминокиселина), који се налазе у њиховом саставу и који бивају изложени на ћелијској мембрани.

Откривање антигена и карактеристичних ниски у њиховом саставу, које могу изазвати имуни одговор, биолошким методама је временски захтјеван и скуп процес. Зато се као допуна овим методима користе рачунарске методе, које омогућавају откривање ових ниски великом брзином и са високом прецизношћу[1]. Избор епитопа који се користе за прављење антитуморских вакцина одређен је бројним биолошким карактеристикама, али примарно зависи од структуре протеинског антигена коме епитоп припада. Нађено је да су ниске епитопа повезане са концептом уређене и неуређене структуре протеина[2]. Насупрот ранијим схватањима, показано је да протеини који немају уређену терцијарну структуру имају значајну улогу у мрежи интеракција међу протеинима у физиолошким процесима[3][4]. Такође, нађено је да су неке болести као што су канцер, аутоимуне болести, алергије и неуродегенеративне болести повезане са функцијом неуређених протеина.

У раду се дефинишу правила придруживања која могу да се користе за одређивање различитих карактеристичних ниски у зависности од њихове врсте, околине у којој се налазе и састава. У раду се такође испитује и ефикасност различитих приступа при одређивању правила придруживања у зависности од врсте карактеристичних ниски.

2. Протеини

2.1 Структура

Аминокиселине су једињења која садрже амино-групу (-NH2) и карбоксилну групу (-COOH). Уколико су обе групе везане за исти, α атом угљеника, тада су у питању α-аминокиселине. α атом угљеника је асиметричан (хиралан) код свих аминокиселина које улазе у састав протеина осим код глицина. Због тога ове аминокиселине могу да се јаве у облику два енантиомера: L и D. Аминокиселине међусобно могу да се везују пептидним везама – карбоксилна група једне аминокиселине се везује за амино-групу друге киселине, при чему се ослобађа молекул воде. Пептиди су једињења мале молекулске масе сачињени од низа аминокиселина везаних пептидним везама. Полипептиди и протеини су већи молекули, који у низу садрже више од 100 аминокиселина. У изградњи протеина учествује 20 L α-аминокиселина: аргинин (R), хистидин (H), леуцин (L), изолеуцин (I), лизин (K), метионин (M), фенилаланин (F), треонин (T), триптофан (W), валин (V), аланин (A), аспарагин (N), аспарагинска киселина (D), цистеин (C), глутаминска киселина (E), глутамин (Q), глицин (G), пролин (P), серин (S) и тирозин (Y).

Протеини су кључни градивни елементи сваког живог организма. Они учествују у свим ћелијским и међућелијским процесима. Структура протеина је одређена редоследом аминокиселина у полипептидном ланцу и од ње директно зависи функција протеина. Протеини имају четири структруна нивоа који одређују њихов изглед у простору: примарни, секундарни, терцијарни и кватернарни.

- Примарну структуру чини редослијед аминокиселина у полипептидном ланцу.
- Секундарну структуру дефинишу обрасци водоничне везе између пептидних веза централног ланца полипептида. Секундарна структура представља просторни распоред сусједних аминокиселина у ланцу. Најчешће секундарне структуре су α-хеликси и β-равни.
- Терцијарна структуру чини тродимензионални распоред атома у једном протеину.
- Уколико се протеин састоји из више полипептида, тада кватернарна структура представља просторни распоред његових полипептида.

2.2. Уређеност структуре протеина

До средине 20. вијека научници су вјеровали да је јединственост терцијарне структуре протеина предуслов за његову функционалност. Међутим, током посљедњих деценија, откривени су биолошки активни протеини који немају стабилну секундарну или терцијарну структуру. Овакви протеини су названи неуређеним протеинима (енг. *intrinsically disordered proteins*). Ови протеини у потпуности, или на свом дијелу показују неуређеност, тј. позиције атома се мијењају у току времена.

Нормална физиологија живих организама је заснована на скупу високо координисаних протеинских интеракција. Координација је контролисана препознавањем јединственог идентификационог региона који се често налазе унутар неуређених региона протеина. Зато су многи неуређени протеини укључени у регулацију, препознавање, сигнализирање и контролу

разних догађаја у ћелији. Једна од јединствених функционалних карактеристика неуређених протеина је њихова способност да се вежу са више других макромолекула – партнера. У мрежи протеинских интеракција, овакви протеини представљају чворове (енг. *hubs*) и од кључног су значаја за њену функционалност и стабилност[4].

Уређеност протеина зависи од редослиједа аминокиселина[3]. Утврђено је да ниска средња вриједност хидрофобности, као и велика количина наелектрисања код аминокиселина промовишу неуређеност полипептидног ланца који сачињавају. Зато се, по утицају на уређеност ланца, аминокиселине дијеле на:

- Аминокиселине које промовишу неуређеност: E, K, R, G, Q, S, P, A.
- Аминокиселине које промовишу уређеност: L, V, W, I, Y, C, F, N.

Постоји више експерименталних метода помоћу којих може да се утврди уређеност протеина односно региона у протеину. Ови методи се ослањају на разлике у молекулским величинама, густинама и хидродинамичком отпору. Најчешће се користе:

- Нуклеарно магнетна резонантна спектроскопија
- Дифракциона кристалографија X-зрацима
- Циркуларни дихроизам

Постоји више од 20 биофизичких и биохемијских метода које су фокусиране на препознавање неуређених региона у протеинима. Ове методе су скупе, временски захтјевне и често је потребно више од једне да би се неуређени протеини потпуно окарактерисали. Неуређени региони у протеинима често имају заједничка својства. На тој чињеници је развијен велики број рачунарских метода за предвиђање неуређених региона. Ови методи се по приступу могу подијелити на:

- Методе код којих се предивиђање заснива на самој секвенци аминокиселина коришћењем техника машинског учења. У ову групу спадају RONN, DISOPRED, DisEMBL, VSL2.
- Методе код којих се предвиђање заснива на физичким особинама аминокиселина. У ову групу спадају FoldUnFold, FoldIndex, IUPred, PONDR, PreLINK.
- Методе које комбинују неколико алгоритама. У ову групу спадају MD, GeneSilico MetaDisorder, PONDR-FIT, metaPrDOS.

3. Имуни одговор

Имуност је способност организма да препозна патогене и да их елиминише на различите начине. Патогени су најчешће микроорганизми или токсини. Супстанце које изазивају имуни одговор организма се називају антигенима (енг. *antigene* од првобитног *anti*body *gen*erator). Имуни систем може да се подијели на неспецифичан (урођени) имуни систем и специфичан (стечени) имуни систем.

Урођени имуни систем представља прву линију одбране организма и дјелује без претходног сусрета организма са патогеном. Разликује стране од властитих материја, али не препознаје врсту страног агенса.

Стечени имуни систем развија специфичан имуни одговор на сваки антиген и има способност да памти и препозна сваки нови контакт са одређеним антигеном. Имуни одговор може бити хуморални и ћелијски. Хуморални одговор настаје при појави антигена у ванћелијском простору, што се најчешће дешава у присуству бактерија, паразита или токсина, док ћелијски настаје у случају појаве антигена у самим ћелијама, што се дешава у случају појаве вируса или тумора.

Све ћелије организма осим оних без језгра презентују сопствене антигене или антигене патогена који продру у њих (попут вируса). Презентација се врши тако што се протеини разлажу на мање секвенце, а затим се неки од тих пептида, тзв. епитопи везују за МНС (енг. *major histocompatibility complex*) молекуле I класе који се налазе у ћелијској мембрани. Постоје и ћелије које су специјализоване за презентацију антигена, које се називају професионалним антиген-презентујућим ћелијама (енг. *professional antigen-presienting cells*). Међу њих спадају Б ћелије, дендритске ћелије и неки макрофаги. Оне презентирају антигене, углавном стране или из ванћелијског простора, уз помоћ МНС молекула II класе као и других ко-стимулаторних молекула.

Т ћелије у својој мембрани имају рецепторе који реагују на неке од презентираних епитопа, и тада, у зависности од тога ког су типа, реагују на одговарајући начин. Наивне цитотоксичне Т ћелије (CD8+ Т ћелије) се активирају уколико се њихов рецептор веже за епитоп везан за МНС молекуле I класе. Оне се потом дијеле на велики број ефекторних ћелија које трагају за ћелијама које презентирају пронађени епитоп-МНС комплекс да их униште. Највећи број ових ћелија умире након престанка инфекције, оне које преживе постају меморијске Т ћелије спремне да брзо реагују уколико се исти антиген поново појави. Да би се избјегла уништавања превеликог броја ћелија, цитотоксичне Т ћелије се активирају у строго контролисаним условима. Неопходан је или јак сигнал са ћелије, или додатна сигнализација са помоћних Т ћелија (CD4+ ћелија). Рецептор помоћних Т ћелија реагује на епитопе везане за МНС молекуле II класе. Уколико се веже за овај комплекс, помоћна ћелија почиње да производи супстанце којима утиче на понашање других ћелија имуног система, па и на ћелију која јој је презентовала антиген, нпр. Б ћелију.

Б ћелије посједују Б ћелијске рецепторе који реагују на антигене. За разлику од Т ћелијских рецептора који препознају антиген након разлагања, Б ћелијски рецептори могу да препознају антигене у његовој оригиналној форми. Након препознавања антигена, Б ћелија га обухвата, разлаже и презентира уз помоћ МНС молекула II класе. Уколико помоћна Т ћелија препозна овај комплекс, она стимулише Б ћелију да се диференцира у извршну плазма ћелију. Плазма ћелије живе кратко и луче антитијела. Ова антитијела се везују за антигене и на тај начин их чине лакшим метама за фагоците. Око 10% плазма ћелија преживљава и постају меморијске Б ћелије које брзо реагују лучењем антитијела ако се специфичан антиген поново појави.

Код човјека се МНС молекули још називају и HLA (енг. *human leukocyte antigen*). Постоји 6 врста молекула HLA I класе: A, B, C, E, F, G, и 5 врста молекула HLA II класе: DP, DQ, DR, DM, DO. HLA молекули I класе се синтетишу на основу HLA-A, HLA-B, HLA-C, HLA-E, HLA-F или HLA-G генетских локуса. HLA молекули II класе се састоје из α и β ланца, и за сваки од њих је потребан по један локус. Основ за синтезу DP и DQ молекула су HLA-DPA1, HLA-DPB1 и HLA-DQA1, HLA-DQB1 локуси. Основ за синтезу α ланца DR молекула је HLA-DRA локус, а за синтезу β ланца основ су HLA-DRB1, HLA-DRB3, HLA-DRB4 и HLA-DRB5 локуси, при чему једна особа може да да синтетише само 3 од ова 4 ланца.

