УНИВЕРЗИТЕТ У БЕОГРАДУ

ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

КАТЕДРА ЗА РАЧУНАРСКУ ТЕХНИКУ И ИНФОРМАТИКУ

ЕФИКАСНОСТ МЕТОДА ЗА СМАЊЕЊЕ МУЛТИФАКТОРСКИХ ДИМЕНЗИОНАЛНОСТИ ЗА ОТКРИВАЊЕ БОЛЕСТИ У ПРИСУСТВУ НЕБАЛАНСИРАНИХ СКУПОВА ПОДАТАКА

- мастер рад -

ментор:
проф. др Вељко Милутиновић

студент:
Милан Бранковић

Београд, 2012

Садржај

1 Геномске варијације

Геномске варијације су разлике у секвенци ДНК од једне особе до друге. То је исто као када посматрамо двоје људи и кажемо да су различити, исто тако, са одговарајућим лабораторијским експериментима можемо погледати геноме те две особе и рећи да су и они различити такође. У суштини људи су јединствени као јединке зато што су им геноми јединствени.

1.1 Колико се разликују људски геноми од човека до човека?

Ако посматрамо две особе, што су оне ближе (гледајући родбинске односе), то су им и геноми сличнији. Научници процењују да се геноми особа које нису у родбинском односу разликују на једној од 1200 до 1500 ДНК база. Постоји преко три милиона разлика између генома две посматране особе. Са друге стране ми смо 99,9% слични!

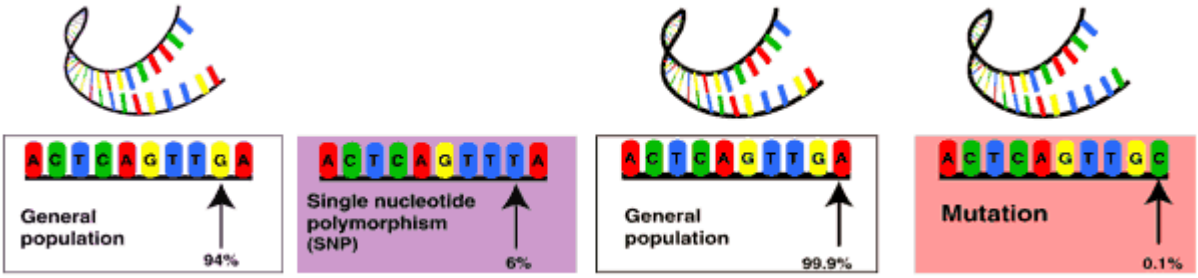
1.2 Зашто је сваки људски геном различит?

Сваки људски геном је различит због мутација – „грешки“ које се јављају повремено у ДНК секвенци. Када се ћелија подели на два дела, она прави копију свог генома, да би свака ћелија имала своју копију. Теоријски, целокупна геномска секвенца је идентична копија оригиналне секвенце, али у пракси погрешна база је убачена у ДНК секвенцу ту и тамо. Ове грешке мењају, или боље рећи промене, се називају мутације.

1.3 Које врсте геномских варијација постоје?

Геномске варијације укључују мутације и поломорфизме. Полиморфизам је ДНК варијација код које се свака могућа секвенца јавља код барем 1% људи.

Термин мутација се често користи када се говори о штетној геномској мутацији која је повезана са неком специфичном болести, док се термин полиморфизам користи када жели да се укаже на варијацију која није ни штетна ни добра. Међутим, научници сада уче да многи полиморфизми заправо утичу на људске карактеристике, на један сложен и понекад неочекиван начин.



(a) (b)

Слика 1. Врсте геномских мутација: (а) једно нуклеотидни полиморфизам (b) мутација

Око 90% геномских варијација долази од једно нуклеотидних полиморфизама (СНП-ова). Како им име говори ове варијације укључују само један нуклеотид, односно базу.

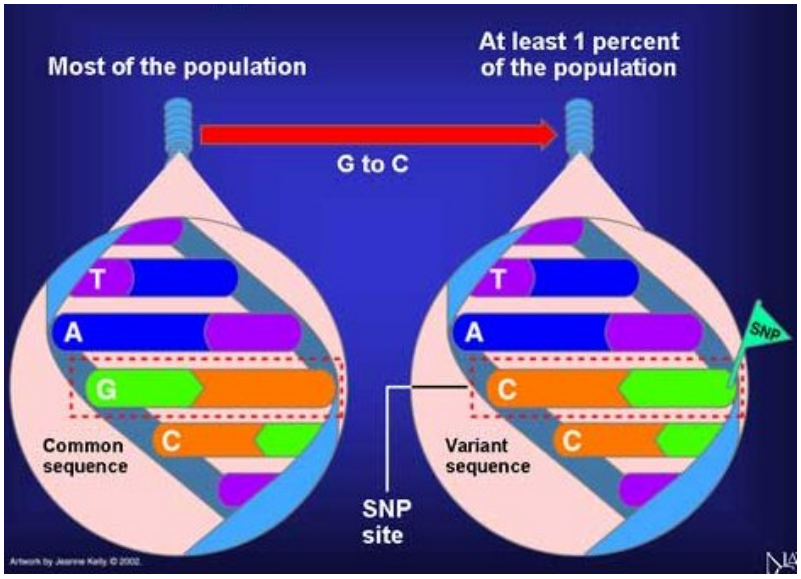
1.4 Зашто проучавати геномске варијације?

Са једне стране, познавање геномских варијација има много практичних апликација. Неке од тих апликација фокусирају се на варијације које утичу на карактеристике неке особе. Друге, користе обе варијанте и оне које утичу и оне које не утичу на крактеристике. Све оне укључују, на један или други начин, анализирање људског генома на присуствно специфичних варијација. И такве анализе нису прецизне или информативне уколико нису поткрепљене основним знањем: колико варијација постоји на геному, где је лоцирана, ...

Научници проучавају геномске варијације да би добили сазнање о људским особинама, посебно оним какве су личност, тежина, или осетљивост на срчана обољења. Мишљење је да на такве особине утиче много различитих гена, тако да студије варијација су једини начин да се разумеју.

Научници верују да ће им многе од ових свеприсутних и малих варијација помоћи да разумеју људску осетљивост на дијабетес, рак, ...

2 Singlenucleotide polymorphisms



Једно нуклеотидни полиморфим (енг. Single-nucleotide polymorphism), СНП, је варијација на ДНК секвенци која се дешава када је један од нуклеотида – А (аденин), Т (тимин), Ц (цитозин) или Г (гуанин) – у генетској секвенци измењен. На пример, два фрагмента ДНК која су у низу од две различите индивидуе АА**А**ГГЦТАА и А**Т**ГГЦТАА, разликују се у само једном нуклеотиду. У том случају кажемо да постоје две алеле А и Т. Скоро сви чести једно нуклеотидни полиморфизми имају две алеле. Да би се варијација сматрала СНП-ом мора да се појави на најмање 1% популације. СНП, који чини око 90% свих људских генских варијација, дешава се на сваких 100 до 300 база међу 3-билионском људском геному. Два од три СНП-а укључују замену цитозина тимином.

Иако су више од 99% људских днк секвенци исте, варијације у днк секвенцама имају велики утицај на то како људи реагују на болест, факторе из окружења као што су бактерије, токсини и хемикалије, лекови и други видови терапија. То чини СНП важним за биомедицинска истраживања и развој фармацеутских производа или медицинске дијагностике. СНП је стабилан – не мења се много од генерације до генерације – што га чини погодним за праћење студија о популацији.

Многи СНП-ови немају утицај на функцију ћелија, али научници верују да други могу преупредити људске болести или њихову реакцију на лекове. Научници такође верују да СНП мапе могу да помогну да се идентификују бројни гени повезани са болестима као што су рак, дијабетес, васкуларне болести и неки облици менталних болести. Ове везе је тешко урадити користећи се конвеционалним методама за уочавање тих гена јер само један промењен ген може допринети развоју болести.

Осим што СНП мапе помажу фармакогенској дијагностици и биомедицинским истраживањима они помажу и идентификовању хиљада додатних маркера у геному.

2.1 Номенклатура

Једно нуклеотидни полиморфизам, је најчешћи тип генских варијација код људи. Сваки појединачни СНП представља разлику у блоку ДНКа који се зове нуклеотид. На пример СНП може заменити нуклеотид цитозин, нуклеотидом тимином у ДНКа низу.

СНП се нормална појава у ДНКа ланцу неке особе. Појава се манифестује у просеку на сваких 300 нуклеотида што значи да има око 10 милиона СНП-ова у генима људи. Најчешће се ове разлике јављају код гена ДНКа. Могу се понашати као биолошки маркери и могу да помогну научницима да пронађу гене повезане са неким болестима. Када се догоди СНП на гену или регији близу гена, онда могу играти велику улогу у болести тако што утиче на функцију гена.

Многи СНП-ови немају утицај на здравље и развој организма. Неке од тих генетских разлика су доказане да су важне за људско здравље. Истраживачи су пронашли СНП који може да помогне приликом предвиђања како ће нека особа реаговати на одређени лек, осетљивост на токсине из околине и ризик од неких болести. СНП се такође може користити за праћење наслеђивања болесних гена у породици. Будуће студије ће се базирати на идентификовању СНП који су повезани са срчаним болестима, раком и дијабетеса.

2.2 Коришћење и значај

СНП-ови не изазивају болести – они су фактори ризика за болести. И много различитих СНП-ова може утицати на ризик на неку болест. СНП-ови не изазивају болест, али могу помоћи да се утврди вероватноћа да ће неко имати одређену болест. Један од гена повезан са Алцхајмеровом болешћу, аполипротеин Е (енг. apolipoprotein E) - *ApoE* - је добар пример како СНП може утицати на развој болести. *ApoE* садржи два СНП-а који резултују са 3 алеле за овај ген- Е2, Е3 и Е4. Свака алела се разликује од днка базе и сваки протеин тог гена се разликује својом амино киселином.

Свака индивидуа наслеђује једну мајчинску копију АпоЕ и једну очинску копију АпоЕ. Истраживања су показала да особа која наследи најмање једну Е4 алелу има већу шансу да добије Алцхајмерову болест. Очигледно промена амино киселине у Е4 протеину мења његову структуру и функцију толико да дозвољава да се болест развије. Наслеђивање Е2 алеле показује да је код те особе мала вероватноћа да ће имати Алцхајмерову болест.

Наравно СНП није апсолутни индикатор развоја болести. Неко које наследио две Е4 алеле можда никада неће имати Алцхајмера док неко ко је наследио две Е2 алеле може да оболи. АпоЕ је само један ген који је повезан са Алцхајмером. Као и чешћи хронични поремећаји као што су болести срца, дијабетес и рак, Алцхајмер је болест коју могу изазвати варијације неколико гена. Полигенска природа ових поремећаја је то што чини генетичко тестирање тако компликованим.

Варијације у ДНК секвенцама код људи могу утицати на то како се код њих развијају болести и одговари организма на патогене, хемикалије, лекове, вакцине... СНП такође могу бити кључни у перонализованој медицини.

2.3 Игла у пласту сена

Проналажење једне нуклеотидне промене у људским генима чини се застрашујуће, али у протеклих 20 година биомедицински истраживачи развили су бројне технике да то учине могућим. Свака техника користи другу методу да упореди одабране регије днк ланца више људи који деле исте карактеристике болести. У сваком тесту резултат показује физичке разлике у узорку днк само када је СНП уочен код једног а не код другог човека.

Многе болести код људи нису изазване генетском варијацијом ни једног гена, него на њих утичу и фактори средине како и интеракције између гена. Иако бројни фактори из средине и начина живота дају велику могућност да се нека болест развије, тренутно је тешко измерити и проценити њихов свеукупни ефекат на процес болести. Зато се овај рад углавном базира на предиспозицијама гена и потенцијалном развоју болести на основу генске мутације.