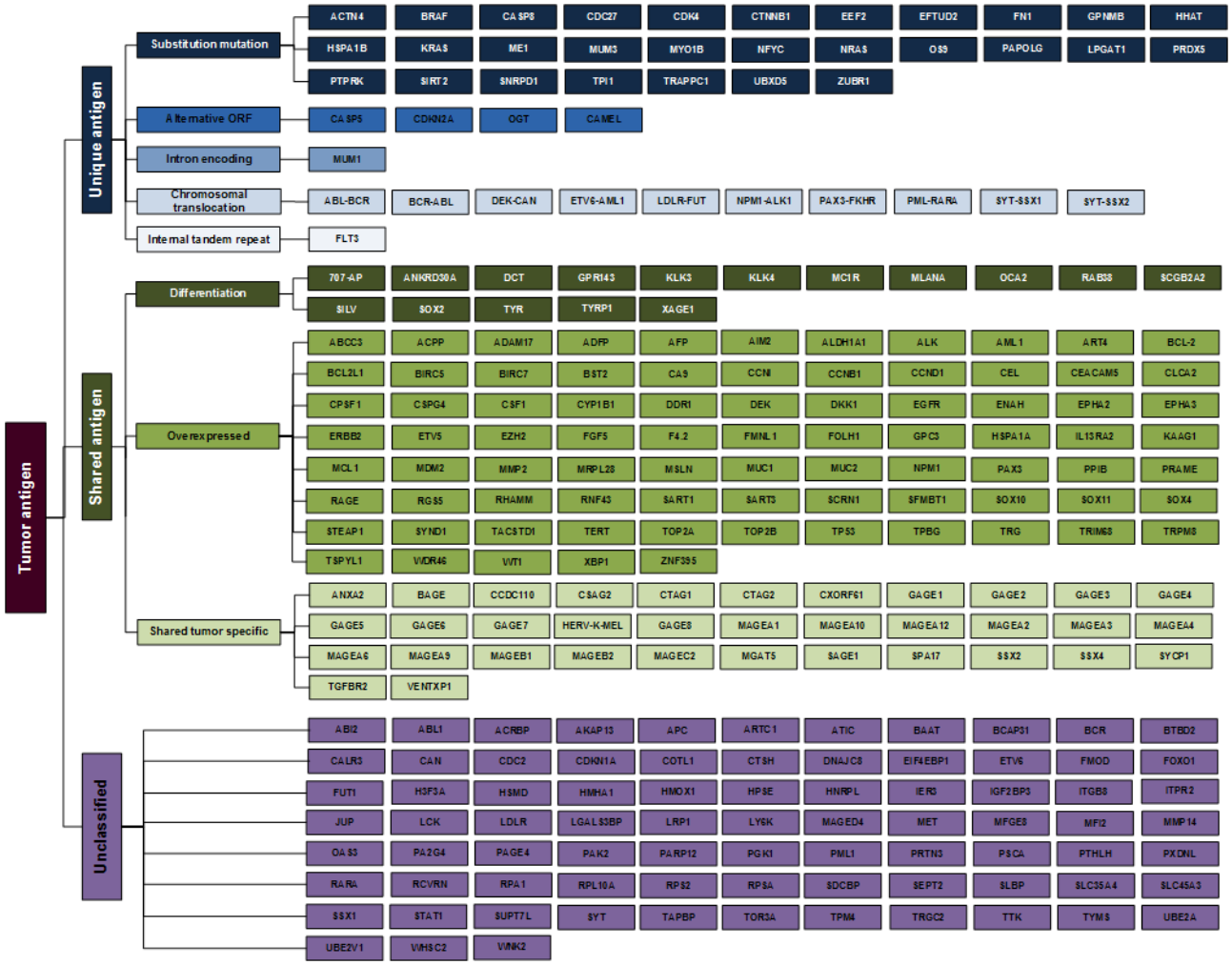
На генетским локусима HLA молекула може да се нађе велики број различитих алела, тј. HLA молекули спадају међу генетски најваријабилније у људском организму. Ознака за алел се састоји из ознаке за локус звјездицом одвојене од парног броја цифара, при чему цифре које се налазе ближе крају ниске специфичније одређују алел. Алели који имају заједичке прве четири цифре су синонимни, тј основ су за синтезу идентичних протеина, тј. HLA молекула.

Реакција антитијела крвног серума особе на ћелије друге особе може да се искористи за типизирање HLA молекула. Овакви типови се називају серотиповима[5]. Њихов број је ограничен бројем познатих антитијела, а развијањем метода за подизање осјетљивости крвног серума, повећан је и број серотипова које је могуће открити.

Испитивањем појединачних HLA молекула I класе показало се да постоје фамилије ових молекула насталих на основу различитих алела који су склони везивању за исте пептиде. Овакве фамилије молекула се називају супертиповима.

4. Туморски антигени

Имунотерапија тумора има за циљ активирање адаптивног имуног система да уништава тумор и да спријечава његову поновну појаву. Цитотоксичне Т ћелије које препознају туморске антигене би имале кључну улогу у овоме. Проблем је у томе што постоји могућност развоја аутоимунитета тј. да цитотоксичне Т ћелије почну да уништавају ћелије нормалног ткива као посљедица смањене толеранције на антигене које оне презентују.



Слика 4.1 Класификација туморских антигена[6]

Туморски антигени се дијеле на[7][8]:

- јединствене (енг. *unique*) антигене – они настају као посљедица мутације гена који се испољавају у свим ћелијама. Овакви антигени могу да играју битну улогу у развијању природног анти-туморског имуног одговора, али је тешко базирати имунотерапију на њима пошто су специфичне за пацијента.
- заједнички (енг. *shared*) антигени – антигени који се јављају код многих независних тумора. Могу додатно да се подијеле на:
  - заједничке антигене специфичне за тумор (енг. *shared tumor-specific*) – они

сејављају само у туморима, а у нормалном ткиву једино у трофобласима плаценте и у герминалним (eng. *germ*) ћелијама тестиса. Пошто ћелије ових нормалних ткива не излажу МНС молекуле I класе на својој мембрани, за ове антигене се може рећи да су строго специфичне за тумор. Гени на основу којих се ови антигени синтетишу се називају и канцер-тестис генима.

- развојни (енг. *differentiation*) антигени – јављају се и у туморима и у нормалном ткиву из којег је тумор настао. Могућности имунотерапије у овом случају су ограничене због могућности појаве имуног одговора и на нормално ткива. Имунотерапија долази у обзир једино у случају да је нормално ткиво непотребно, или да ће бити уклоњено, као на примјер у случају канцера простате.
- прекомјерно испољени (енг. *overexpressed*) – антигени који се јављају у разним ткивима, али се нарочито испољавају у туморима. У овој групи је тешко предвидјети безбједност имунотерапије, тј. да ли ће и смањена испољеност антигена у нормалном ткиву изазвати имуни одговор.

- вирални антигени – антигени који настају усљед вирусне инфекције ткива. Неки вируси као што су хумани папилома вирус или Епштајн-Баров вирус су повезани са развојем тумора код људи. Ови антигени имају висок потенцијал као мете имунотерапије.

5. Правила придруживања

Нека је I скуп неких објеката i_1, i_2,... ,i_d и нека је T торка сачињена од t_1, t_2,..., t_N подскупова скупа I. Елементи скупа I се још називају ставкама, а елементи торке T се називају трансакцијама. За скуп ставки X се може рећи да припада трансакцији t_i уколико је $X \subseteq t_i$. Скуп ставки има сљедећа својства.

- Бројач подршке (eng. *support count*) - σ(X) - број трансакција којима скуп ставки припада.
- Подршка - s(X) - учесталост појављивања скупа ставки у трансакцијама:

  $s(X) = \sigma(X) / N$
- За X се може рећи да је чест уколико је $s(X) \geq minsup$, гдје је minsup неки одабрани праг.

Правила придруживања су правила облика X -> Y, гдје су X и Y дисјунктни подскупови скупа I такви да у T постоји t_i такво да је $X \cup Y \subseteq t_i$. Тада се може рећи да у трансакцији t_i важи правило придруживања X -> Y. Значење правила придруживања је: уколико се у једној трансакцији нађе скуп ставки X, са одређеном вјероватноћом ће се у истој трансакцији наћи и скуп ставки Y. Најважнија својства правила придруживања су:

- подршка – количник броја трансакција за које важи правило придруживања и укупног броја трансакција:

  $s(X \rightarrow Y) = \sigma(X \cup Y) / N$

  Ова мјера је значајна јер, уколико је подршка неког правила мала, постоји могућност да су се неке ставке јавиле заједно случајно, а такође може да упућује на везе међу ставкама које нису од значаја.
- поузданост – количник броја трансакција које садрже ставке и из X и из Y и броја свих трансакција које садрже ставке из X.

  $c(X \rightarrow Y) = \sigma(X \cup Y) / \sigma(X)$

  Поузданост представља вјероватноћу јављања Y ако се појавио X – P(X|Y).

Алгоритми за проналажење правила придруживања проналазе сва правила која имају подршку која је већа или једнака од изабраног прага minsup, и поузданост већу од изабраног прага minconf. Ови алгоритми често као резултат дају велики број правила. Задатак истраживања правила придруживања је проналажење правила која откривају интересантне везе у скупу ставки. Често велики број правила није од значаја. Таква су, на примјер, правила која повезују независне ставке, или правила која повезују ставке чија је повезаност већ позната. Да би се олакшао посао потраге за интересантним правилима, постоје објективне мјере интересантности правила које су засноване на статистикама над подацима. Постоји велики број оваквих мјера са различитим својствима и примјенама у различитим условима. Примјери оваквих мјера су подршка, поузданост и корелација.

Нека је A -> B правило придруживања са сљедећом табелом контингената:

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1x}
$\overline{A}$	f_{01}	f_{00}	f_{0x}
	f_{x1}	f_{x0}	N

Нека је M(A,B) мјера интересантности тог правила. Према Пиатетски-Шапироу[10], добра мјера мора да има сљедеће 3 особине:

- M(A, B) = 0 ако су A и B статистички независне.
- M(A, B) монотоно расте са P(AB) када P(A) и P(B) остају непромјењене.
- M(A, B) монотоно опада са P(A) односно са P(B) када P(A, B) и P(B) односно P(A) остају непромјењене.

Важна својства која мјере могу да имају[9]:

- симетричност – M(A, B) = M(B, A) – за симетричне мјере није од значаја који скуп ставки је придружен којем. Примјери симетричних мјера су подршка, лифт, збирна снага, косинусна, Жакардова мјера, итд. Асиметричне мјере су поузданост, убјеђење, Лапласова мјера, J-мјера, итд.
- инваријантност у односу на инверзију – M(A, B) = M($\overline{A}$, $\overline{B}$) – мјере које посједују ово својство нису погодне у случају асиметричних података, јер дају једнак значај заједничком одсуству и заједничком присуству скупова ставки у трансакцијама. Примјер мјера инваријантних у односу на инверзију: однос шанси, збирна снага, капа, корелација. Мјере које нису инваријантне у односу на инверзију: Пиатетски-Шапироова, косинусна, Жакардова и лифт.
- инваријантност у односу на скалирање – M([f_{11} , f_{10} ; f_{10} , f_{00}]) = M([$k_1k_3f_{11}$, $k_1k_4f_{10}$; $k_2k_3f_{01}$, $k_2k_4f_{00}$]) за нека k_1, k_2, k_3, $k_4 > 0$ – Интересантност правила не би требало да се мијења уколико се промијени број трансакција тако да се скалира врста или колона табеле контигенције. На примјер, уколико се број трансакција у којима се јавља A повећа тако да однос f_{11}/f_{10} остане исти, мјера интересантности правила A -> B би требало да остане иста. Једино однос шанси има ово својство.
- инваријантност у односу на додавање празних слогова – M([f_{11} , f_{10} ; f_{10} , f_{00}]) = M([f_{11} , f_{10} ; f_{01} , f_{00}+r]) - На примјер, уколико правило указује на заједничко појављивање двије ријечи у документима, интересантност правила може да се измијени додавањем докумената која не садрже ни једну од ове двије ријечи. Мјере које имају ово својство су косинусна мјера и Жакардов коефицијент, а немају га лифт, Пиатетски-Шапироова мјера, коефицијент корелације и однос шанси.