Генетски фактори могу допринети развоју болести и одредити њен напредак. Ми још увек не знамо све факторе који су умешани у ове промене гена, научницима је тешко да осмисле тестове за бројне болести и поремећаје. Осим што они истражују саме генетске промене као резултат истраживања умешали су се и фактори који нису генски као што су понашање, исхрана, начин живота и физичка активност, а који доста утичу на развој болести.

Генетски фактори могу утицати на реакцију организма на терапију, тако да днк полиморфизми као што је СНП могу бити корисни у помагању истраживачима да одреде и разумеју зашто се индивидуе разликују по питању апсорбовања неког лека као и зашто организам неке особе испољава контраефекта дејства неког лека. Скорашња достигнућа СНП-а дају допринос не само у процесу детекције болести већ и у њеној превентиви и лековитој медицини.

2.4 СНП и дијагноза болести

Генетски материјал сваке особе садржи јединствен СНП шаблон који је сачињен од бројних генетских варијација. Истраживачи су открили да већина СНП-ова није одговорна за стање болести. Они служе као биолошки маркер за обележавање неке болести на људској мапи генома јер су углавном лоцирани поред гена који је/су повезан/и са неком болешћу. Понекад СНП може изазвати болест.

Само је питање времена пре него што се ће лекари бити у могућности да директно испитују ДНК секвенце да би открили неку болест или могућност да пацијент оболи од неке болести.

Да би направили генетски тест који може да лоцира болест код кога је ген који изазива болест већ лоциран, научници су скупили узорке крви од групе људи који су оболели од одређене болести и анализирали њихов ДНК и СНП шаблон. Затим су упоредили шаблоне тако што су анализирали ДНК од групе људи који немају болест. На крају ће добити СНП профиле за карактеристичне болести и обољења. Тада ће само бити питање тренутка пре него што научници буду могли да анализирају ДНК неке особе тражећи одређени СНП шаблон који ће указати на неку болест.

2.5 СНП и лекови

Како је споменуто раније, СНП-ови могу бити повезани са начином реакције на одређени лек. Тренутно не постоји неки једноставан начин да се открије како ће неки пацијент реаговати на лек. Третман који се показао ефикасан код једног пацијента може се показати неефикасним код другог. Још горе, неки пацијенти могу доживети негативну имунолошку реакцију на неки лек. Данас, фармацеутске компаније су ограничене на производњу лекова који делију на „просечног“ човека. Као резултат тога, многи лекови који би помогли малој групи пацијената никад не бивају пласирани на тржиште.

У будућности, најодговарајући лек за неку особу може бити откривен анализом СНП профила те особе. Поступак за проналажење најбољег лека за оболелу особу назива се „персонализована медицина“, и он би омогућио фармацеутским компанијама да пласирају много више лекова на тржиште и дозволио докторима да препишу персонализоване терапије за потребе пацијената.

3 Мутације

3.1 Шта је генетска мутација и како долази до мутација?

У молекуларној биологији и генетици, мутације су случајне промене у ДНК секвенци. Ове случајне секвенце могу се дефинисати као изненадне и спонтане промене у ћелији. Радијација, вируси и мутагене хемикалије су узроци мутација, као и грешке које се догађају приликом копирања ДНК. Генетска мутација је трајна промена у ДНК секвенци која чини ген. Величина мутације може варирати од једног градивног блока ДНК (ДНК база) до великог сегмента неког хромозома.

До генетских мутација долази на два начина: могу бити наслеђене од родитеља или стечене током живота. Мутације које се преносе са родитеља на дете зову се наслеђене мутације или мутације заметка (зато што се оне налазе у јајашцу или сперматозоиду, који се такође називају ћелије заметка). Овај тип мутација је присутан током целог живота појединца у готово свакој ћелији у телу.

Мутације до којих долази на ћелији заметку, или оне које се десе одмах након оплодње, зову се нове (de novo) мутације. De novo мутације могу објаснити генетске поремећаје код којих дете има мутацију на свакој ћелији, али нема забележеног поремећаја у породичној историји болести и обољења.

Стечене мутације – ове мутације се дешавају на ДНК појединачне ћелије током живота неке особе. До ових промена може доћи под утицајем фактора окружења (какво је УВ зрачење, радијација...) или уколико је направљена грешка током клонирања ДНК током ћелијске деобе. Стечене мутације у соматским ћелијама (ћелије које нису ћелије јајашца или сперматозоиди) не могу се пренети на следећу генерацију.

До мутација може доћи на појединачним ћелијама у раној фази ембриоланог развитка. Како се све ћелије деле током раста и развоја, јединка ће имати неке ћелије са мутацијама и неке без генетских промена. Ова ситуација се назива мозиацизам (енг. mosaicism).

Неке генетске промене су веома ретке, док су друге честе. Генетске промене које се појављују на више од 1% популације називају се полиморфизми. Они су довољно чести да се сматрају нормалним варијацијама унутар ДНК. Полиморфизми су одговорни за многе разлике међу људима какве су рецимо боја очију, боја косе, крвна група... Иако већина полиморфизама нема негативне ефекте на људско здравље, неке од ових варијација могу утицати на ризик појаве и развитка одређених поремећаја.

3.2 Како генетске мутације могу утицати на здравље и развој јединке?

Да би исправно функционисала, свака ћелија зависи од хиљаде и хиљаде протеина који раде свој посао на правим местима у право време. Некад, генетска мутација онемогући један или више протеина да одради свој посао ваљано. Мењајући генетске инструкције за

изграђивање протеина, мутација може изазвати да протеин буде неисправан или да уопште не буде изграђен. Када мутација мења протеин који игра критичну улогу у телу, она може пореметити нормалан развој или узроковати неко обољење. Обољење узроковано мутацијом на једном или више гена назива се генетски поремећај.

У неким случајевима, генетске мутације су толико озбиљне да онемогућавају ембриону да преживи до порођаја. Ове промене се догађају на генима који су суштински битни за развој, и често ометају ембрион да се развије у најранијој фази.

Важно је да се примети да гени сами по себи не узрокују настанак болести – генетски поремећаји су изазвани мутацијама које онемогућавају исправно функционисање гена. На пример, када људи кажу да неко има ген цистичне фиброзе (енг. cystic fibrosis gene), они углавном говоре о мутираној верзији CFTR гена који изазива болест. Али сви људи, укључујући и оне без гена цистичне фиброзе имају неку верзију CFTR гена.

3.3 Да ли све мутације утичу на здравље и развој?

Не; само мали проценат мутација изазива генетске поремећаје – већина нема утицај на здравље и развој. На пример, неке мутације мењају ДНК секвенцу, али не мењају функцију протеина од којих је ген сачињен.

Често, генетске мутације које могу изазвати генетске поремећаје бивају поправљене од стране неких ензима, пре него што је протеин изграђен. Свака ћелија има ензиме који препознају и поправљају грешке начињене у ДНК. Како ДНК може бити оштећена или мутирана на разне начине, поправак ДНК је важан процес у коме тело само себе штити од болести.

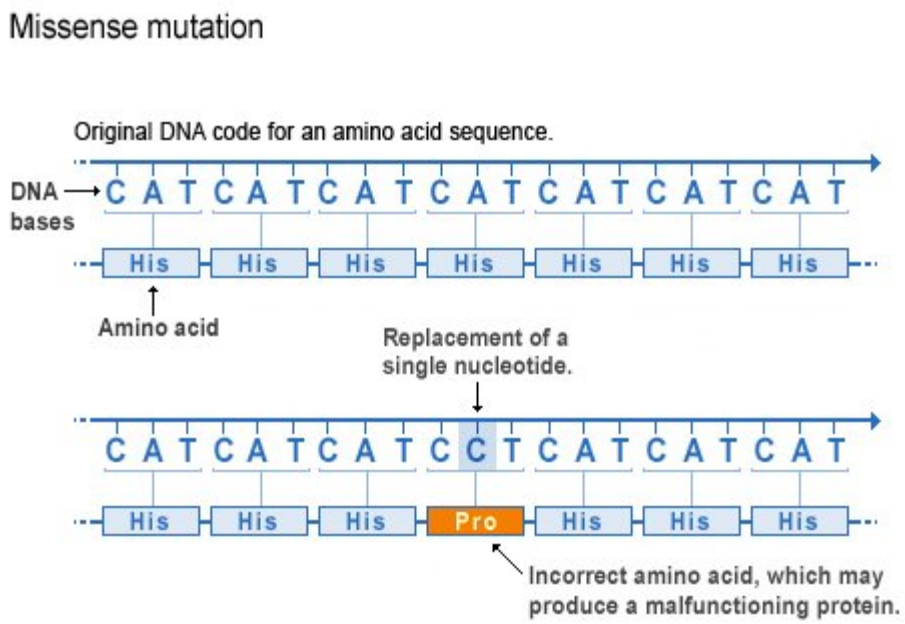
Заправо веома мали проценат мутација има позитиван ефекат. Ове мутације доводе до изградње нових верзија протеина који помажу организму и његовим будућим генерацијама да се прилагоди променама у окружењу. На пример, делотворна мутација може направити протеин који штити организам од нових бактерија.

3.4 Које врсте генетских мутација постоје?

ДНК секвенца може бити измењена на разне начине. Генетске мутације имају различит утицај на здравље, у зависности од тога где се појаве и уколико мењају функцију неких важних протеина. Постоје следеће врсте мутација:

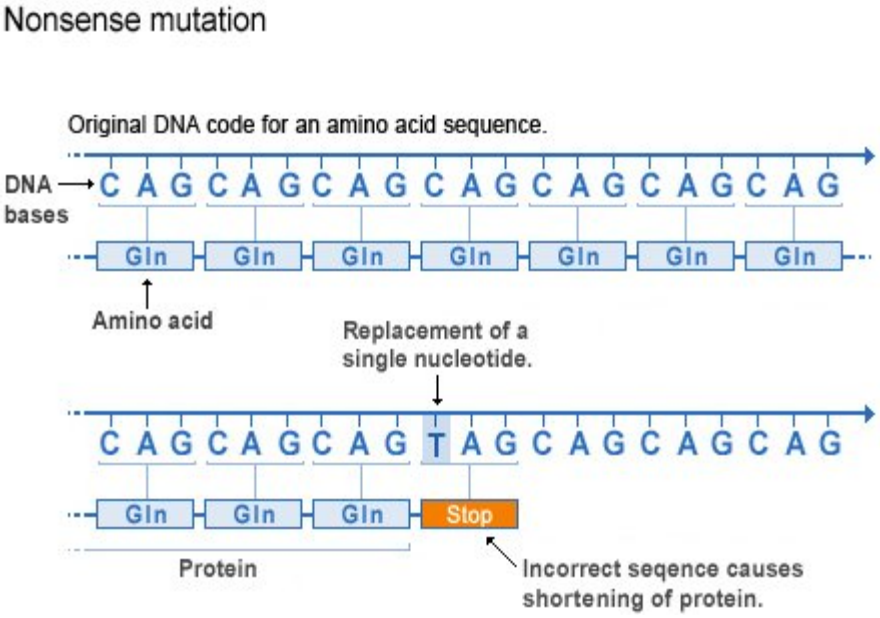
Missense мутација

> Ова врста мутације мења једну базу ДНК која резултира у промени једне аминокиселине другом.