Симетричне мјере:

Мјера (Симбол)	**Дефиниција**
Корелација (ϕ)	$\frac{Nf_{11}-f_{1x}-f_{x1}}{\sqrt{f_{1x}f_{x1}f_{0x}f_{x0}}}$
Однос шанси (енг. *Odds ratio*) (α)	$(f_{11}f_{00})/(f_{10}f_{01})$
Капа (κ)	$\frac{Nf_{11}+Nf_{00}-f_{1x}f_{x1}-f_{0x}f_{x0}}{N^2-f_{1x}f_{x1}-f_{0x}f_{x0}}$
Интересовање (eng. *interest*) (I)	$(Nf_{11})/(f_{1x}f_{x1})$
Косинусна (IS)	$(f_{11})/(\sqrt{f_{1x}f_{x1}})$
Пиатетски-Шапироова (PS)	$\frac{f_{11}}{N}-\frac{f_{1x}f_{x1}}{N^2}$
Збирна снага (eng. *collective strength*) (S)	$\frac{f_{11}+f_{00}}{f_{1x}f_{x1}+f_{0x}f_{x0}}\cdot\frac{N-f_{1x}f_{x1}-f_{0x}f_{x0}}{N-f_{11}-f_{00}}$
Жакардова мјера (ζ)	$f_{11}/(f_{1x}+f_{x1}-f_{11})$
Потпуна поузданост (енг. *All-confidence*) (h)	$min\left[\frac{f_{11}}{f_{1x}},\frac{f_{11}}{f_{x1}}\right]$

Асиметричне мјере:

Мјера (симбол)	**Дефиниција**
Гудмен-Крускалова мјера (λ)	$(\sum_j max_k f_{jk}-max_k f_{xk})/(N-max_k f_{xk})$
Заједничка информација (енг. *mutual information*) (M)	$(\sum_i\sum_j \frac{f_{ij}}{N}\log\frac{Nf_{ij}}{f_{ix}f_{xj}})/(-\sum_i \frac{f_{ix}}{N}\log\frac{f_{ix}}{N})$
J-мјера (J)	$\frac{f_{11}}{N}\log\frac{Nf_{11}}{f_{1x}f_{x1}}+\frac{f_{10}}{N}\log\frac{Nf_{10}}{f_{1x}f_{x0}}$
Гини индекс (G)	$\frac{f_{1x}}{N}\cdot\left[\left(\frac{f_{11}}{f_{1x}}\right)^2+\left(\frac{f_{10}}{f_{1x}}\right)^2\right]-\left(\frac{f_{x1}}{N}\right)^2$ $+\frac{f_{0x}}{N}\cdot\left[\left(\frac{f_{01}}{f_{0x}}\right)^2+\left(\frac{f_{00}}{f_{0x}}\right)^2\right]-\left(\frac{f_{x0}}{N}\right)^2$
Лапласова мјера (L)	$(f_{11}+1)/(f_{1x}+2)$
Убјеђење (енг. *conviction*) (V)	$(f_{1x}f_{x0})/(Nf_{10})$
Фактор сигурности (енг. *certainty factor*) (F)	$(\frac{f_{11}}{f_{1x}}-\frac{f_{x1}}{N})/(1-\frac{f_{x1}}{N})$

Додата вриједност (енг. *added value*) (AV)	$\frac{f_{11}}{f_{1x}} - \frac{f_{x1}}{N}$

6. Циљ рада

Циљ овог рада је да се искористе правила придруживања да би се пронашле повезаности међу различитим својствима туморских антигена, њихових карактеристичних ниски, и аминокиселина које их сачињавају. Међу значајна својства спадају припадност карактеристичних ниски уређеним или неуређеним регионима антигена, њихова хидрофобност, положај у нисци антигена. Занимљиво је провјерити способност везивања ниске за МНС молекуле I или II класе, алел на основу којих је тај МНС молекул изграђен, њихов серотип или супертип. Такође, биће испитано постојање повезаности ових својстава са појединим типовима туморских антигена.

Одређивање позиције Т-ћелијских епитопа у дијеловима протеина који могу да формирају одређене биохемијске или структурне обрасце, могло би да допринесе бољем предвиђању епитопа, а самим тим и развоју анти-туморских вакцина.

7. Материјал

7.1 TANTIGEN

Подаци о туморским антигенима су преузети са сајта TANTIGEN[6] базе 20.01.2014. За преузимање података коришћени су Perl скриптови, са LWP::Simple и HTML::TreeBuilder модулима. На првој страници овог сајта се налази слика на којој су категорисани антигени, и кликом на одређени дио слике отвара се страница одговарајућег антигена. То је постигнуто коришћењем map и area ознака. Примјер:

```
<mapname="antigen">
  <area shape="rect" coords="250,6,314,24" alt="ACTN4" href="/cvccgi/tadb/second.pl?name=ACTN4">
  <area shape="rect" coords="320,6,382,24" alt="BRAF" href="/cvccgi/tadb/second.pl?name=BRAF">
  <area shape="rect" coords="390,6,452,24" alt="CASP8" href="/cvccgi/tadb/second.pl?name=CASP8">
  <area shape="rect" coords="460,6,522,24" alt="CDC27" href="/cvccgi/tadb/second.pl?name=CDC27">
.
.
.
</map>
```

На основу друге и четврте координате се може утврдити висина линка на слици, а на основу тога је могуће и одредити категорију и поткатегорију антигена. Постојала је грешка на страници јер линкови за 'ETV5', 'HSPA1A', 'SYND1', 'TRIM68', 'XBP1', 'HMHA1' нису били у одговарајућој категорији, а 'SUPT7L' је био у двије. Категоризација за те антигене је морала да се одради експлицитно.

Страница антигена садржи табелу у којој су наведене различите форме у којима се тај антиген јавља. Форме су описане или потпуном секвенцом аминокиселина или фрагментом. Преузети су само подаци који садрже потпуну секвенцу аминокиселин.

Странице појединачних форми садрже табелу са подацима као што су идентификатор протеина, датум уноса у базу, име антигена, идентификатори у UniProt и NCBI Gene бази, листе епитопа, HLA лиганада, изоформи, мутација и синонима. На основу ове странице се формирају CSV датотеке – main.csv датотека која садржи све атрибуте са атомичним вриједностима и по једна CSV датотека за сваки атрибут који садржи листу вриједности. Тако настају epitope.csv, isoform.csv, ligand.csv, mutation.csv и synonym.csv. Свака од ових CSV датотека се користи за учитавање података у одговарајуће табеле у бази података над којом ће се вршити обрада: Main, T_cell_epitope, Isoform, HLA_ligand, Mutation, Synonym. За детаљан опис ових табела погледати додатак.

Неке грешке су откривене и исправљене у табелама:

- Start и End атрибути нису одговарали дужини секвенце у табелама HLA_ligand:

AG_ACCESSION	E_SEQ	START	END	REFERENCE
Ag000168	LLMWITQCFLPVFLAQPPSGQRR	157	180	15661941
Ag000168	LLMWITQCFLPVFLAQPPSGQRR	157	180	15661941
Ag000168	LLMWITQCFLPVFLAQPPSGQRR	157	180	15661941
Ag000168	WITQCFLPVFLAQPPSGQRR	160	180	15534491
Ag000168	LLMWITQCFLPVFLAQPPSGQRR	157	180	15661941
Ag001733	GVLVGVALI	693	702	10940913

Ag001735	GVLVGVALI	411	420	10940913
Ag000009	GVLVGVALI	693	702	10940913
Ag001734	GVLVGVALI	693	702	10940913

T_cell_epitope:

AG_ACCESSION	E_SEQ	START	END	REFERENCE
Ag000457	FLAEDALNTV	903	913	12750359
Ag000458	FLAEDALNTV	866	876	12750359
Ag002089	FLAEDALNTV	866	876	12750359

Дужине су преправљене тако што је Start атрибут увећан за један

- У табели HLA_ligand је била нетачно уписана референца:

AG_ACCESSION	ALLELE	START	END	REFERENCE
Ag000337	DRB1*0401 or DRB1*03011	2053	2072	1626703

Референца је преправљена у 16267033.

7.2 Додатне табеле

Додате су табеле Allele, Aminoacids, Disorder и Disorder_numeric.

Табела Allele служи за сврставање алела наведених у табелама HLA_ligand i T_Cell_Epitope. Сврставање се врши у серотипове. Додатно, HLA алели I класе се сврставају у супертипове, а HLA алели II класе се сврставају у групе. Група алела је означена називом локуса алела који је звјездицом одвојен од прве двије цифре ознака за алеле који припадају групи.

Атрибути Sydney и Harjanto садрже супертипове за наведене алеле по истраживањима[11], односно[12] или null вриједности уколико наведени алел није обрађен у одговарајућем истраживању. Serotype садржи серотип за наведени алел према[5].

Атрибут Allele за поједине епитопе садржи серотип. Тада се на основу алела који припадају датом серотипу одређује супертип. У неким случајевима HLA алела II класе овај атрибут садржи групу. Тада се одређује серотип на основу алела који припадају овој групи.

Атрибут Allele_real садржи само ознаке алела из атрибута Allele. Уколико је у Allele дата група или серотип, Allele_real ће садржати null вриједност. Атрибут Supertype_S садржи сврставање HLA-A и HLA-B алела и серотипова према [11] и сврставање HLA алела II класе у групе. Атрибут Supertype_H садржи сврставање HLA-A и HLA-B алела и серотипова према [12] и сврставање HLA алела II класе у групе. Уколико сврставање није дефинисано у [12] користи се сврставање из [11].

Табела Aminoacids садржи својства и ознаке аминокиселина које учествују у изградњи протеина. Значајна дата особина је хидрофобност која је дата по Кајт-Дулитлу и по Хоп-Вудсу.

Табела Disorder садржи податке о почетку и крају уређених и неуређених региона у протеинима. Табела садржи податке за 10 предиктора.

Табела Disorder_numeric за сваки од 10 предиктора и сваку позицију у ланцу аминокиселина протеина антигена садржи нумеричку оцјену уређености. За детаље погледати додатак.

Табеле Disorder и Disorder_numeric не садрже податке за четири антигена дата у TANTIGEN-у: Ag000202, Ag000227, Ag004220 и Ag000538. За предиктор DISOPRED2 недостају још и Ag000163, Ag000453, Ag000308, Ag000310, Ag000459, Ag002115 и Ag000401.

7.3 Помоћне табеле

Помћне табеле су Protein_AA, Epitope, и Disorder_all.

Табела Protein_AA садржи ознаку аминокиселине за сваку позицију у сваком датом протеину из TANTIGEN-а. Ова табела се користи за валидацију података у другим табелама које садрже ознаке аминокиселина, као што је Disorder_numeric или Disorder_all

Табела Epitope представља унију табела HLA_ligand i T_cell_epitope. При том је додат атрибут Type који одређује из које од ове двије табеле је узет епитоп.

Табела Disorder_all је табела која је изграђена на основу табеле Disorder_numeric. Она садржи по двије колоне за сваки предиктор из те табеле. Једна колона садржи ознаку уређености, док друга садржи нумеричку оцјену. Оцјене и ознаке су дате за сваку аминокиселину свих антигена.

7.4 Исправке података у припреми за истраживање података

Наредне измјене нису примјењене на табеле из TANTIGENa, већ само на нове и помоћне.

При изради табеле Allele која служи за категоризацију алела, откривено је да се ознака за алел Cw*0601 не користи, зато што је њом означен алел за који се грешком сматрало да је нови. Ова ознака за алел је измјењена, умјесто ње се користи се ознака Cw*06020101[13].

Неки епитопи имају двоструке ознаке за алел:

- 'DRB1*0401 or DRB1*03011'
- 'DRB1*0401 or DRB1*0301'
- 'DRB1*0401 or DRB1*1301'

Епитопи са оваквим ознакама су уклоњени, а за сваки наведени алел додати су епитопи који су по осталим атрибутима исти као и уклоњени.

У табели Epitope откривени су епитопи који једино имају различит Type атрибут, по свему осталом су исти. Пошто се подразумјева да је сваки Т-ћелијски епитоп истовремено и HLA лиганд, међу овим епитопима уклоњени су сви они који су HLA лиганди.

7.5 Припрема табела за рачунање правила придруживања

Табела Consensus, за сваки протеин антигена и сваку позицију у његовом ланцу аминокиселина рачуна консензусе предиктора уређености. Атрибут Cons1 односно Cons2 садрже ознаку припадности аминокиселине уређеном или неуређеном простору амиинокиселина уколико постоји поклапање на 7 односно 9 од 9 предиктора уређености у ознаци. У супротном, аминокиселина добија ознаку да не постоји поклапање довољног броја предиктора око њене уређености. ANCHOR је предиктор који не учествује рачунању консензуса, зато што он предвиђа неуређене регионе склоне везивању за други протеин-партнер.

Табела Epitope_AA садржи податке о појединачним амиинокиселинама које се налазе у епитопима, ти подаци су: редни број аминокиселине у самом епитопу, нумеричка ознака и оцјена хидрофобности и по Кајт-Дулитлу, и по Хоп-Вудсу, оцјене уређености из табеле Consensus, релативни положај у односу на епитоп.

Табела Epitope_Data садржи податке о епитопу као цјелини: његову уређеност према консензусима – да ли комплетно припада уређеном, неуређеном, региону без консензуса или се налази у више региону; његову хидрофобност – да ли има више хидрофобних или хидрофилних аминокиселина, просјек нумеричких оцјена хидрофобности и по Кајт-Дулитлу и по Хоп-Вудсу. Ту су још укључени подаци о томе да ли се епитоп налази у првих или задњих 30% ланца, или је у средњих 40%, ту су категорије његовог HLA-алела – класа, серотип, супертип, и ту су имена и категорије антигена којима припада.

Детаљи о свим табелама се налазе у додатку.

8. Резултати

8.1 Приступ

При задавању параметара алгоритма за израчунавање правила придруживања и филтрирању правила води се рачуна о сљедећем:

- минимална подршка треба да буде довољно малена да би правилима биле обухваћене законитости међу ријетко заступљеним ставкама. На примјер, епитопи који су у потпуности у неуређеним регионима према консензусу 7 од 9 предиктора чине око 2% укупног броја епитопа.
- минимална поузданост - занимају нас правила високе поузданости, али је потребно да минимална поузданост буде нешто нижа да бисмо могли да отпратимо како се мијењају подршка и поузданост када се дода или одузме ставка из тијела правила.
- лифт треба да буде већи од 1 да би смо уклонили везе међу независним или негативно корелираним ставкама.
- максималан број ставки у тијелу правила треба да буде довољан да обухвати све атрибуте који се користе у израчунавању.

При избору атрибута који ће учествовати у обради води се рачуна о томе да се не узима више од једног атрибута из скупа у ком су они функционално зависни. Функционална зависност може да постоји зато што један атрибут представља категорију другог, као на примјер серотип или супертип за алел. Такође, узима се само један атрибут из скупа атрибута који описују исту величину. Такав је, на примјер, скуп атрибута који одређује уређеност региона ком епитоп припада, или скуп атрибута који одређују хидрофобност епитопа.

Атрибути табеле Epitope_data се дијеле на сљедеће скупове из којих се узима највише један атрибут:

{Type}, {Epitope_d1, Epitope_d2}, {Epitope_KD_maj, Epitope_HW_maj, Epitope_KD_avg, Epitope_HW_avg}, {Pos_in_protein}, {Allele_real, Supertype_S, Supertype_H, Serotype, MHC_class}, {Ag_accession, Antigen_name, Sub_category, Category}

Атрибути Start i End се не користе, они су заступљени кроз атрибут Pos_in_protein. Атрибут Allele се не користи зато што су у њему помијешани алели и њихове категорије – серотипови. Умјесто овог атрибута се користе Allele_real и Serotype.

Атрибути табеле Epitope_AA се дијеле на сљедеће скупове:

{Ag_accession}, {Type}, {Pos_in_epitope}, {Epitope_len}, {Rel_pos_in_epitope}, {AA}, {Cons1, Cons2}, {KD_l, KD_v, HW_l, HW_v}

Атрибути Start, End, Allele и Position се не користе.

Идеја анализе правила је да правила буду груписана према глави правила, а потом према броју елемената који чине тијело, а након тога да буду сортирана према поузданости. На овај начин можемо да пратимо мијењање вриједности поузданости и подршке при додавању или одузимању ставке у тијело правила. Као један од параметара за филтрирање може да се искористи производ подршке и поузданости – није за очекивати да се додавањем ставке у тијело правила ниске подршке и ниске поузданости добије правило високе поузданости са довољно високом подршком. Такође, мали скуп правила је корисно сортирати према нискама карактера које чине тијело правила. Овакво сортирање нам омогућава да лакше отпратимо

присуство или одсуство ставки у тијелу правила.

8.2 Анализа правила

Атрибути табеле Epitope_data укључени у израчунавање правила:

- Type
- Epitope_d1
- Epitope_KD_maj
- Pos_in_protein
- MHC_class
- Sub_category

Опције израчунавања правила:

- подршка ≥ 0,2%
- поузданост ≥ 75%

Филтрирање:

- лифт ≥ 1,1

Правила којима главу чини Epitope_d1 = 'O' – епитоп припада уређеном региону протеина према консензусу 7 од 9 предиктора:

Филтрирање:

- подршка ≥ 0,5%
- поузданост ≥ 95%
- производ поузданости и подршке ≥ 0,005%

▼ Body	Head	Support	Confidence	Lift	Sup...	Abs...
[SUB_CATEGORY=shared_tumor_specific]+[POS_IN_PROTEIN=E]+[EPITOPE_KD_MAJ=HF]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	0,9863%	95,0549%	1,36	0,01	346
[SUB_CATEGORY=shared_tumor_specific]+[EPITOPE_KD_MAJ=HF]+[MHC_CLASS=1]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	7,0296%	95,1389%	1,36	0,07	2.466
[SUB_CATEGORY=differentiation]+[POS_IN_PROTEIN=E]+[EPITOPE_KD_MAJ=HF]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	1,4424%	96,3810%	1,38	0,01	506
[SUB_CATEGORY=differentiation]+[POS_IN_PROTEIN=E]+[EPITOPE_KD_MAJ=HF]+[MHC_CLASS=1]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	1,4424%	96,3810%	1,38	0,01	506
[SUB_CATEGORY=differentiation]+[POS_IN_PROTEIN=E]+[EPITOPE_KD_MAJ=HF]+[MHC_CLASS=1]	[EPITOPE_D1=O]	1,4994%	96,5138%	1,38	0,01	526
[SUB_CATEGORY=differentiation]+[POS_IN_PROTEIN=E]+[EPITOPE_KD_MAJ=HF]	[EPITOPE_D1=O]	1,4994%	96,5138%	1,38	0,01	526
[POS_IN_PROTEIN=M]+[SUB_CATEGORY=unclassified]+[TYPE=HLA ligand]	[EPITOPE_D1=O]	0,6072%	95,5157%	1,37	0,01	213
[POS_IN_PROTEIN=M]+[SUB_CATEGORY=shared_tumor_specific]+[EPITOPE_KD_MAJ=HF]+[MHC_CLASS=1]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	6,0462%	97,3829%	1,39	0,06	2.121
[POS_IN_PROTEIN=M]+[SUB_CATEGORY=shared_tumor_specific]+[EPITOPE_KD_MAJ=HF]+[MHC_CLASS=1]	[EPITOPE_D1=O]	6,0462%	97,3829%	1,39	0,06	2.121
[POS_IN_PROTEIN=M]+[SUB_CATEGORY=differentiation]+[EPITOPE_KD_MAJ=HF]	[EPITOPE_D1=O]	1,0747%	95,2020%	1,36	0,01	377
[MHC_CLASS=2]+[TYPE=HLA ligand]	[EPITOPE_D1=O]	0,8979%	100,0000%	1,43	0,01	315
[MHC_CLASS=2]+[EPITOPE_KD_MAJ=HL]+[TYPE=HLA ligand]	[EPITOPE_D1=O]	0,7839%	100,0000%	1,43	0,01	275
[MHC_CLASS=2]+[EPITOPE_KD_MAJ=0]+[SUB_CATEGORY=shared_tumor_specific]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	0,6043%	100,0000%	1,43	0,01	212
[MHC_CLASS=2]+[EPITOPE_KD_MAJ=0]+[SUB_CATEGORY=shared_tumor_specific]	[EPITOPE_D1=O]	0,6043%	100,0000%	1,43	0,01	212
[MHC_CLASS=2]+[EPITOPE_KD_MAJ=0]+[POS_IN_PROTEIN=M]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	0,6956%	100,0000%	1,43	0,01	244
[MHC_CLASS=2]+[EPITOPE_KD_MAJ=0]+[POS_IN_PROTEIN=M]+[SUB_CATEGORY=shared_tumor_specific]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	0,5131%	100,0000%	1,43	0,01	180
[MHC_CLASS=2]+[EPITOPE_KD_MAJ=0]+[POS_IN_PROTEIN=M]+[SUB_CATEGORY=shared_tumor_specific]	[EPITOPE_D1=O]	0,5131%	100,0000%	1,43	0,01	180
[MHC_CLASS=2]+[EPITOPE_KD_MAJ=0]+[POS_IN_PROTEIN=M]	[EPITOPE_D1=O]	0,8096%	100,0000%	1,43	0,01	284
[MHC_CLASS=2]+[EPITOPE_KD_MAJ=0]	[EPITOPE_D1=O]	1,1288%	95,1923%	1,36	0,01	396
[EPITOPE_KD_MAJ=0]+[SUB_CATEGORY=shared_tumor_specific]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	0,6842%	96,0000%	1,38	0,01	240
[EPITOPE_KD_MAJ=0]+[SUB_CATEGORY=shared_tumor_specific]	[EPITOPE_D1=O]	0,6842%	96,0000%	1,38	0,01	240
[EPITOPE_KD_MAJ=0]+[POS_IN_PROTEIN=M]+[SUB_CATEGORY=shared_tumor_specific]+[TYPE=T-cell epitope]	[EPITOPE_D1=O]	0,5644%	95,1923%	1,36	0,01	198
[EPITOPE_KD_MAJ=0]+[POS_IN_PROTEIN=M]+[SUB_CATEGORY=shared_tumor_specific]	[EPITOPE_D1=O]	0,5644%	95,1923%	1,36	0,01	198

Слика 8.1 Филтрирана правила са Epitope_d1='O' у глави

Слика 8.2 Значајна правила са Epitope_d1='O' у глави

Значајна правила:

- уколико је епитоп означен као HLA лиганд и везује се за MHC молекул II класе, са 100% поузданости и подршком од 0,89% (315 случајева) се налази у уређеном региону.
- хидрофобни епитопи који везују се за MHC молекуле II класе и припадају групи заједничких антигена специфичних за тумор, са поузданошћу од 95,14% и подршком од 7,03% (2.466 случајева) се налазе у уређеном региону. Уколико се налазе у средњих 40% ланца протеина, са поузданошћу од 97,38% и подршком од 6,05% (2.121 случај) припадају уређеном региону. Сви наведени епитопи су и Т-ћелијски епитопи.
- епитопи који се везују за MHC молекуле II класе и имају једнак број хидрофобних и хидрофилних аминокиселина, са поузданошћу од 95,19% и подршком 1,13% (396 случајева) се може рећи да припадају уређеном региону. Уколико се налазе у средњих 40% ланца протеина, са поузданошћу од 100% и подршком 0,81% (284 случаја) припадају уређеном региону

Правила којима главу чини Epi_KD_Maj = 'HL' – већина аминокиселина епитопа је према Кајт-Дулитлу хидрофилна.