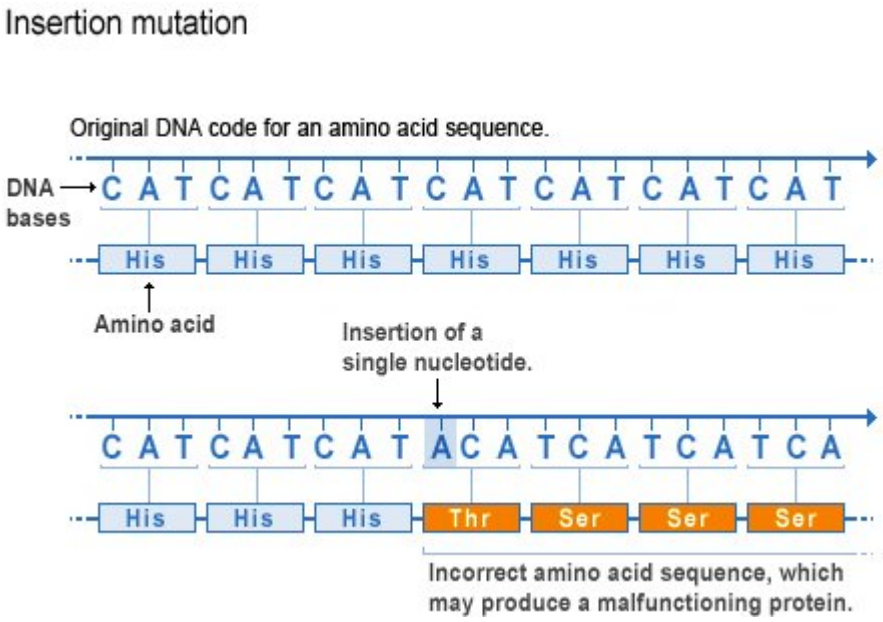
Nonsense мутације

Ова мутација је такође промена једне ДНК базе. Али уместо замене једне аминокиселине другом, промењена ДНК секвенца пре времена сигнализира ћелији да стане са изградњом протеина. Овај тип мутације резултује у скраћивању протеина који може функционисати неправилно или не функционисати уопште.



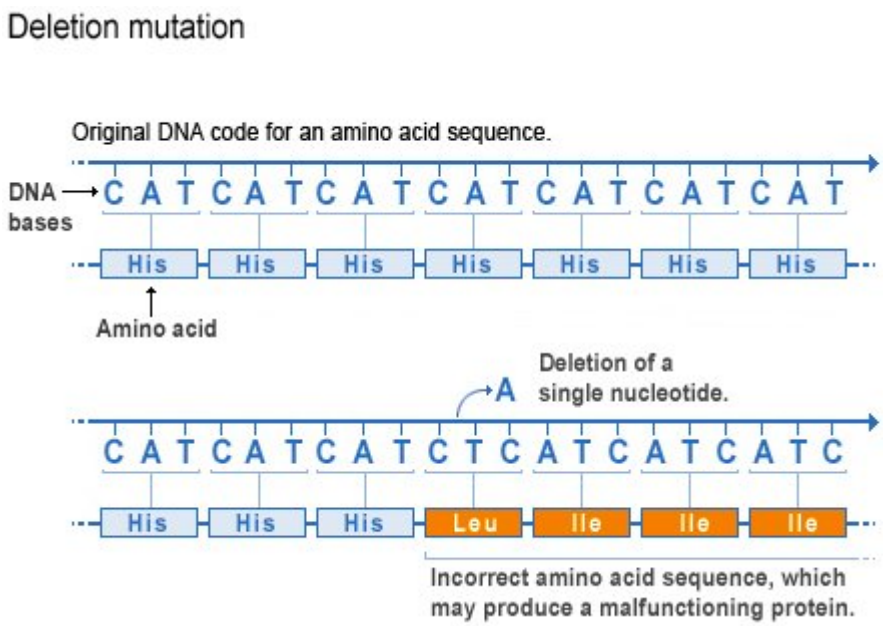
Уметање (енг. Insertion)

Уметање мења број ДНК база додавањем још једног делића ДНК. Као резултат, протеин може да функционише неправилно.



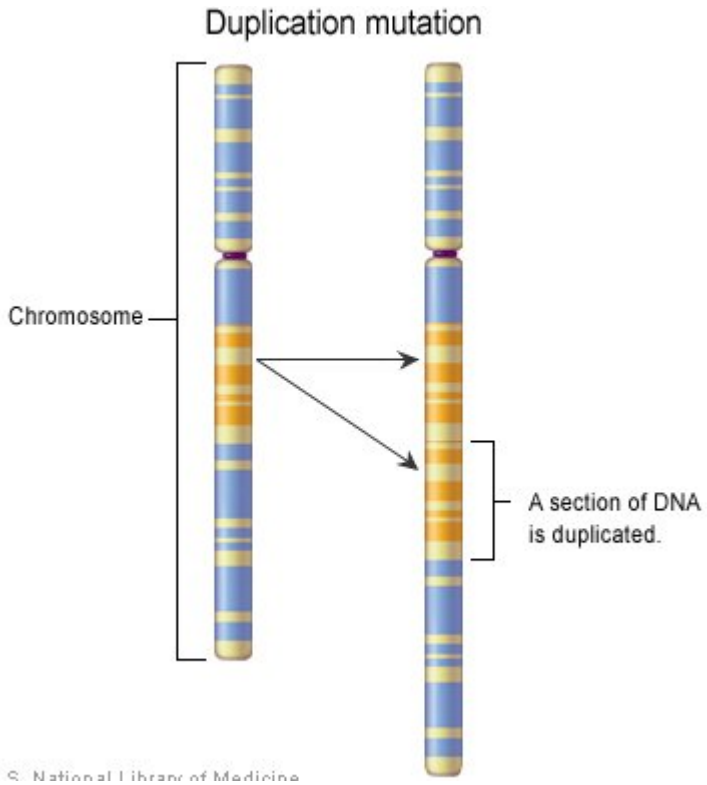
Брисање (енг. Deletion)

Брисање мења број ДНК база уклањајући део ДНК. Брисање може да уклони једну или неколико база, ово је такозвано мало брисање, док велико брисање може уклонити читаве гене или неколико суседних гена. Таква ДНК може изменити функцију резултујућег протеина.



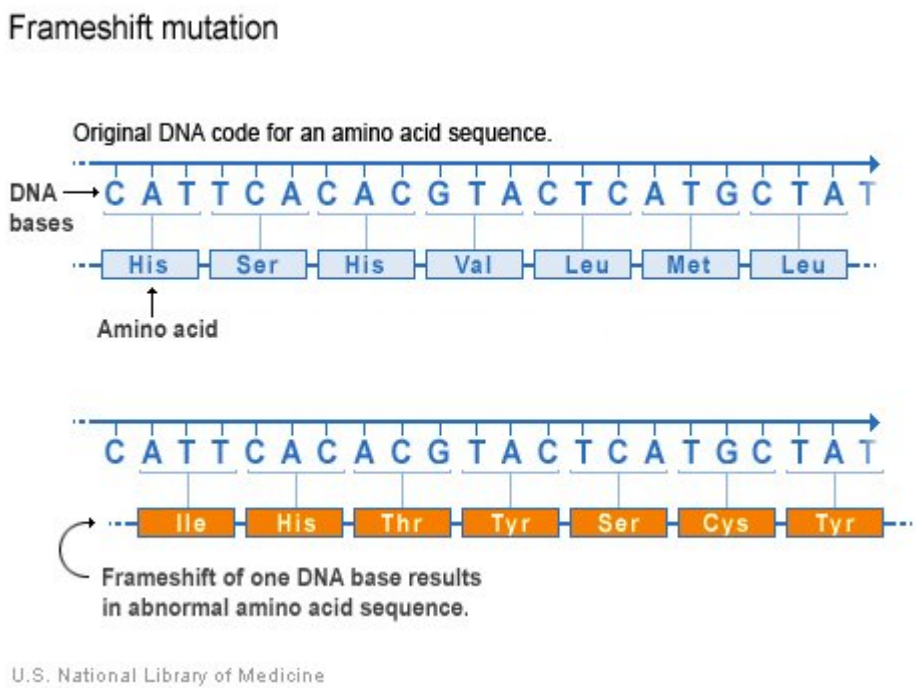
Удвајање (енг. Duplication)

Удвајање се огледа у томе да се делови ДНК ненормално копирају један или више пута. Овај тип мутације може изменити функцију резултујућег гена.



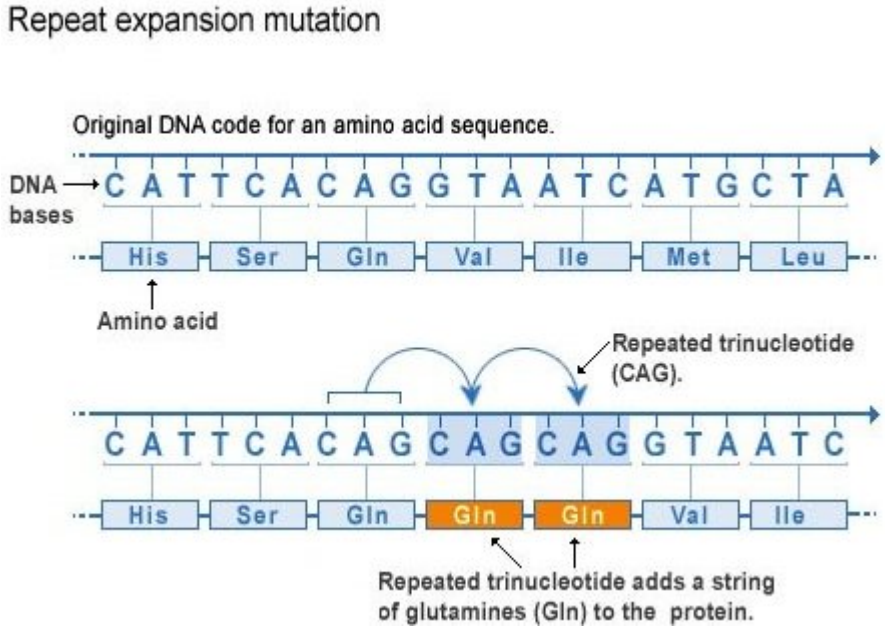
Frameshift мутација

Овај тип мутације се јавља када се додавањем или брисањем ДНК база мења оквир за читање. Оквир за читање се састоји од групе од 3 базе, од којих свака кодира једну аминокиселину. Овај тип мутације смењује груписање база и мења код за аминокиселину. Резултујући протеин је најчешће нефункционалан.



Понављајућа експанзија (енг. Repeat expansion)

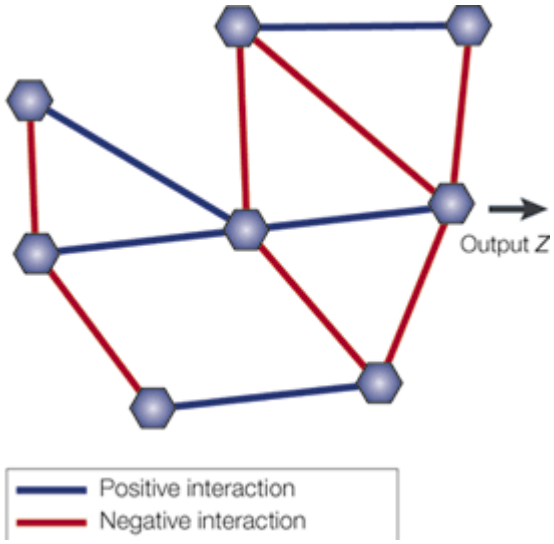
Нуклеотидна понављања су кратке ДНК секвенце које бивају поновљене више пута. Ова врста мутације повећава број понављања неке кратке ДНК секвенце. Он може изазвати да резултујући протеин не функционише правилно.



4 Епистазе

4.1 Шта су епистазе?

Епистаза је интеракција између гена на два или више локуса, таква да се фенотип разликује од онога који је очекиван да су локуси изражени независно. За ген се каже да је епистазан када његово присуство спречава дејство гена на другом локусу. Епистазни гену се понекад називају инхибирајући (енг. inhibiting) гени због њиховог дејства на друге гене који су означени као хипостатички (енг. hypostatic). Треба приметити да епистаза није исто што и доминација. Код епистаза мутација једног гена маскира изражавање неког другог гена, док код доминације један алел гена маскира изражавање другог алела истог гена.