Филтрирање:

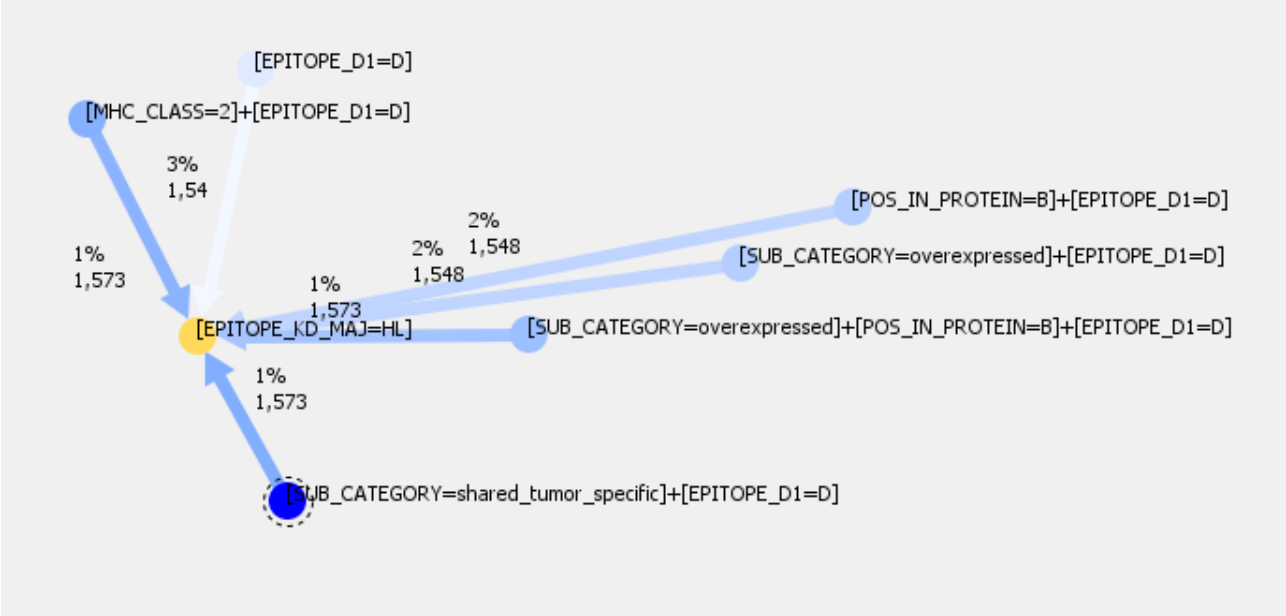
- подршка ≥ 0,5%
- поузданост ≥ 90%
- производ поузданости и подршке ≥ 0,005%
- тијело садржи дефинисан Epitope_d1 атрибут

Body	Head	Support	Confidence	Lift	Su...	A...
[EPITOPE_D1=D]	[EPITOPE_KD_MAJ=HL]	2,6739%	97,9123%	1,54	0,03	938
[EPITOPE_D1=D] +[MHC_CLASS=1]	[EPITOPE_KD_MAJ=HL]	1,9441%	97,1510%	1,53	0,02	682
[EPITOPE_D1=D] +[MHC_CLASS=1] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	1,5165%	96,3768%	1,52	0,01	532
[EPITOPE_D1=D] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	2,2463%	97,5248%	1,53	0,02	788
[MHC_CLASS=2] +[EPITOPE_D1=D]	[EPITOPE_KD_MAJ=HL]	0,7298%	100,0000%	1,57	0,01	256
[MHC_CLASS=2] +[EPITOPE_D1=D] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,7298%	100,0000%	1,57	0,01	256
[MHC_CLASS=2] +[POS_IN_PROTEIN=B] +[EPITOPE_D1=D]	[EPITOPE_KD_MAJ=HL]	0,6727%	100,0000%	1,57	0,01	236
[MHC_CLASS=2] +[POS_IN_PROTEIN=B] +[EPITOPE_D1=D] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,6727%	100,0000%	1,57	0,01	236
[MHC_CLASS=2] +[SUB_CATEGORY=overexpressed] +[EPITOPE_D1=D]	[EPITOPE_KD_MAJ=HL]	0,5131%	100,0000%	1,57	0,01	180
[MHC_CLASS=2] +[SUB_CATEGORY=overexpressed] +[EPITOPE_D1=D] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,5131%	100,0000%	1,57	0,01	180
[MHC_CLASS=2] +[SUB_CATEGORY=overexpressed] +[POS_IN_PROTEIN=B] +[EPITOPE_D1=D]	[EPITOPE_KD_MAJ=HL]	0,5131%	100,0000%	1,57	0,01	180
[MHC_CLASS=2] +[SUB_CATEGORY=overexpressed] +[POS_IN_PROTEIN=B] +[EPITOPE_D1=D] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,5131%	100,0000%	1,57	0,01	180
[POS_IN_PROTEIN=B] +[EPITOPE_D1=D]	[EPITOPE_KD_MAJ=HL]	1,7446%	98,3923%	1,55	0,02	612
[POS_IN_PROTEIN=B] +[EPITOPE_D1=D] +[MHC_CLASS=1]	[EPITOPE_KD_MAJ=HL]	1,0718%	97,4093%	1,53	0,01	376
[POS_IN_PROTEIN=B] +[EPITOPE_D1=D] +[MHC_CLASS=1] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,8552%	96,7742%	1,52	0,01	300
[POS_IN_PROTEIN=B] +[EPITOPE_D1=D] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	1,5279%	98,1685%	1,54	0,01	536
[SUB_CATEGORY=overexpressed] +[EPITOPE_D1=D]	[EPITOPE_KD_MAJ=HL]	1,7389%	98,3871%	1,55	0,02	610
[SUB_CATEGORY=overexpressed] +[EPITOPE_D1=D] +[MHC_CLASS=1]	[EPITOPE_KD_MAJ=HL]	1,2258%	97,7273%	1,54	0,01	430
[SUB_CATEGORY=overexpressed] +[EPITOPE_D1=D] +[MHC_CLASS=1] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,8267%	96,6667%	1,52	0,01	290
[SUB_CATEGORY=overexpressed] +[EPITOPE_D1=D] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	1,3398%	97,9167%	1,54	0,01	470
[SUB_CATEGORY=overexpressed] +[POS_IN_PROTEIN=B] +[EPITOPE_D1=D]	[EPITOPE_KD_MAJ=HL]	1,2172%	100,0000%	1,57	0,01	427
[SUB_CATEGORY=overexpressed] +[POS_IN_PROTEIN=B] +[EPITOPE_D1=D] +[MHC_CLASS=1]	[EPITOPE_KD_MAJ=HL]	0,7041%	100,0000%	1,57	0,01	247
[SUB_CATEGORY=overexpressed] +[POS_IN_PROTEIN=B] +[EPITOPE_D1=D] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	1,0006%	100,0000%	1,57	0,01	351
[SUB_CATEGORY=shared_tumor_specific] +[EPITOPE_D1=D]	[EPITOPE_KD_MAJ=HL]	0,6043%	100,0000%	1,57	0,01	212
[SUB_CATEGORY=shared_tumor_specific] +[EPITOPE_D1=D] +[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,6043%	100,0000%	1,57	0,01	212

Слика 8.3 Филтрирана правила са Epitope_KD_maj='HL' у глави и дефинисаним Epitope_d1='D' у тијелу

Значајна правила за **епитопе у неуређеном региону**:

- Са поузданошћу од 97,91% и подршком од 2,67% (938 случајева) су хидрофилни
- Уколико се везују за MHC молекуле II класе, тада су хидрофилни са поузданошћу 100% и подршком 0,73% (256 случајева). Сви ови епитопи су и Т-ћелијски епитопи.
- Уколико припадају категорији прекомјерно испољених антигена, тада су са поузданошћу од 98,39% и подршком од 1,74% (610 случајева) хидрофилни. Уколико се налазе у првих 30% ланца протеина, тада су хидрофилни са поузданошћу од 98,39% и подршком од 1,74 (612 случајева). Епитопи су хидрофилни са поузданошћу 100% и подршком 1,21% (427 случајева) уколико за њих важе оба претходно наведена својства.
- Уколико припадају категорији заједничких антигена специфичних за тумор, тада су хидрофилни са поузданошћу 100% и подршком 0,6% (212 случајева). Сви ови епитопи су и Т-ћелијски епитопи.