Слика 2. Интеракција између гена

„Када фактор једног пара маскира изражавање фактора неког другог пара, каже се да је он епистатичан... Именица формирана од прилога епистатичан је епистаза. Епистаза је према томе исти ефекат који се испољва између фактора два или више парова“ - Laurence Snyder, Ph.D. (1901-1986)

Израз „епистатичко“ је први пут употребио Bateson 1909 да би описао маскирајући ефекат где алел на једном локусу спречава алел на другом локусу да испољи своје дејство. Ово је виђено као проширење концепта доминације алела на једном локусу. Претпоставимо да имамо два локуса, B и G који утичу на неку особину, рецимо боју косе. Локус B има две могуће алеле *B* или *b*, и локус G такође има две могуће алеле, *G* или *g*. Могући фенотипски исходи (бела, црна или сива коса) за дати генотип дата је у табели 1. Видимо да, без обзира на генотип на локусу B, индивидуе са било којом копијом *G* алеле имају сиву косу. На локусу G, алела *G* је доминантна у односу на алелу g, маскирајући било који ефекати алеле g. Такође можемо приметити ако је генотип на локусу *G* g/g, тада индивидуа са било којом копијом B алеле има црну косу, тако да на локусу B, алела B је

доминантна у односу на b. Ако генотип на локусу G није g/g, тада ефекат на локусу B приметан, јер индивидуе са било којом копијом G алеле има сиву косу, без обзира на генотип на локусу B. Ефекат на локусу B је маскиран од оног на локусу G – за локус G се каже да је епистатичан у односу на локус B (или специфичније, алела G на локусу G је епистатична у односу на алелу B на локусу B).

	Генотип на локусу **G**		
Генотип на локусу B	***g/g***	***g/G***	***G/G***
b/b	Бела	Сива	Сива
b/B	Црна	Сива	Сива
B/B	Црна	Сива	Сива

Табела 1

У људској генетици, интересантан фенотип је често квалитативан, указујући на присуство или одсуство болести. Математички модели уједињених акција на два или више локуса се углавном фокусирају на могућности за развој болести на датом генотипу. Претпоставимо да имамо два локуса A и B. Локус A има две могуће алеле A или a, док локус B има такође две могуће алеле B или b. Претпоставимо такође да је предиспонирајућа алела обавезна на оба локуса да би се испољила нека особина болести, једна или више копија обе алеле A или B је обавезна. У том случају, оба локуса се узимају у обзир. Резултат овога је приказан у Табели 2. У овој табели, ефекат алеле A се може посматрати једино када је алела B такође присутна. Ефекат на локусу A се чини да је маскиран од оног на локусу B. У аналогији са примером из Табеле 1, можемо рећи да је локус B епистатичан у односу на локус A, јер када је генотип b/b присутан на локусу B, ефекат алеле на локусу A није приметан. Међутим, исто тако можемо рећи да је локус A епистатичан у односу на локус B, јер када је генотип a/a присутан на локусу A, ефекат алеле на локусу B није приметан.

	Генотип на локусу B		
Генотип на локусу A	***b/b***	***b/B***	***B/B***
a/a	0	0	0
a/A	0	1	1
A/A	0	1	1

Табела 2

Овде ћемо дати дефиницију два типа епистаза: биолошке и статистичке епистазе. Биолошке епистазе су оне епистазе које су резултат физичких интеракција између биомолекула (нпр. ДНК, РНК, протеини, ензими...) и јављају се на ћелијском нивоу код индивидуе. Статистичке епистазе, са друге стране, јављају се на нивоу популације, а реализују се када постоји индивидуална варијација у ДНК секвенци. Веза између биолошких и статистичих епистаза је често збуњујућа, али важно је да се разуме да ми желимо да изведемо неки биолошки закључак из статистичких резултата.

Епистазе играју важну улогу у генетичком инжињерству код честих људских болести и могу се сагледати из две различите перспективе, биолошке и статистичке, свака изведена из другачијих претпоставки и истраживачких стратегија. Биолошке епистазе су оне

епистазе које су резултат физичких интеракција између биомолекула у оквиру регулационих генетских мрежа и биохемијских путева код неке индивидуе, такве да ефекти гена неког фенотипа су зависни од једног или више других гена. Насупрот томе статистичке епистазе се дефинишу као одступање у неком математичком моделу, сумирајући однос између мултилокусних генотипова и фенотипских варијација у популацији.

Статистичке епистазе се тешко откривају и карактеришу у студијама случаја људског организма, због нелинеарности интеракција. У екстремној форми, епистазе се могу јавити у одсуству детектованих независних ефеката једног полиморфизма. Ово прави неколико врло тешких рачунских и статистичких изазова, посебно у контексту широкопојасне генетске студије. Прво, моделовање нелинеарних интеракција захтева специјалне аналитичке методе зато што параметарско-статистички приступ, као на пример логистичка регресија, може имати мање успеха при проналажењу интеракција. Друго, недостатак главних независних ефеката има значајне импликације код широкопојасних генетских студија са стотинама и хиљадама СНП-ова, зато што стратегије за претрагу укључују похлепне hill-climbing алгоритме, који обављају многа израчунавања. Како открити неку специфичну нелинеарну интеракцију када је потребно проценити небројено много комбинација? Исцрпно претраживање није могуће, а похлепни алгоритми нису ефективни. На крају споменимо и то да модели засновани на комбинацијама СНП-ова ће бити тешки за тумачење због великог броја димензија.

Епистазе се појављују у следећим стањима:

- Када два или више локуса интерагују да би створили нови фенотип
- Када алела на једном локусу маскира ефекат алеле на једном или више других локуса
- Када алела на једном локусу мења ефекат једне или више алела на другим локусима

4.2 Колико су честе епистазе код болести?

Епистатичке интеракције су можда много чешће него што ми мислимо. Неки научници мисле да су епистазе свеприсутне у биологији и да су дуго биле игнорисане у студијама људских болести. Гени једне јединке не делују самостално, већ је јасно да функционишу заједно у ћелијском окружењу. Стога, очекивано је да се догоди понека интеракција између гена.

Епистазне интеракције могу да загорчају живот научницима који се баве истраживањем који су гени одговорни за појаву и развој људских болести. На пример, резултати већине студија које се фокусирају на неком обећавајућем кандидат гену, нису биле у могућности да потпуно објасне болест када се у обзир узело више пацијената са истом болести. Ово је довело до закључка да је више гена укључено и да они интерагују међусобно да би или повећали или умањили дејство болести. Уколико је ген који носи болест маскиран дејством неког другог гена, проналажење тог гена постаје веома компликовано. Штавише, уколико постоји више епистазних интеракција, проналажење гена и веза које утичу на

њега постаје готово нерешив задатак. Постоје, међутим, бројни начини да се истраже епистазе употребом неких специфичних метода.

Када размишљамо о факторима који утичу на формирање неке болести, често мислимо само на неке специфичне мутације на генима и на факторе из окружења који подстичу развој болести. Али веома је важно узети у обзир и епистазе, које укључују интеракцију између два или више гена. У ствари, разумевање епистазних интеракција је кључ за разумевање болести.

4.3 Разумевање болести

Сада је могуће пронаћи везе између гена, као и епистатичке интеракције. Данас, на располагању су разни алати за експериментисање који мере молекулске и биохемијске податке. На пример, ДНК низови дозвољавају научницима да сакупе стотине хиљада података који потичу из неке ћелије, са нивоом транскрипције који је довољан да би се измерио неки фенотип. Онда може да се употреби неки рачунарски и/или биоинформатички метод да се прореде и сортирају ове енормне количине података да би се фокусирали на епистатичке интеракције. Једном када пронађемо и размемо епистатичке везе, можемо употребити добијено знање да побољшамо дијагнозу и третман неке болести.

4.4 Примери епистаза

Ћелавост. Иако ген за ћелавост код мушкараца није лоциран на X хромозому, пол индивидуе одређује да ли је алела за ћелавост доминантна или рецесивна. Код мушкараца ова алела је доминантна, док је код жена она рецесивна. Алел за ћелавост ће бити представљена са b, док ће алела за „косматост“ бити представљен са W. Ако би генотип особе био bbW_ или BbW_, ћелавост код особе би била испољена.

Албинизам. Алела за албинизам ће бити представљена са a, док ће алела за боју косе бити представљена са B (смеђа) и b (плава). Ако би генотип особе био B_aa, албинизам би маскирао смеђу боју косе. Да је генотип био bbaa, албинизам би маскирао плаву боју косе. Алела која утиче на албинизам би такође маскирала боју очију код особе (доводећи до губитка пигмента), као и боју коже. Особе са албинизмом ће бити беле косе, црвених очију и веома светле пути.

Црвена коса. Алела за смеђу боју косе ће бити представљен са B, плава са b, нормална са R и црвена са r. Ако би генотип особе садржао rr, особа би имала црвену косу. Ово се дешава зато штп алел за црвену боју косе маскира алел за плаву или смеђу боју косе.

Дијабетес. На ову болест утичу како епистазе тако и фактори из окружења. Иако се зна да особе са дијабетесом имају низак ниво инсулина и висок ниво шећера у крви, специфични фактори који доводе до појаве болести још увек нису откривени. Интеракције су откривене на локусима хромозома 2 и 15, као и на локусима хромозома 1 и 10.

Крвна група. Врло редак ген, односно његов рецесиван алел h, када се нађе у хомозиготном стању (hh), кочи стварање антигена A и B у крви човека. Тада се не

испољавају крвне групе А, В и АВ јер, иако особе имају алеле А или В (или оба), њихово дејство је маскирано овим рецесивним геном h. Тако нпр. иако особа има генотип ААhh она неће имати крвну групу А јер је алел А маскиран. Особа ће тада имати О крвну групу.

Још неки примери епистаза су: кардиоваскуларне болести, хипертензија, аутизам, шизофренија и остали неуролошки поремећају, као и рак дојке, рак бешике и други типови рака.

5 МДР

Проналажење и карактеризација осетљивости гена на комплексне болести које погађају људски род је један он највећих изазова са којима се сусрећу генетичари данашњице. Овај изазов бива још већи знајући да су параметарско-статистички методи ограничени (нпр. они у којима се поставља нека хипотеза на основу вредности статистичких параметара) за проналажење генетских ефеката који су делимично или потпуно зависни од интеракција са другим генима и/или изложености утицају животне средине. На пример, метод логистичке регресије се често користи за моделовање везе између дискретних предиктора, какви су геноми/генотипови, и дискретних клиничких исхода. Међутим овај метод, као и већина параметарско-статистичких метода, је мање практичан при проналажењу резултата код мултидимензионалних скупова података. Када се моделују интеракције, многи подаци недостају (ћелије су празне), што може да доведе до великих грешака при процени. Једно решење овог проблема је да се сакупи довољан број узорака и да се дозволи робусна процена ефеката интеракције, међутим, количина узорака који су потребни је често прескупа опција. Алтернатива је да се побољша метод за идентификовање ових ефеката на релативно малом узорку.

Овде ћемо кратко представити метод за смањивање количине информација о мултилокусима чији је основни циљ да побољша идентификацију полиморфних комбинација које су повезане са ризичним обољењима и болестима. Метод смањења мултифакторским димензионисањем (енг. multifactor-dimensionality reduction), у даљем тексту МДР, је непараметарски (не уводи се ниједна хипотеза о вредностима статистичких параметара), model-free (не претпоставља ни један посебан модел наслеђивања) и употребљив је директно на case-control и discordant-sib-pair студије случаја.