Слика 8.4 Значајна правила са Epitope_KD_maj='HL' у глави и дефинисаним Epitope_d1='D' у тијелу

▲ Body	Head	Support	Confidence	Lift	Su...	Ab...
[EPITOPE_D1=N]+[SUB_CATEGORY=overexpressed]+[TYPE=HLA ligand]+[POS_IN_PROTEIN=B]	[EPITOPE_KD_MAJ=HL]	3,1243%	93,1181%	1,47	0,03	1.096
[EPITOPE_D1=N]+[SUB_CATEGORY=overexpressed]+[TYPE=HLA ligand]+[POS_IN_PROTEIN=B]+[MHC_CLASS=1]	[EPITOPE_KD_MAJ=HL]	3,1243%	93,1181%	1,47	0,03	1.096
[EPITOPE_D1=N]+[SUB_CATEGORY=shared_tumor_specific]+[POS_IN_PROTEIN=E]	[EPITOPE_KD_MAJ=HL]	1,1003%	93,4625%	1,47	0,01	386
[EPITOPE_D1=N]+[SUB_CATEGORY=shared_tumor_specific]+[POS_IN_PROTEIN=E]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	1,1003%	95,5446%	1,50	0,01	386
[EPITOPE_D1=N]+[TYPE=HLA ligand]+[POS_IN_PROTEIN=B]	[EPITOPE_KD_MAJ=HL]	3,2526%	92,0904%	1,45	0,03	1.141
[EPITOPE_D1=N]+[TYPE=HLA ligand]+[POS_IN_PROTEIN=B]+[MHC_CLASS=1]	[EPITOPE_KD_MAJ=HL]	3,2526%	92,0904%	1,45	0,03	1.141
[MHC_CLASS=2]+[EPITOPE_D1=N]+[POS_IN_PROTEIN=B]	[EPITOPE_KD_MAJ=HL]	0,7070%	96,1240%	1,51	0,01	248
[MHC_CLASS=2]+[EPITOPE_D1=N]+[POS_IN_PROTEIN=B]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,7070%	96,1240%	1,51	0,01	248
[MHC_CLASS=2]+[EPITOPE_D1=N]+[POS_IN_PROTEIN=E]	[EPITOPE_KD_MAJ=HL]	3,4635%	100,0000%	1,57	0,03	1.215
[MHC_CLASS=2]+[EPITOPE_D1=N]+[POS_IN_PROTEIN=E]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	3,4635%	100,0000%	1,57	0,03	1.215
[MHC_CLASS=2]+[EPITOPE_D1=N]+[SUB_CATEGORY=overexpressed]	[EPITOPE_KD_MAJ=HL]	3,2497%	99,1304%	1,56	0,03	1.140
[MHC_CLASS=2]+[EPITOPE_D1=N]+[SUB_CATEGORY=overexpressed]+[POS_IN_PROTEIN=E]	[EPITOPE_KD_MAJ=HL]	1,9812%	100,0000%	1,57	0,02	695
[MHC_CLASS=2]+[EPITOPE_D1=N]+[SUB_CATEGORY=overexpressed]+[POS_IN_PROTEIN=E]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	1,9812%	100,0000%	1,57	0,02	695
[MHC_CLASS=2]+[EPITOPE_D1=N]+[SUB_CATEGORY=overexpressed]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	3,2497%	99,1304%	1,56	0,03	1.140
[MHC_CLASS=2]+[EPITOPE_D1=N]+[SUB_CATEGORY=shared_tumor_specific]+[POS_IN_PROTEIN=E]	[EPITOPE_KD_MAJ=HL]	0,9407%	100,0000%	1,57	0,01	330
[MHC_CLASS=2]+[EPITOPE_D1=N]+[SUB_CATEGORY=shared_tumor_specific]+[POS_IN_PROTEIN=E]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,9407%	100,0000%	1,57	0,01	330
[MHC_CLASS=2]+[POS_IN_PROTEIN=M]+[EPITOPE_D1=N]+[SUB_CATEGORY=overexpressed]	[EPITOPE_KD_MAJ=HL]	0,8808%	100,0000%	1,57	0,01	309
[MHC_CLASS=2]+[POS_IN_PROTEIN=M]+[EPITOPE_D1=N]+[SUB_CATEGORY=overexpressed]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,8808%	100,0000%	1,57	0,01	309
[MHC_CLASS=2]+[POS_IN_PROTEIN=M]+[SUB_CATEGORY=differentiation]+[EPITOPE_D1=N]	[EPITOPE_KD_MAJ=HL]	0,5872%	100,0000%	1,57	0,01	206
[MHC_CLASS=2]+[POS_IN_PROTEIN=M]+[SUB_CATEGORY=differentiation]+[EPITOPE_D1=N]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,5872%	100,0000%	1,57	0,01	206
[MHC_CLASS=2]+[SUB_CATEGORY=differentiation]+[EPITOPE_D1=N]	[EPITOPE_KD_MAJ=HL]	1,1916%	100,0000%	1,57	0,01	418
[MHC_CLASS=2]+[SUB_CATEGORY=differentiation]+[EPITOPE_D1=N]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	1,1916%	100,0000%	1,57	0,01	418
[POS_IN_PROTEIN=M]+[EPITOPE_D1=N]+[SUB_CATEGORY=unclassified]	[EPITOPE_KD_MAJ=HL]	0,7355%	90,5263%	1,42	0,01	258
[POS_IN_PROTEIN=M]+[EPITOPE_D1=N]+[SUB_CATEGORY=unclassified]+[MHC_CLASS=1]	[EPITOPE_KD_MAJ=HL]	0,6927%	90,0000%	1,42	0,01	243
[POS_IN_PROTEIN=M]+[EPITOPE_D1=N]+[SUB_CATEGORY=unclassified]+[MHC_CLASS=1]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,6927%	90,0000%	1,42	0,01	243
[POS_IN_PROTEIN=M]+[EPITOPE_D1=N]+[SUB_CATEGORY=unclassified]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,7355%	90,5263%	1,42	0,01	258
[POS_IN_PROTEIN=M]+[SUB_CATEGORY=differentiation]+[EPITOPE_D1=N]	[EPITOPE_KD_MAJ=HL]	0,7526%	90,1024%	1,42	0,01	264

Слика 8.5 Филтрирана правила са Epitope_KD_maj='HL' у глави и дефинисаним Epitope_d1='N' у тијелу

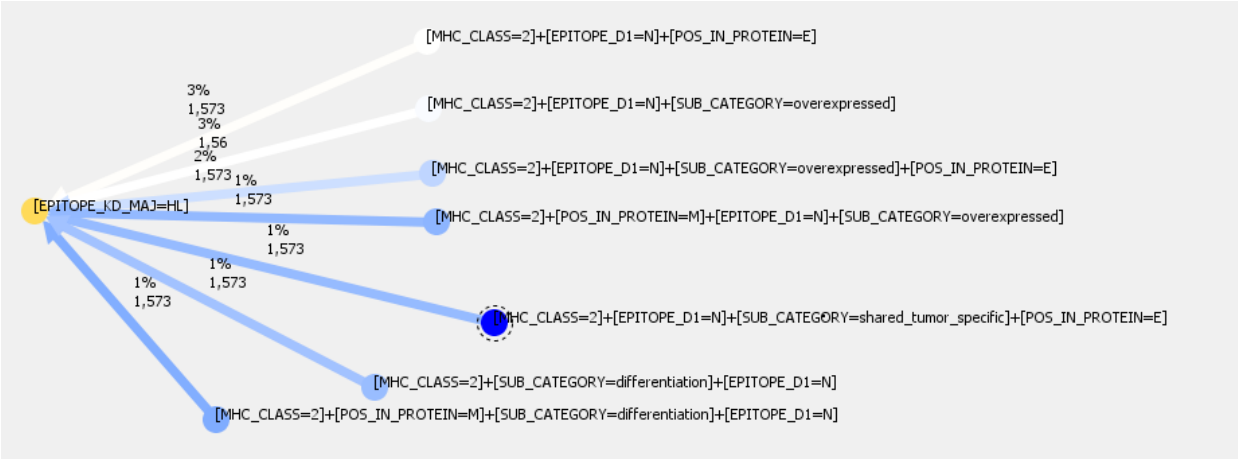
Значајна правила за **епитопе који се налазе на граници уређеног, неуређеног или мјешовитог региона**:

- Хидрофилни су са поузданошћу од 100% са подршком 3,25% (1.141 случај) уколико се везују за MHC молекуле II класе и уколико се налазе у задњих 30% ланца протеина. Сви су Т-ћелијски епитопи.
- Уколико припадају прекомјерно испољеним антигенима и везују се за MHC молекуле II класе, хидрофилни су са поузданошћу 99,13% и подршком 3,25% (1.140 случајева). Ако су при том и у задњих 30% ланца протеина, тада су хидрофилни са поузданошћу 100% и подршком 1,98% (625 случајева), а ако се налазе у средњих 40% ланца,

хидрофилни су са поузданошћу 100% и подршком 0,88% (309 случајева). Сви су Т-ћелијски епитопи.

– Уколико припадају развојним антигенима и везују се за МНС молекуле II класе, хидрофилни су са поузданошћу 100% и подршком 1,19% (418 случајева). Сви су Т-ћелијски епитопи.

– Уколико припадају заједничким антигенима специфичним за тумор, везују се за МНС молекуле II класе и налазе се у задњих 30% ланца протеина, хиддрофилни су са 100% поузданости и подршком од 0,94% (330 случајева). Сви су Т-ћелијски епитопи.



Слика 8.6 Значајна правила са Epitope_KD_maj='HL' у глави и дефинисаним Epitope_d1='N' у тијелу

▲ Body	Head	Support	Confidence	Lift	Su...	Ab...
[EPITOPE_D1=O]+[SUB_CATEGORY=substitution_mutation]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,6129%	95,9821%	1,51	0,01	215
[MHC_CLASS=2]+[EPITOPE_D1=O]+[POS_IN_PROTEIN=B]	[EPITOPE_KD_MAJ=HL]	1,9755%	100,0000%	1,57	0,02	693
[MHC_CLASS=2]+[EPITOPE_D1=O]+[POS_IN_PROTEIN=B]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	1,7930%	100,0000%	1,57	0,02	629
[MHC_CLASS=2]+[EPITOPE_D1=O]+[POS_IN_PROTEIN=E]	[EPITOPE_KD_MAJ=HL]	3,8113%	90,5213%	1,42	0,03	1.337
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=overexpressed]	[EPITOPE_KD_MAJ=HL]	4,2531%	91,1423%	1,43	0,04	1.492
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=overexpressed]+[POS_IN_PROTEIN=B]	[EPITOPE_KD_MAJ=HL]	1,0291%	100,0000%	1,57	0,01	361
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=overexpressed]+[POS_IN_PROTEIN=B]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	1,0291%	100,0000%	1,57	0,01	361
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=overexpressed]+[POS_IN_PROTEIN=E]	[EPITOPE_KD_MAJ=HL]	0,9151%	95,2522%	1,50	0,01	321
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=overexpressed]+[POS_IN_PROTEIN=E]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,7098%	93,9623%	1,48	0,01	249
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=overexpressed]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	4,0479%	93,1148%	1,46	0,04	1.420
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=shared_tumor_specific]+[POS_IN_PROTEIN=E]	[EPITOPE_KD_MAJ=HL]	2,0182%	92,1875%	1,45	0,02	708
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=shared_tumor_specific]+[POS_IN_PROTEIN=E]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	2,0182%	92,1875%	1,45	0,02	708
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=substitution_mutation]	[EPITOPE_KD_MAJ=HL]	0,5217%	100,0000%	1,57	0,01	183
[MHC_CLASS=2]+[EPITOPE_D1=O]+[SUB_CATEGORY=substitution_mutation]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	0,5217%	100,0000%	1,57	0,01	183
[MHC_CLASS=2]+[POS_IN_PROTEIN=M]+[EPITOPE_D1=O]+[SUB_CATEGORY=overexpressed]+[TYPE=T-cell epitope]	[EPITOPE_KD_MAJ=HL]	2,3090%	90,1001%	1,42	0,02	810

Слика 8.7 Филтрирана правила са Epitope_KD_maj='HL' у глави и дефинисаним Epitope_d1='O' у тијелу

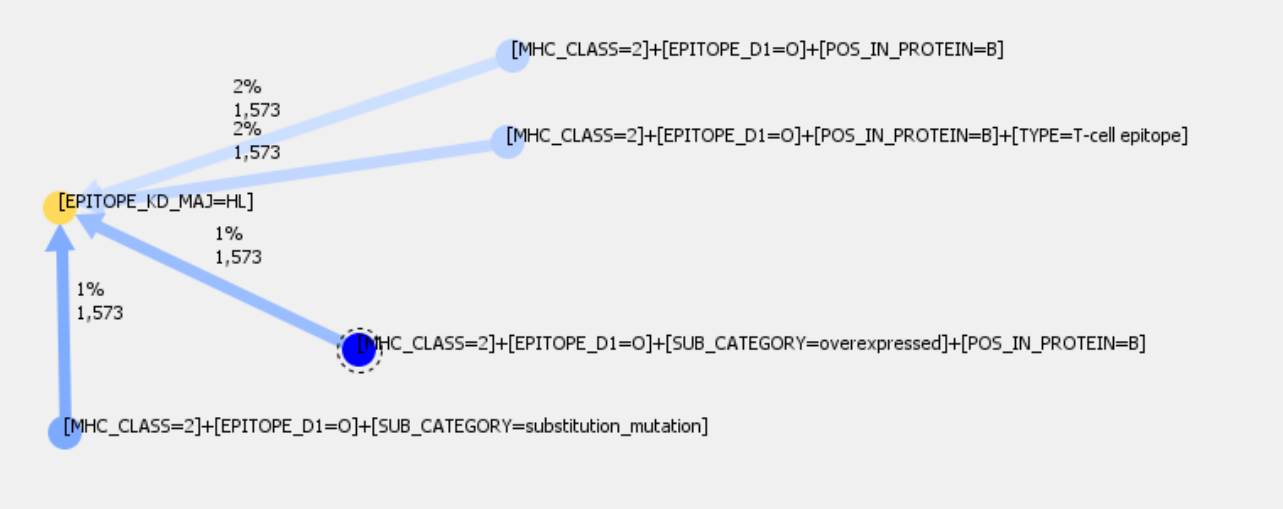
Значајна правила за **епитопе који се налазе у уређеним регионима протеина**:

– Уколико се везују за МНС молекуле II класе и налазе се у посљедњих 30% ланца протеина, може се рећи да су хидрофилни са поузданошћу од 100% и подршком од 1,98% (693 случаја).

– Уколико се везују за МНС молекуле II класе, налазе се у посљедњих 30% ланца протеина и припадају категорији прекомјерно испољених антигена, хидрофилни су са поузданошћу 100% и подршком 1,03% (361 случај).

– Уколико се везују за МНС молекуле II класе и припадају категорији јединствених

антигена насталих мутацијом (супституцијом), хидрофилни су са поузданошћу 100% и подршком 0,52% (183 случаја).



Слика 8.8 Значајна правила са Epitope_KD_maj='HL' у глави и дефинисаним Epitope_d1='O' у тијелу

9. Закључак

Анализом правила придуживања на датом скупу атрибута епитопа показало се да се у правилима која указују на хидрофилност са 100% поузданости, довољно да се појави својство везивања за MHC молекуле II класе и припадност крају протеина. За епитопе који припадају консензусу неуређених региона је чак довољно својство везивања за MHC молекуле класе II да укаже на хидрофилност са 100% поузданости. Правила која имају хидрофилност у глави и садрже својство епитопа да припада региону за којег нема консензуса нису наведена у раду у недостатку „занимљивих“, али би било занимљиво видјети какав је њихов однос са својством везивања са MHC молекулима класе II и положајем у протеину према хидрофилности. Када би постојала таква снажна веза, онда би постојала и снажна веза везивања за MHC молекуле II класе и положаја на крају ланца протеина са хидрофилношћу у којој уређеност региона ком епитоп припада уопште не би фигурисала.

Занимљиво је и својство да епитопи, који су само HLA лиганди заједно са својством везивања за MHC молекуле II класе указују на уређеност региона у ком се налазе. То би значило да су HLA лиганди II класе, уколико се налазе у неуређеном, неодређеном или граничном региону, аутоматски и Т-ћелијски епитопи.

Свако од наведених, „занимљивих“, правила поставља додатна питања, за чије се одговоре треба вратити истраживању података.

Литература

[1] Sette A., Rappuoli R., *Reverse Vaccinology: Developing Vaccines in the Era of Genomics*, Immunity, 2010, doi: 10.1016/j.immuni.2010.09.017.

[2] Landry, S.J., *Helper T-cell epitope immunodominance associated with structurally stable segments of hen egg lysozyme and HIV gp120*, J Theor Biol, 2000, doi:10.1006/jtbi.1999.1056

[3] Јеличић М., *Повезаност дужине епитопа и уређености делова протеина*, мастер рад, Математички факултет, Београд, 2012

[4] Patil A, Kinoshita K, Nakamura H. *Hub promiscuity in protein-protein interaction networks*, Int J Mol Sci. 2010, doi: 10.3390/ijms11041930

[5] *Genotype, Serotype and Supertype Classification,* http://www.soc-bdr.org/content/rds/authors/unit_tables_conversions_and_genetic_dictionaries/genotype_serotype_and_supertype_classification/ (14.12.2014)

[6] *TANTIGEN: Classification of tumor antigens,* http://cvc.dfci.harvard.edu/tadb/HTML/classification.php (01.12.2009)

[7] Vigneron N., *Human Tumor Antigens and Cancer Immunotherapy*, BioMed Research International Volume 2015, doi: 10.1155/2015/948501

[8] van der Bruggen P, Stroobant V, Vigneron N, Van den Eynde B., *Peptide database: T cell-defined tumor antigens*, Cancer Immun, 2013, http://www.cancerimmunity.org/peptide/ (13.10.2015)

[9] Tan P.N., Steinbach M., Kumar V., *Introduction to Data Mining*, Addison-Wesley Longman Publishing Co., 2005, ISBN: 0321321367

[10] Piatetsky-Shapiro G., *Discovery, analysis, and presentation of strong rules*, Knowledge Discovery in Databases, 1991

[11] Sidney J., Peters B., Frahm N., Brander C., Sette A, *HLA class I supertypes: a revised and updated classification*, BMC Immunology, 2008, doi: 10.1186/1471-2172-9-1

[12] Harjanto S, Ng L.F.P., Tong J.C., *Clustering HLA Class I Superfamilies Using Structural Interaction Patterns,* PLoS ONE, 2014, doi: 10.1371/journal.pone.0086655

[13] *IPD – IMGT/HLA* http://www.ebi.ac.uk/ipd/imgt/hla/

Додатак – опис табела

Main

Садржи податке за сваки протеин дат потпуним низом амиинокиселина у TANTIGEN бази.

- Ag_accession – идентификатор протеина антигена у TANTIGEN-у
- Entered – датум уноса у базу
- Updated – датум посљедње измјене
- Antigen_Name – име протеина засновано на HUGO номенклатури гена
- Category – категорија којој антиген припада. Може бити:
  - shared_antigen
  - unique_antigen
  - unclassified
- Sub_category – подкатегорија
  - За категорију shared_antigen може имати вриједности:
    - differentiation
    - overexpressed
    - shared_tumor_specific
  - За категорију unique_antigen може имати вриједности:
    - alternative_orf
    - internal_tandem_repeat
    - intron_encoding
    - substitution_mutation
  - За категорију unclassified може имати вриједности:
    - unclassified
- Common_Name – име антигена које се користи у имунологији тумора
- Full_Name – потпун назив протеина из GeneCard базе или литературе
- Isoform_Name – назив изоформе антигена додијељен у TANTIGEN-u
- Isoform_Synonym – синоними изоформе
- UniProt_ID – идентификатор у UniProt bazi
- NCBI_Gene_ID – идентификатор у NCBI Gene бази
- GeneCard_ID – идентификатор у GeneCard бази
- Gene_express_prf – Профил генетског израза (енг. *gene expression profile)* изведен анализом броја ознака изражених секвенци (енг. *expressed sequence tag*, *EST)* из UniGene базе.
- Comment – напомене о антигену
- Annotation – означава како је дат низ аминокиселина антигена. У бази над којом се врши обрада се налазе само протеини дати потпуним низом аминокиселина
- Sequence – низ аминокиселина антигена

Protein_AA
Помоћна табела, изведена из Main табеле. Садржи податке о свим аминокиселинама у антигену.

- Ag_accession – идентификатор антигена
- Position – позиција у низу аминокиселина
- AA – једнословна ознака аминокиселине

Synonym
Садржи друга имена протеина или гена антигена.

- Ag_accession – идентификатор антигена
- Synonym – синоним

Isoform
Садржи идентификаторе алтернативних облика антигена. Садржи и идентификаторе антигена дате фрагментом низа амино-киселина.

- Ag_accession – идентификатор антигена
- Isoform – идентификатор протеина изоформе TANTIGEN-у

Mutation
Садржи идентификаторе мутација антигена. Садржи и идентификаторе антигена дате фрагментом низа амино-киселина.

- Ag_accession – идентификатор антиген
- Mutation – идентификатор протеина мутације у TANTIGEN-у

T_cell_epitope
Садржи податке о Т-ћелијским епитопима.

- Ag_accession – идентификатор антигена ком епитоп припада
- Sequence – низ аминокиселина епитопа
- Start – редни број прве аминокиселине епитопа у низу аминокиселина антигена
- End - редни број посљедње аминокиселине епитопа у низу аминокиселина антигена
- Allele – Алел HLA молекула за који се епитоп везује. Може да садржи и серотип ком припада умјесто алела.
- Reference – PubMed идентификатор рада који садржи резултат валидације Т-ћелијског епитопа

HLA_ligand
Садржи податке HLA лигандима.

- Ag_accession – идентификатор антигена ком HLA лиганд припада
- Sequence – низ аминокиселина HLA лиганда
- Start – редни број прве аминокиселине HLA лиганда у низу аминокиселина антигена
- End - редни број посљедње аминокиселине HLA лиганда у низу аминокиселина антигена
- Allele – Алел HLA молекула за који се епитоп везује. Може да садржи и серотип ком

припада умјесто алела.
- Reference – PubMed идентификатор рада који садржи резултат откривања HLA лиганда

Epitope
Помоћна табела, изведена из табела T_cell_epitope и HLA_ligand. Садржи унију података о HLA лигандима и Т-ћелијским епитопима.
- Ag_accession – идентификатор антигена ком епитоп припада
- Sequence – низ аминокиселина епитопа
- Start – редни број прве аминокиселине епитопа у низу аминокиселина антигена
- End - редни број посљедње аминокиселине епитопа у низу аминокиселина антигена
- Allele – Алел HLA молекула за који се епитоп везује. Може да садржи и серотип ком припада умјесто алела.
- Reference – PubMed идентификатор рада који садржи резултат везивања HLA лиганда или резултат валидације Т-ћелијског епитопа
- Type – Врста епитопа. Може имати вриједности:
  - T-cell epitope – уколико је у питању Т-ћелијски епитоп
  - HLA ligand – уколико је у питању HLA лиганд

Disorder
Садржи податке о уређеним и неуређеним регионима структуре протеина антигена за разне предикторе.
- Ag_accession – идентификатор антигена
- Start - редни број прве аминокиселине региона
- End – редни број посљедње аминокиселине региона
- Order_level – Ознака уређености структуре протеина. Може имати вриједности:
  - O – означава уређени регион
  - D – означава неуређени регион
- Predictor – Предиктор уређености структуре протеина. Може имати вриједности:
  - ANCHOR
  - DISOPRED2
  - DisEMBL_Hot_loops
  - DisEMBL_Remark465
  - IUPred-L
  - IUPred-S
  - IsUnstruct
  - OnDCRF
  - RONN
  - VSL2b

Disorder_numeric
Садржи оцјену уређености за сваку аминокиселину у протеину антигена за сваки од предиктора.
- Ag_accession – идентификатор антигена
- Position – позиција у низу аминокиселина
- AA – једнословна ознака аминокиселине

- Value – нумеричка оцјена уређености за дату аминокиселину по датом предиктору
- Order_level – Ознака уређености амино-киселине. Може имати вриједности:
  - O – аминокиселина припада уређеном региону
  - D – аминокиселина припада неуређеном региону
- Predictor – Предиктор уређености структуре протеина.Може имати вриједности:
  - ANCHOR
  - DISOPRED2
  - DisEMBL_Hot_loops
  - DisEMBL_Remark465
  - IUPred-L
  - IUPred-S
  - IsUnstruct
  - OnDCRF
  - RONN
  - VSL2b

Disorder_all

Помоћна табела, изведена из Disorder_numeric табеле. За сваки предиктор је дата по колона за ознаку и за нумеричку оцјену уређености структуре протеина.