МДР је data mining стратегија за детекцију и карактеризацију комбинација атрибута или независних променљивих/атрибута (нпр. СНП, пушење, пол...) који интерагују, а које утичу на неку зависну променљиву. МДР је направљен специјално за идентификацију интеракција између дискретних променљивих које утичу на бинарни исход и сматра се непараметарском алтернативом за традиционалне методе (нпр. метод логистичке регресије). МДР софтверски пакет комбинује избор атрибута, конструкцију атрибута и класификацију са унакрсном валидацијом (енг. cross-validation) да обезбеди моћан приступ за моделовање интеракција.

База МДР метода је конструкција индукованог алгоритма који конвертује два или више независних атрибута у један. Овај процес конструкције новог атрибута мења представу простора података. Циљ је да се направи или открије неки приказ који олакшава детекцију нелинеарних и неадитивних интеракција између атрибута таквих да предвиђање класног атрибута представља побољшану верзију првобитне представе података.

Са МДР-ом, генотипови су подељени у две групе: високо и ниско ризичне групе, смањујући н-димензионални простор предикција на само једну димензију! Ова димензија је у ствари процена да се класификује и предвиди статус неке болести користећи се унакрсним испитивањима и тестирањем помоћу пермутација.

Прикажимо то на једном простом примеру, узмимо функцију ексклузивно ИЛИ (XOR). Ексклузивно ИЛИ је логички оператор који се често користи у data mining-у и код машинског учења (енг. machine learning). Следећа табела приказује једноставан сет података где релације између атрибута (X1 и X2) и класне променљиве је дефинисана функцијом ексклузивно ИЛИ: Y = X1 XOR X2.

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Табела 3

Било који data mining алгоритам би требао да открије апроксимацију функције ексклузивно ИЛИ да би тачно предвидео Y, користећи информације о атрибутима X1 и X2. Алтернативна стратегија била би да се прво промени репрезентација података користећи се конструктивном индукцијом. МДР алоритам би променио репрезентацију података (X1 и X2) на следећи начин. Најпре се бирају два атрибута. У овом простом примеру, X1 и X2. Свака комбинација вредности атрибута X1 и X2 се испитује и број појављивања Y=1 и/или Y=0 се броји. У овом примеру Y=1 се не појављује ниједном, а Y=0 се појављује једном за комбинацију X1=0 и X2=0. Са МДР-ом овај однос се израчунава и упоређује са фиксним прагом. Овде је тај однос 0/1 што је мање од нашег прага 1. Пошто је однос 0/1 < 1, уводимо нови атрибут (Z) са вредношћу 0. Када је однос већи од један, вредност новог атрибута је 1. Овај процес се понавља за све јединствене комбинације вредности X1 и X2. Табела 4 представља нашу трансформацију података.

Z	Y
0	0
1	1
1	1
0	0

Табела 4

Сада data mining алгоритам има много мање посла да би пронашао добру предиктивну функцију. У овом простом примеру, функција Y = Z има тачност класификације 100%. Лепа одлика конструктивне индукције је могућност да се искористи било који data mining или machine learning метод да се анализира нова репрензетација података.

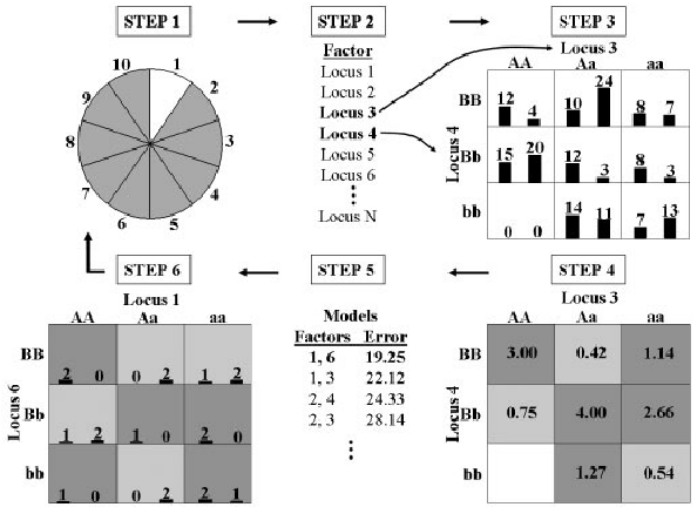
5.1 Data mining са МДР-ом

Како је показано у претходном примеру, база алгоритма конструктивне индукције код МДР-а је веома једноставна. Међутим, његова имплеметација за проналажење образаца користећи се реалним подацима, може бити веома сложена. Као код било којег data mining алгоритма увек постоји бојазан од претренирања. Ови алоритми могу наћи образац користећи било какве случајно изабране податке. Веома често је тешко да се утврди да ли је добијени образац стваран или је само могућност. Једна могућност је да се процени

генералност модела, користећи се независним скуповима података, а уз помоћ укрштене валидације. Друга могућност је да се генеришу многобројне случајне пермутације података и да се види какав ће образац data mining алгоритам наћи, али укључујући могућност претренирања.

5.2 Алгоритам рада МДР алата

Слика 2 илуструје шест корака која су укључена у имплементацију МДР метода за студије случаја case-control. У првом кораку, скуп података се дели на сет података за тренирање (90%) и независни скуп података за тестирање (10%). У кораку 1а, бира се скуп од н генетских и/или фактора окружења из скупа свих расположивих фактора. У корацима 2 и 3, тих н фактора и њихове могуће мултифакторске класе бивају представљене у н-димензионалном простору; нпр. за два локуса са тру генотипа за сваки, постоји девет могућих дво-локусних генотипских комбинација. Затим се израчунава однос cases/controls за сваку мултифакторску класу. У кораку 4, свака ћелија у н-димензионалном простору се обележава као високо- (ако је однос cases/controls већи или једнак од неке границе), или као ниско-ризичан. На овај начин, модел за cases и controls се формира тако што се праве групе од ћелија високог и ниског ризика. Овај начин смањује н-димензионални модел на само једну димензију. У коришћеној имплементацији МДР, потребне ус избалансиране студије случаја case-control. У корацима 5 и 6, грешка предикције за сваки модел се израчунава и бира се најбољи модел на основу израчунате грешке. У овом кораку се скуп дели насумично на 10 једнаких делова. МДР модел је развијен за сваку могућност од 9/10 субјеката, а затим се користи да би предвиђао статус болести на преосталих 1/10 субјеката. Проценат испитаника за које је направљена погрешно предвиђање јесте процена грешке предвиђања.



Слика 2. Генерални приказ алгоритма рада МДР алата

За студије које укључују више од два фактора, ови кораци се понављају за сваку могућу величину модела, ако је то рачунарски изводљиво. Резултат је сет модела, један за сваку величину.

5.3 Начин функционисања МДР-а

5.3.1 Како обрадити податке који недостају?

МДР алат изискује комплетан скуп података, без података који недостају, да би могао да спроведе анализу. Постоје три начина да се проблем података који недостају реши:

- Најједноставније је избацити редове који не садрже довољан број атрибута. Ово је најнепожељније решење зато што на крају одбацујемо половину сета
- Друго решење је да се недостајући подаци означе неком вредношћу која се не користи. На пример уколико су вредности атрибута 0, 1 и 2, недостајуће податке можемо кодовати са 4 или 17. На овај начин се подаци укључују у модел. Ова опција је најпримењивија све док постоји мали број недостајућих генома, и сви они су разбацани у посматраном сету података
- Треће решење је да се користи неки статистики модел који предвиђа недостајуће податке. МДР алат користи овај метод, предвиђајући највероватнију вредност гледајући број понављања генома. Ово је такође најпримењивија опција ако имамо мало генома који су разбацани у посматраном скупу података. Добра страна овог решења је та да целокупан сет података остаје нетакнут, у смислу да сви редови остају у скупу, односно ништа се не одбацује

5.3.2 Филтрирање података

МДР алат спроводи исцрпљујућу претрагу свих могућих комбинација двојки, тројки,..., н-торки атрибута (СНП-ова). Овај начин је добар за студије случаја где постоји до 100 СНП-ова, зато што ће резултати бити добијени у неком разумном временском року.

Када постоји више од 100 СНП-ова, овај приступ је непрактичан, осим уколико не желимо да чекамо данима или недељама да би смо добили резултат. Када број СНП-ова премаши 10 000, исцрпна претрага свих комбинација тројки постаје неисплатива, чак и на паралелној машини.

Тренутно постоје два начина широкопојасне анализе генома. Помоћу прве издвајамо 100 или 1000 СНП-ова, из сета од 100 000 кандидата. Ово је такозвани приступ прочишћивањем (енг. filter). Други начин је да се користи неки стохастички алгоритам да се нађе оптимална комбинација. Ово је такозвани омотачки (енг. wrapper) приступ.

Постоје многобројни филтери који могу бити примењени на СНП податке да се смањи број атрибута који су потребни да би се извршила систематична анализа уз помоћ МДР-а. МДР алат користи Relief фамилију алгоритама: ReliefF, Tuned ReliefF и Spatially Uniform ReliefF.

Статистички и биолошки филтери могу нам помоћи да смањимо број СНП-ова који ће бити коришћени у исцрпној анализи. Добра страна овог приступа је да је он рачунарски праћен. Лоша страна је да листа филтрираних СНП-ова је добра колико и филтер који се користи. Свакако постоји могућност да ће неки добри СНП-ови бити избачени из анализе.

5.3.3 Анализа

Након трансформације сета података и филтрирања истог, можемо извршити анализу.

Први корак је да учитамо фајл у коме се налази наш сет за анализу. Уколико смо успешно учитали фајл, можемо видети целокупан сет ако кликнемо на дугме View Datafile. У пољу Datafile Information можемо пронаћи информације о фајлу, као и о броју СНП-ова. Тренутна верзија може да учита сет података који садржи до 1000 СНП-ова.

Када смо спремни да извршимо анализу, потребно је само да кликнемо на дугме Run Analysis. Анализа ће бити извршена користећи подразумевана подешавања, која је могуће применити уколико је потребно да извршимо неку специфичну анализу.

5.3.4 Преглед резултата

Након извршене анализе, можемо погледати резултате. Постоје више резултата које можемо посматрати, а који су одвојени у засебне табове.

Analysis таб – у овом одељку налази се табела са четири колоне. Прва колона говори који је најбољи модел за сваки изабрани случај. Требало би да видимо четири реда за сваки од модела са једним, два, три и четири локуса. Тачност тренинга (енг. Training Accuracy) налази се у другој колони. Ова тачност се користи да ви се изабрао најбољи модел из прве колоне. Свака вредност која је већа од 0,55 сматра се за интересантну. Треба напоменути да је тачност пропорција инстанци или субјеката који су коректно класификовани као case или као control. Формална дефиниција је Accuracy = (TP + TN)/(TP + TN + FP + FN), где је TP - true positives, TN - true negatives, FP - false positives и FN - false negatives. Тачност ради најбоље док год је број cases и controls мање-више балансиран или једнак. Овде се користи балансирану тачност (енг. balanced accuracy) која се показала као најбоља за сетове података де је једна класа много већа од друге. Ова тачност се дефинише на следећи начин (sensitivity+specificity)/2 где је sensitivity = TP/(TP+FN) и specificity = TN/(FP+TN). Овај приступ нам даје процењену тачност која није пристрасна већој класи. Ово комбинујемо са прагом за дефинисање високо и ниско ризичних комбинација генотипова која је једнака односу cases наспрам controls у изабраном скупу података. Тачност и балансирана тачност дају исте резултате када је сет података балансиран. Последња колона Summary табеле показује CV Consistency (CVC). Ово је статистика направљена да прати број понављања истог модела који је пронађен користећи се различитим поделама скупа података на различите сегменте. „Прави сигнал“ треба да буде увек пронађен независно од начина поделе скупа података. Добар модел обично има CVC око 9 или 10 (уколико користимо поделу на десет CV преклопа). CVC је користан из два разлога. Прво, може се користити да нам помогне да се одлучимо за најбољи модел (ако постоји више модела који имају исте резултате), и други, ако је CVC<10, то нам указује да је тај модел можда пристрасан неком CV интервалу.