- Ag_accession – идентификатор антигена
- Position – позиција у низу аминокиселина
- AA – једнословна ознака аминокиселине
- ANCHOR_v – нумеричка оцјена по ANCHOR prediktoru
- ANCHOR_l - ознака по ANCHOR предиктору
- DISOPRED2_v - нумеричка оцјена по DISOPRED2 предиктору
- DISOPRED2_l - ознака по DISOPRED2 предиктору
- DISEMBLHL_v - нумеричка оцјена по DisEMBL_Hot_loops предиктору
- DISEMBLHL_l - ознака по DisEMBL_Hot_loops предиктору
- DISEMBLRM_v - нумеричка оцјена по DisEMBL_Remark465 предиктору
- DISEMBLRM_l - ознака по DisEMBL_Remark465 предиктору
- IUPREDL_v - нумеричка оцјена по IUPred-L предиктору
- IUPREDL_l - ознака по IUPred-L предиктору
- IUPREDS_v - нумеричка оцјена по IUPred-S предиктору
- IUPREDS_l - ознака по IUPred-S предиктору
- ISUNSTRUCT_v - нумеричка оцјена по IsUnstruct предиктору
- ISUNSTRUCT_l - ознака по IsUnstruct предиктору
- ONDCRF_v - нумеричка оцјена по OnDCRF предиктору
- ONDCRF_l - ознака по OnDCRF предиктору
- RONN_v - нумеричка оцјена по RONN предиктору
- RONN_l - ознака по RONN предиктору
- VSL2B_v - нумеричка оцјена по VSL2b предиктору
- VSL2B_l - ознака по VSL2b предиктору

Aminoacids

- Name – назив аминокиселин
- Code1 – једнословна ознака

- Code3 – трословна ознака (прво слово је велико, остала су мала)
- Codons – број различитих кодона који бивају преведени у аминокиселину
- Code3u – трословна ознака великим словима
- KDHidrotip – хидрофобност амино киселине према Кајт-Дулитлу. Може имати вриједности:
  - Hydrophobic – аминокиселина је хидрофобна
  - Hydrophilic – аминокиселина је хидрофилна
  - Unknown – непозната хидрофобност аминокиселине
- KDHidrofobnost – нумеричка оцјена хидрофобности аминокиселине према Кајт-Дулитлу
- KDHydrophobic - хидрофобност амино киселине према Кајт-Дулитлу. Може имати вриједности:
  - HF - аминокиселина је хидрофобна
  - HL - аминокиселина је хидрофилна
  - un - непозната хидрофобност аминокиселине
- HWHidrotip - хидрофобност амино киселине према Хоп-Вудсу. Може имати вриједности:
  - Hydrophobic – аминокиселина је хидрофобна
  - Hydrophilic – аминокиселина је хидрофилна
  - Unknown – непозната хидрофобност аминокиселине
- HWHidrofobnost – нумеричка оцјена хидрофобности аминокиселине према Хоп-Вудсу
- HWHydrophobic - хидрофобност амино киселине према Хоп-Вудсу. Може имати вриједности:
  - HF - аминокиселина је хидрофобна
  - HL - аминокиселина је хидрофилна
  - un - непозната хидрофобност аминокиселине

Allele

Садржи категоризацију HLA алела.

- Ag_accession – идентификатор антигена
- Position – позиција у низу аминокиселина
- Allele – ознака за алел у TANTIGEN-у
- Allele_real – ознака за алел. Уколико у TANTIGEN-у није дат алел, вриједност није дефинисана.
- Supertype_S – садржи ознаку супертипа за алеле HLA класе I и ознаке групе за HLA класе II. Ознаке супертипа за HLA-A и HLA-B алеле су узете по Сиднију
- Supertype_H – садржи ознаку супертипа за алеле HLA класе I и ознаке групе за HLA класе II. Ознаке супертипа за HLA-A и HLA-B алеле су узете по Харјантоу, а уколико су недефинисане, узете су по Сиднију
- Serotype – серотип
- Sydney – класификација HLA-A и HLA-B алела по Сиднију
- Harjanto – класификација HLA-A и HLA-B алела по Харјантоу

Consensus

- Ag_accession – идентификатор антигена
- Position – позиција у низу аминокиселина

- Cons1 - ознака консензуса 7 од 9 предиктора уређености (свих осим ANCHOR-a). Може имати вриједности:
  - O – постоји консензус да је аминокиселина у уређеном региону
  - D – постоји консензус да је аминокиселина у неуређеном региону
  - M – не постоји консензус
- Cons2 - ознака консензуса предиктора свих предиктора уређености осим ANCHOR-a. Може имати вриједности:
  - O – постоји консензус да је аминокиселина у уређеном региону
  - D – постоји консензус да је аминокиселина у неуређеном региону
  - M – не постоји консензус

Epitope_AA

- Ag_accession – идентификатор антигена
- Start – редни број прве аминокиселине епитопа у низу аминокиселина антигена
- End - редни број посљедње аминокиселине епитопа у низу аминокиселина антигена
- Allele – Алел HLA молекула за који се епитоп везује. Може да садржи и серотип ком припада умјесто алела.
- Type – Врста епитопа. Може имати вриједности:
  - T-cell epitope – уколико је у питању Т-ћелијски епитоп
  - HLA ligand – уколико је у питању HLA лиганд
- Position – позиција у низу аминокиселина
- Pos_in_epitope – редни број аминокиселине у епитопу
- Epitope_len – дужина епитопа
- Rel_pos_in_epitope – релативна позиција аминокиселине у епитопу. Узима вриједности из интервала [0, 1] – 0 уколико је на почетку епитопа, 1 уколико је на крају
- AA – једословна ознака аминокиселине
- Cons1 - ознака консензуса 7 од 9 предиктора уређености (свих осим ANCHOR-a). Може имати вриједности:
  - O – постоји консензус да је аминокиселина у уређеном региону
  - D – постоји консензус да је аминокиселина у неуређеном региону
  - M – не постоји консензус
- Cons2 - ознака консензуса предиктора свих предиктора уређености осим ANCHOR-a. Може имати вриједности:
  - O – постоји консензус да је аминокиселина у уређеном региону
  - D – постоји консензус да је аминокиселина у неуређеном региону
  - M – не постоји консензус
- KD_l - хидрофобност амино киселине према Кајт-Дулитлу. Може имати вриједности:
  - HF - аминокиселина је хидрофобна
  - HL - аминокиселина је хидрофилна
- KD_v - нумеричка оцјена хидрофобности аминокиселине према Кајт-Дулитлу
- HW_l- хидрофобност амино киселине према Хоп-Вудсу. Може имати вриједности:
  - HF - аминокиселина је хидрофобна
  - HL – аминокиселина је хидрофилна
- HW_v - нумеричка оцјена хидрофобности аминокиселине према Хоп-Вудсу

Epitope_data

- Ag_accession – идентификатор антигена
- Start – редни број прве аминокиселине епитопа у низу аминокиселина антигена
- End - редни број посљедње аминокиселине епитопа у низу аминокиселина антигена
- Allele – Алел HLA молекула за који се епитоп везује. Може да садржи и серотип ком припада умјесто алела.
- Type – Врста епитопа. Може имати вриједности:
  - T-cell epitope – уколико је у питању Т-ћелијски епитоп
  - HLA ligand – уколико је у питању HLA лиганд
- Epitope_d1 – ознака уређености цијелог епитопа према Cons1 из Epitope_AA табеле:
  - O – постоји консензус да се цијели епитоп налази у уређеном региону
  - D – постоји консензус да се цијели епитоп налази у неуређеном региону
  - M – не постоји консензус ни за једну амино-киселину епитопа
  - N – аминокиселине епитопа су различито означене
- Epitope_d2 – ознака уређености цијелог епитопа према Cons2 из Epitope_AA табеле:
  - O – постоји консензус да се цијели епитоп налази у уређеном региону
  - D – постоји консензус да се цијели епитоп налази у неуређеном региону
  - M – не постоји консензус ни за једну амино-киселину епитопа
  - N – аминокиселине епитопа су различито означене
- Epitope_KD_maj – ознака хидрофобности већине аминокиселина у епитопу по Кајт-Дулитлу. Може имати сљедеће ознаке:
  - HF – већина аминокиселина је хидрофобна
  - HL – већина аминокиселина је хидрофилна
  - 0 – једнак је број хидрофилних и хидрофобних аминокиселина у епитопу
- Epitope_HW_maj – ознака хидрофобности већине аминокиселина у епитопу по Хоп-Вудсу. Може имати сљедеће ознаке:
  - HF – већина аминокиселина је хидрофобна
  - HL – већина аминокиселина је хидрофилна
  - 0 – једнак је број хидрофилних и хидрофобних аминокиселина у епитопу
- Epitope_KD_avg – просјечна вриједност хидрофобности аминокиселина у епитопу по Кајт-Дулитлу
- Epitope_HW_avg – просјечна вриједност хидрофобности аминокиселина у епитопу по Хоп-Вудсу
- Pos_in_protein – положај епитопа у комплетном протеину. Може имати сљедеће ознаке:
  - B – епитоп се налази у првих 30% аминокиселина протеина
  - M – епитоп се налази у средњих 40% аминокиселина протеина
  - E – епитоп се налази у задњих 30% аминокиселина протеина
- Allele_real – ознака за алел. Уколико у TANTIGEN-у није дат алел, вриједност није дефинисана.
- Supertype_S – садржи ознаку супертипа за алеле HLA класе I и ознаке групе за HLA класе II. Ознаке супертипа за HLA-A и HLA-B алеле су узете по Сиднију
- Supertype_H – садржи ознаку супертипа за алеле HLA класе I и ознаке групе за HLA класе II. Ознаке супертипа за HLA-A и HLA-B алеле су узете по Харјантоу, а уколико су недефинисане, узете су по Сиднију
- Serotype – серотип
- MHC_class – класа MHC молекула за који се епитоп везује

- Antigen_name – име протеина засновано на HUGO номенклатури гена
- Category – категорија којој антиген припада. Може бити:
  - shared_antigen
  - unique_antigen
  - unclassified
- Sub_category – подкатегорија
  - За категорију shared_antigen може имати вриједности:
    - differentiation
    - overexpressed
    - shared_tumor_specific
  - За категорију unique_antigen може имати вриједности:
    - alternative_orf
    - internal_tandem_repeat
    - intron_encoding
    - substitution_mutation
  - За категорију unclassified може имати вриједности:
    - unclassified