У дну прозора можемо видети до шест табова, који могу бити коришћени за интерпретацију резултата:

1. Graphical Model таб приказује дистрибуцију cases (лева страна)/controls (десна страна) за сваки генотип или комбинацију нивоа. Поља високог ризика су осенчена

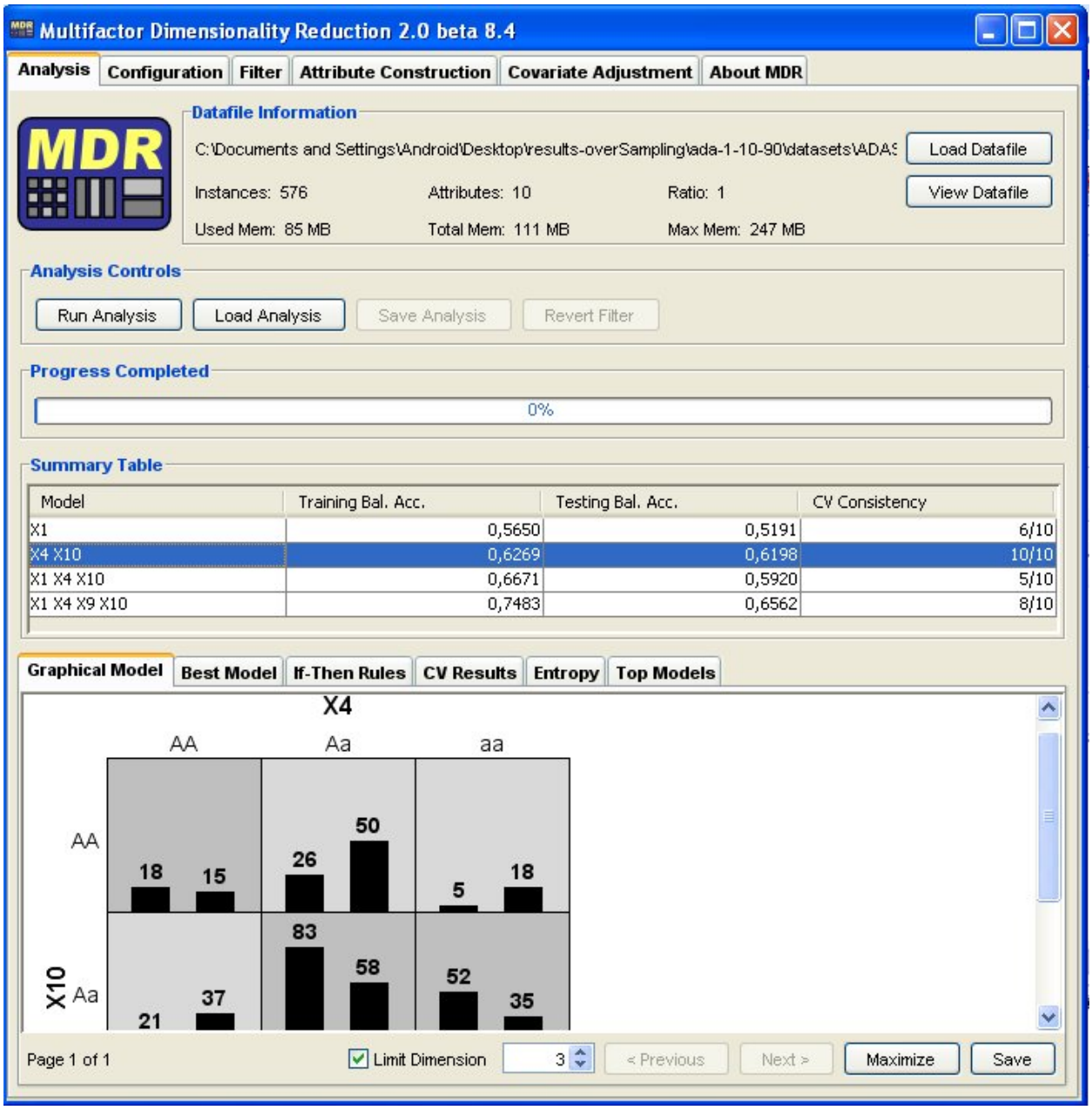
тамно сивом, док су поља ниског ризика осенчена светло сивом бојом. Треба приметити да МДР комбинује и високо и ниско ризичне комбинације правећи један нови атрибут користећи конструктивну индукцију. Над овим атрибутом се врши статистичка обрада.

2. Best Model таб даје неке додатне статистичке информације за најбољи модел који је одабран у Summary Table. Када је изабран најбољи модел може се направити статистика за целокупан сет података.
3. IF-THEN правила је другачији начин приказа МДР модела. Ово су правила која су коришћена се издвоје високо и ниско ризични геноми/генотипови.
4. CV Results таб нам даје статистику и најбољи модел идентификован на неком CV интервалу.
5. Entropy таб се користи за статистичку интерпретацију МДР модела. Дендограм користи теорију информација и анализу кластера да идентификује синергетске, независне, и редундантне генетске ефекте.
6. Ако је изабрана опција Landscape (у табу Configuration) овај таб би требао да буде видљив. Ово је место где се могу проучавати сви модели добијени помоћу МДР анализе. Ово је веома корисна опција за упоређивање првог са другим, трећим или чак четвртим најбољим моделом.

Configuration таб – у овом одељку се налазе параметри који се могу подешавати ради извршавања неког специфичног експеримента. Овде можемо изабрати тип претраге, која у неким случајевима функционише брже и боље уколико се изабере неки други тип. Пример бржег рада претраге: постоји велики скуп података (преко 10000 података); тада исцрпна претрага дуго траје и боље је изабрати неки други тип.

Filter таб – у овом одељку можемо изабрати неки филтер и извршити филтрирање података уколико имамо неки велики скуп података, а желимо да га смањимо.

Постоје још два таба – Attribute construction и Covariate adjustment. Ови одељци нису коришћени приликом извршавања експеримената.



Слика 3. МДР алат

5.3.5 Интерпретација

МДР је data mining приступ који користи конструктивну индукцију (енг. constructive induction) или конструкцију атрибута (енг. attribute construction) да олакша проналажење интеракције између генома, при одсуству стварног ефекта. Најчешћа критика data mining и machine learning метода је та да ови методи користе улазне параметре и генеришу излазни параметар/е без разумевања процеса у средини који обрађује добијене информације (модел „црне кутије“ (енг. black box model)). Феномен „црне кутије“ може довести до тога да се модел пронађен помоћу неког data mining метода тешко тумачи.

Важан концепт у data mining-у је добитак на информацијама (енг. information gain) који се базира на измереној ентропији. Односно колико информација је добијено о статусу case-control од знања о генотиповима на једном или више СНП-ова? Речено на други начин колико је велика ентропија статуса case-control узимајући у обзир само овај генотип. За сврхе МДР анализе ми желимо да видимо колико информација о case-control статусу добијамо комбинацијом два или више СНП-ова користећи се МДР атрибутивно-конструктивном функцијом. Тачније, да ли добијамо нешто више ако користимо одређене СНП-ове појединачно?

Идеја је једноставна. Комбинацијом два или више СНП-ова, користећи МДР, ако добијемо позитивни исход тада имамо доказ да постоји синергетичка интеракција. Ако добијемо негативан исход, тада је информација изгубљена и бескорисна што се дешава када имамо редундансу или корелацију. Ако немамо добитак ни губитак, можемо закључити да СНП-ови имају независне ефекте. Дендограм у Dendrogram табу је генерисан на следећи начин. Прво се израчунава добитак за сваки СНП у summary табели и онда се исто ради за комбинације парова. Тада се вредности резултата инвертују и чувају у матрици растојања, тако да СНП-ови са већом интеракцијом имају мање растојање. Ова матрица се користи да спроведе хијерархијску анализу кластера чији резултат нам даје дендограм интеракције. Што је краћа линија између два атрибута, већа је интеракција између истих. Боја линије даје нам информацију о типу интеракције. Црвена и наранџаста указују да постоји синергетичка повезаност (нпр. епистаза), жута нам говори о независности атрибута, зелена и плава указују на редундансу или корелацију. На овај начин се брзо може прегледати генерисани дендгорам и индентификовати ефекат епистазе добијен МДР анализом.

5.4 Добре стране МДР-а

Најважнија карактеристика МДР-а је та да олакшава истовремену детекцију и карактеризацију више генетских локуса повезаних са дискретним клиничком обољењем. Ово се постиже на тај начин што се мултидимензионални простор претвара у једнодимензионални. Срж овог приступа је да се генотипови из више локуса и/или дискретне класе из окружења раздвајају у две групе: ниско и високо ризичне групе, у зависности да ли су више слични безбедном или небезбедном стању.

Још једна добра страна МДР-а је да је овај метод непараметарски. Ово је важна карактеристика која је тотално другачија наспрам традиционалним параметарско-статистичким методама, који се базирају на генерализованом линеарном моделу. На пример, код логистичке регресије, сваки додатни ефекат се укључује у модел, тако да број могућих интеракција експоненцијално расте.

Трећа предност је да се не претпоставља никакав одређени генетички модел (енг. model free), односно не треба навести некакав модел наслеђивања. Ово је веома важно за болести, какав је рак дојке, у којој је модел наслеђивања непознат и врло вероватно веома комплексан.

Четврта предност је да false-positive резултати услед вишеструког тестирања бивају минимизирани. Ово се пре свега догађа због унакрсног упаривања коришћеног да се изабере најоптималнији модели. Редукција скупа података и методи за проналажење шаблона (енг. pattern-recognition methods).

Методи за редукцију података (енг. data-reduction) и препознавање шаблона (енг. pattern-recognition) су добри за идентификацију комплексних релација између података. Међутим, прави квалитет неког метода се огледа у његовој способности да направи добро предвиђање над независним скуповима података.

5.5 Недостаци и ограничења МДР-а

Иако МДР превазилази нека ограничења генерализованог линеарног модела, он садржи четири веома важна недостатка:

1. МДР може да захтева превише израчунавања, посебно када желимо да направимо евалуацију где желимо да укључимо више од десет полиморфизама. Претраживање генома са сто или хиљаду полиморфизама захтева робусни алгоритам, зато што све комбинације мултилокуса не могу да се претражују док се не исцрпе све могућности. Међутим, ово није само проблем МДР-а, већ је ово генерални проблем свих алгоритама.
2. МДР модели могу бити тешки за интерпретацију.
3. МДР, у форми у којој је дистрибуиран сада, може бити примењен само на балансиране case-control студије (оне које имају исти број cases и controls)
4. Још један недостатак МДР је његова способност да прави предвиђања за независне скупове података када је величина најбољег модела релативно велика, а узорак је мали. Велика прецизност и мали узорак могу довести да неке од ћелија буду празне или попуњене само са једним податком. Ово није проблем за процену код грешке при класификацији и приликом евалуације конзистентности унакрсне валидације, али је проблем приликом процене грешке предикције. На пример, ако правимо једно посматрање за сваку ћелију у н-димензионалном простору, тада током унакрсне валидације, тада ће ово посматрање завршити или у сету података за тренирање или у скупу података који се користе за процену грешку класификације или у скупу тест података који се користе за процену грешке предвиђања, али не у оба. Ако овај податак заврши у другом скупу, тада неће бити изгенерисан модел да се направи предвиђање (ћелија ће бити празна). Овај проблем ограничава број посматрања за које предвиђања могу да се израчунају у скупу тест података који се користе за процену грешке предвиђања и такође утиче на укупну грешку при предвиђању.

5.6 Апликације

МДР је углавном коришћен за проналажење интеракције између гена или епистаза у студијама генетике људских болести, какве су аутизам, рак дојке, кардиоваскуларне болести, хипертензија, рак простате, шизофренија, рак бешике, и дијабетес типа 2. Међутим, МДР се може применити и на друга подручја (нпр. економија, метерологија, инжињерство...) где интеракција између дискретних атрибута може бити значајна за предвиђање бинарног резултата.

Емпиријске студије са симулираним и стварним подацима доказују је МДР довољно снажан алат за проналажење интеракција између гена. Због овога, МДР је идеалан за епидемиолошке студије које се баве проналажењем и карактеризацијом интеракција између гена.

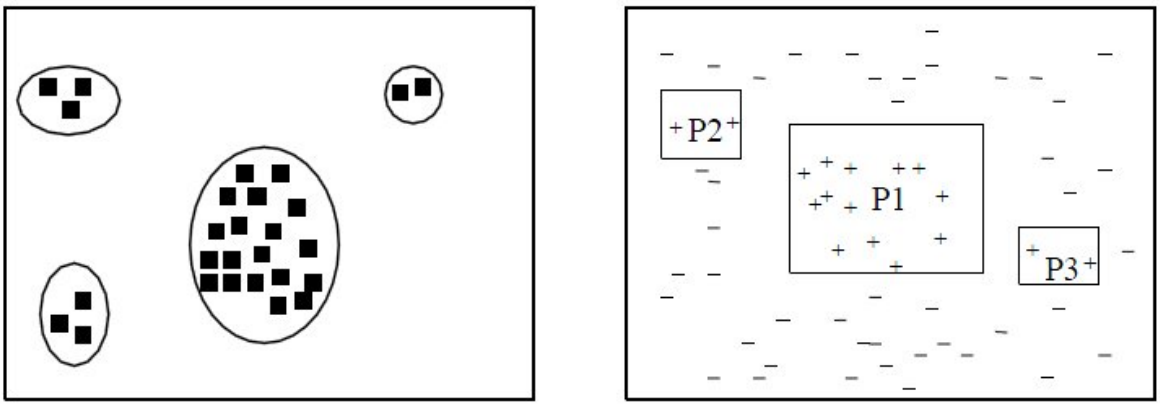
6 Небалансирани скупови података

6.1 Опис проблема

Многи системи за учење углавном претпостављају да су скупови који се користе за тренинг - уравнотежени. Међутим то није увек случај у реалном свету. Једна класа може бити репрезентована великим бројем примерака, док је друга репрезентована много мањим бројем. Овај проблем је познат као проблем небалансираних класа и углавном је препрека за индукцију добрих класификатора од стране алгоритама за машинско учење (енг. Machine Learning). Углавном се проблем небалансираних скупова података јавља када једна класа представља један концепт док друга класа представља пандан тог концепта, тако да примерци из класе дупликата тешко могу да надјачају примерке из позитивне класе.

Небалансиран скуп података се појављује када је класа од интереса јако мало заступљена у односу на друге класе. Овакви скупови података се појављују у многим апликацијама као што су медицинска дијагноза, детекција уљеза или провале, категоризација текста, менаџмент ризика, повраћај информација и филтрирајући задаци. Однос мањинске и већинске класе је у размери 1 према 100, 1 према 1000, 1 до 10 000 или више. Многи стандардни класификатори покушавају да увећају тачност и не улазе у дистрибуцију класа из скупа података за тренинг, који су често преплављени великим класама, и игноришу мањинске класе. РАЗЛОГ ЗАШТО ОВО ТРЕБА ДА БУДЕ ЈАСНО: предвиђање великих класа у небалансираним скуповима података може имати допринос више од 99%!

Учење помоћу класификатора из небалансираних или искривљених (енг. skewed) скупова података је важна тема која се често јавља у пракси приликом класификације проблема. У таквим проблемима све инстанце су типа једне класе, али неке од њих су типа друге класе која је углавном значајнија. Очигледно је да, традиционални класификатори који траже прецизне перформансе, више него опсег случајева, нису у могућности да се суоче са небалансираним скуповима података, јер покушавају да класификују све податке у једну велику класу, која је углавном мање битна класа.



Слика 4. Слабо заступљене класе се често класификују као шум

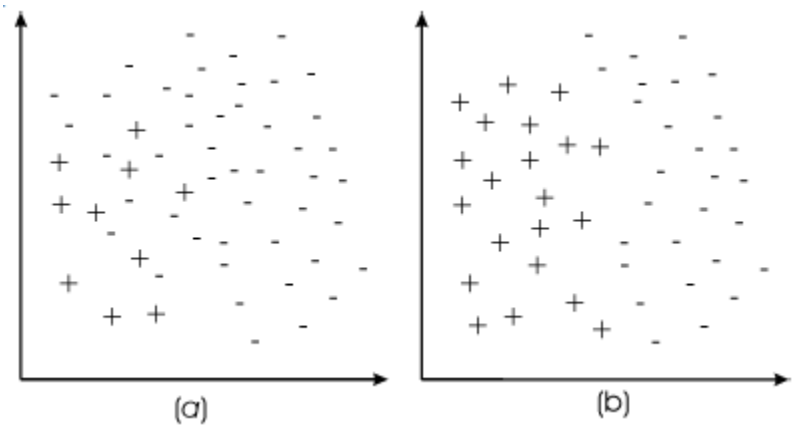
Висок степен дисбаланса се јавља у стварним доменима где систем који доноси одлуке треба да детектује ретку али веома битну класу. На проблем дисбаланса протеклих година ставља се велики акценат. Небалансирани скупови података постоје у многим доменима стварног живота, као што су уочавање непоуздане муштерије у телекомуникацији, детектовање нафтних мрља на сателитским сликама, учење изговора речи, класификација текстова, детектовање лажних телефонских позива, проналажење информација и филтрирање задатака итд.

Учење из небалансираних скупова података се проналази веома често у функционалним геномским апликацијама. У научним апликацијама је уобичајено да постоји само један позитивни примерак над хиљаду негативних инстанци. Нажалост, традиционални систем за машинско учење третирају овај примерак као безначајан. Стандардни поступак за превазилажење ове потешкоће је балансирање скупа података за вежбу тако што их resampling-је. Ипак ово резултира лажним позивитним предвиђањима.

Учење из небалансираног скупа података је релативно нов изазов за данашње data mining апликације. Од веб апликација до категоризације текста за биомедицинске анализе овај изазов се манифестује у две одређене форме: интереси мањине (енг. minority interests) и ретке инстанце (енг. rare instances). Интереси мањине се јављају у доменима где су ретки објекти (узорци мањинских класа) значајнији, и ту је циљ алгоритма за машинско учење да идентификује те мањинске класе што је тачније могуће. На пример, код финансијског инжењерства, важно је препознати преваре са кредитним картицама при великом броју трансакција. Са друге стране, ретке инстанце, односе се на ситуације у којима подаци представљају посебан ограничени догађај у поређењу са другим дистрибуцијама, као што је код проналажења нафтних мрља на сателитским снимцима. Многи проблеми небалансираног учења су настали комбинацијом та два фактора. На пример, у биомедицинским анализама података, узорци података за различите врсте рака су углавном веома ограничени у поређењу са нормалним не канцерогеним случајевима, тако да је однос мањинских класа са већинским веома значајан (1 према 1000 и више). Са друге стране неопходно је предвидети присуство рака и даље класификовати различите типове рака, што је прецизније могуће, због што ранијег и што бољег лечења.

Веза између величине скупа тренинг података и перформансе класификације над небалансираним скуповима података показује да у малим небалансираним скуповима података постоји мањинска класа која је лоше представљена због доста малог броја примерака који нису довољни за учење, поготово када велики степен преклапајућих класа и када је класа даље подељена на мање подкластере. За веће скупове података ефекат ових компликованих фактора је смањен јер је мањинска класа представљена већим бројем примерака.

Учење из небалансираних скупова података је често тежак задатак. У циљу да се боље разуме овај проблем замислите ситуацију илустровану на слици 5. На слици 5а постоји велика небалансираност између мањинских класа(-) и већинских класа(+) и скупови података показују некакав степен преклапања. Много боља ситуација за учење је показана на слици 5b где су класе избалансиране добро дефинисаним кластерима.

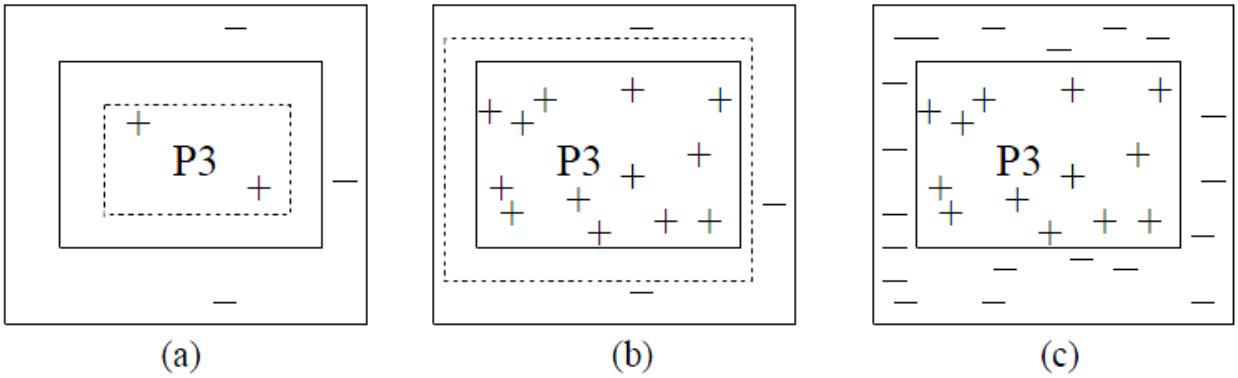


Слика 5

Уопштено говорећи, учење из небалансираних скупова података се дешава кад год неки типови дистрибуције података доминирају над инстанцама других дистрибуција података.

Слабо заступљени случајеви код великих скупова података су често од посебног интереса за истраживаче. Ово је посебно тачно у контексту data mining-a, где се жели пронаћи неки скривени образац који је можда сакривен међу великом количином података. Примери за тражење образаца: учење изговора речи, проналажење нафтних мрља на радарским снимцима, предвиђање кварова телекомуникационе опреме и проналажење зависности међу често куповним артиклима из супермаркета. Слабо заступљени случајеви захтевају посебну пажњу зато што они представљају значајне проблеме за data mining алгоритме.

Слабо заступљени случајеви представљају велики проблем за data mining алгоритме због много разлога. Основи и најочигледнији проблем је проблем недостатка узорака – узорци покривају само неколико примера за тренирање. Недостатак узорака чини откривање ових узорака тешким, чак и кад је неки примерак ове класе откривен он бива класификован погрешно или бива одбачен као шум. Да би размотрили мало боље овај случај погледајмо слику 6. Слика 6a приказује регион *P3*. Слика 6b приказује шта се дешава када се тренинг сет састоји само од једне врсте података, док слика 6c приказује додавање узорака мањинске класе. Границе које се узимају у обзир код учења су приказане испрекиданим линијама. Граница које се користи за учење на слици 6a је толико далеко од стварне границе региона, тако да искључује доста примерака. Укључивање додатних позитивних примера на слици 6b представља проблем апсолутне реткости и изазива да примерци који и не припадају региону *P3* бивају погрешно класификовани. На слици 6c, која укључује додатне примерке, како позитивне, тако и негативне, исправља претходни проблем (граница која се користи за учење скоро је иста као и стварна граница, стога и није приказана). Слике 6b и 6c показују да додатни узорци могу да побољшају проналажење решења код овог проблема (апсолутне реткости). Наравно, у пракси није увек могуће да се добију додатни узорци за тренирање.



Слика 6

Још један проблем који се јавља код учења помоћу малог броја узорака најбоље је описан цитатом – „Тражити иглу у пласту сена“. Тежина не лежи у томе што је игла мала, или што постоји само једна игла, већ зато што се она налази негде измешана са великом количином суве траве. Исто тако, при data mining-у, слабо заступљени узорци могу бити окружени великим бројем узорака неке друге класе. Овај проблем посебно долази до изражаја када су data mining алгоритми засновани на хеуристикама са похлепним (енг. greedy) претрагама које узимају у обзир један по један узорак, јер ретки узорци могу зависити од много услова и било које појединачно изоловано стање не може пружити добре смернице.

6.2 Приступи за решавање проблема

Методологије за решавање проблема небалансираних скупова података се могу поделити у следећих пет категорија:

i. Sampling strategies. Овај метод има за циљ да развије разноразне oversampling и/или undersampling технике да би компензовао дистрибуцију дисбаланса у оригиналним скуповима података
ii. Синтетичка генерација података (енг. Synthetic data generation). Овај приступ има за циљ да превазиђе дисбаланс у оригиналним скуповима података тако што генерише вештачке узорке
iii. Cost-sensitive learning. Уместо стварања уравнотежених дистрибуција података користећи sampling метод или метод генерисања синтетичких података, овај метод има другачији приступ овом проблему: користи матрицу цена (енг. cost-matrix) за различите типове грешака да би осигурао учење из небалансираних скупова података. Cost-sensitive learning не модификује директно дистрибуцију небалансираних података већ означава овај проблем користећи различите матрице трошкова да опише све цене погрешке класификације било којег изабраног узорка.
iv. Активно учење. Технике активног учења конвенционално се користе да реше проблеме коју су повезани са неозначеним подацима који се користе за тренирање. Уместо претраживања целог простора за тренинг овај метод ефектно одабира информативне инстанце из случајног сета популације, и тако у великој мери смањује цену израчунавања када се суочава са великим небалансираним скуповима података.
v. Kernel метода се такође користи за проучавање проблема небалансираних скупова података.

6.3 Технике семпловања

Семпловање (енг. sampling) је вероватно најдиректнији приступ за руковање дисбаланса између класа. Оно мења скуп података за тренинг или повећањем инстанци из мањинских класа (over-sampling) или смањивањем инстанци из већинских класа (under-sampling) или комбинацијом оба приступа. Као резултат дистрибуција класа из скупа података за тренинг се мења, а резултат је уравнотежени скуп података. Иако је генерална имплементација углавном једноставна постоје неки недостаци. Као прво under-sampling може одбацити неке корисне податке што може довести до under-fitting-а података већинске класе. С друге стране, over-sampling може довести до overfitting података мањинске класе, јер већина over-sampling метода генерише идентичне копије постојећих инстанци. Over-sampling повећава величину скупа података за тренинг што ће повећати време потребно да се научи класификатор.

6.3.1 Undersampling

Случајни under-sampling је нехеуристички метод који има за циљ створи баланс у дистрибуцији класа кроз случајну елиминацију неких примерака већинске класе. У позадини свега тога је покушај да се створи баланс у скупу података и да се превазиђе преосетљивност алгоритма за машинског учење (енг. machine learning algorithm). Велики недостатак случајног under-sampling је што овај метод може одбацити потенцијално корисне податке који могу бити важни за индукциони процес. Још један проблем је то што помоћу овог приступа циљ машинског учења је да класификатор процени вероватноћу дистрибуције одређене популације. Пошто је та дистрибуција непозната ми покушавамо да проценимо дистрибуцију популације користећи неки узорак дистрибуције. Статистика нам говори да докле год је узорак узет случајно, овај узорак може бити коришћен да се процени дистрибуција популације из које је узет. Једном када применимо undersampling над већинском класом узорака, целокупан узорак се више не може сматрати случајним.

Овај метод уклања примерке већинских класа да би скуп података постао уравнотежен. Овај метод је одговарајући за апликације великих размера где је број примерака већинских класа јако велики и где смањење узорака за тренинг смањује време и простор за тренинг. Недостатак овог метода је што одбацује потенцијално корисне информације које могу бити корисне за класификаторе.

Овај метод је подељен на случајни и информативни. Случајни Undersampling је једноставан и случајно елиминише примерке већинских класа док скупови података не постану уравнотежени. Информативна Undersampling метода бира само одабране примерке већинске класне, који су базирани на изабраном критеријуму селекције, тако да учини скуп података уравнотеженим. Информативни Undersampling може бити пасивни и активни. Пасивна метода селекције се обично предлаже као предпроцесна техника за селекцију информативног узорка за класификатор. При активној селективној методи информативни узорци се испитују током конструкције класификатора.

6.3.2 Oversampling

Случајни over-sampling је нехеуристички метод која има за циљ да избалансира дистрибуцију класа кроз случајно копирање примерака мањинске класе. Случајни over-

sampling може повећати вероватноћу да се догоди overfitting, јер прави идентичне копије примерака мањинске класе. На тај начин класификатор може изгенерисати правила која су на први поглед тачна, али заправо она покривају само један одговарајући пример. Oversampling може увести додатни рачунарски задатак ако је скуп података прилично велики, али небалансиран.

Oversampling је приступ који балансира скупове података тако што дуплира примерке мањинских класа. Још се назива и upsampling. Предност овог метода је што не постоји губитак података као у undersampling техници. Недостатак ове технике је да може довести до overfitting-а и може увести додатни трошак при израчунавању ако је скуп података велики, али небалансиран.

Као и Undersampling, oversampling је такође подељен на два дела: случајни и информативни Oversampling. Случајни Oversampling је метод који балансира дистрибуцију класа тако што случајно дуплира изабране примерке мањинске класе Информативни Oversampling је метод који синтетички генерише мањинске примерке класа на основу критеријума који су одређени пре тога.

6.4 Метрике за мерење перформанси

Метрика која се користи током data mining процеса, као и за процену резултата може такође да закомпликује ствар. Основна претпоставка ових метода је да се инстанце које се јављају ретко третирају се као шум. То још више дискриминише мањинску класу да како би се постигла што боља тачност код предвиђања. За високо дисбалансиране сетове података, класификатори направљени помоћу ових алгоритама би једноставно лоше предвиђали све време и постогли готово 100% тачност! Ово је бесмислено за апликације које се користе за интернет безбедност, као и код геномске функционалности где је циљ да се открију ове инстанце са одређеном дозом толеранције. Штавише, размотримо начин на који се формирају стабла одлуке. Већина стабала расте од врха надоле, где се тест услови у више наврата израчунавају, и изабира се најбољи. Метрика коришћена да се изабере најбољи тест генерално више воле тестове који дају балансирано стабло где је добијен висок проценат „чистоће“ за већину примерака, од тестова који постижу висок проценат „чистоће“ за мали подскуп података и мали проценат „чистоће“ за остале податке. Због тога мањинска класа ће често бити занемарена приликом генерисања стабла одлуке.

Многи истраживачки радови на тему небалансираних скупова података говоре да су перформансе постојећих класификатора углавном пристрасни већим класама. Разлози за лоше перформансе постојећих класификацијских алгоритама над небалансираним скуповима података су:

1. Њихов циљ је да смање укупне грешке приликом чега мањинске класе доприносе јако мало
2. Они претпостављају да постоји једнака дистрибуција података за све класе
3. Такође претпостављају да грешке које долазе из различитих класа имају исту цену

Перформансе традиционалних алгоритма за класификацију су процењене прецизношћу која је дефинисана као проценат примерака који нису тачно класификовани. Ово није одговарајуће када се ради са небалансираним скуповима података где мањинска класа има мање ефекта на тачност него већинска класа. Ту постоји потреба за другом метриком. Све

ове метрике користе матрицу (енг. confusion matrix). Формула ове матрице је дата испод. представља најпознатију евалуациону метрику. TP и TN означавају број позитивних и негативних примера који су тачно класификовани, док FN и FP означавају број позитивних и негативних примера који су погрешно класификовани. Разлика између ове матрице и матрице цена је та што прва садржи број TP, FN, FP, TN у ћелијама док друга матрица обезбеђује информације о погрешној класификацији цене званој FN, FP. Код ове матрице елементи на дијагонали су нуле.

		Hypothesis output: Y	Hypothesis output: N
True class	p	TP (True Positives)	FN (False Negatives)
	n	FP (False Positives)	TN (True Negatives)

Слика 7. Матрица за евалуацију

Предложене су две метрике за рад са дисбалансом класа. Да би се добила оптимална уравнотежена способност класификовања, сензитивитет и тачност, обично се користи класификација перформанси те две класе одвојено. Сензитивитет се још назива true positive rate или positive class accuracy, док се специфичност назива true negative rate или negative class accuracy. Предложена је G-Mean метрика која се заснива на ове две метрике, која је геометричко средство сензитивитета и тачности. Још су усвојени и прецизност (енг. precision) и опозив (енг. recall) као метрике. F-Measure се користи као комбинација ова два метода у јединствену метрику за добро моделовање.

Overall Accuracy (OA):

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

Precision:

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

Recall:

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

F_Measure:

$$F\_Measure = \frac{(1 + \beta^2) \cdot recall \cdot precision}{\beta^2 \cdot recall + precision}$$

G_mean:

$$G\_mean = \sqrt{PositiveAccuracy \times NegativeAccuracy} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

Слика 8. Опште прихваћене мере. β код F_Measure је коефицијент за подешавање релативног значаја прецизности наспрам одзива (најчешће се узима да је β = 1)

Друге метрике Receiver Operating Characteristics (ROC) curve и Area Under ROC (AUC) , се широко користи када се ради са небалансираним скуповима. Receiver Operating Characteristics (ROC) curve даје визуелну индицију да је класификатор супериорнији другом класификатору у односу на широк спектар операција. Area Under ROC curve (AUC) сумира перформансе класификатора у јединствену метрику. Што је већи простор испод боље су перформансе.

6.5 Закључак

Sampling технике су директан приступ руковања дисбаланса између класа, и могу бити коришћене у комбинацији са неким другим класификатором. under-sampling нуди решење за огромнан број инстанци у скупу података за тренинг чија величина треба да буде смањена. Коришћење sampling стопа дозвољава се фино подешавање дисбаланса класа и као резултат класификације се добијају бољи класификациони модели и перформансе.

Oversampling и Undersampling су ефектне методе које се користе приликом рада са небалансираним скуповима података. Али undersampling приступ функционише боље него oversampling метод над великим скуповима података. Oversampling је бољи за мале скупове података.

7 Алгоритми за балансирање

KEEL (Knowledge Extraction based on Evolutionary Learning) је open source (GPLv3) алат за оцењивање алгоритама за решавање Data Mining проблема као што су регресија, класификација, подела на кластере, проналажење образаца... Садржи велику колекцију класичних алгоритама за екстраховање знања, предпроцесирање, рачунарска интелигенција заснована на алгоритмима за учење, укључујући учење засновано на правилу еволуције (енг. evolutionary rule learning) користећи различите приступе (Питсбург, Мичиген и IRL, ...), и хибридне моделе какви су генетички фази (енг. fuzzy) системи, неуралне мреже итд. Овај широк дијапазон алгоритама и додатака нам омогућава да урадимо комплетну анализу било ког модела за учење, упоређујући га са било којим другим постојећим, укључујући и модуле за статистичко поређење.

Штавише, KEEL је развијен за две категорије корисника: истраживаче и студенте, од којих свака има другачији скуп потреба:

- *KEEL као алат за истраживање*. Најчешћа употреба овог алата за истраживања је код аутоматизације експеримената и при прегледу статистичких резултата. Експеримент је уобичајено мешавина алгорит(а)ма, статистичког прегледа и неке од техника вештачке интелигенције. Како садашњи стандарди у машинском учењу потражују тешка израчунавања, алат није направљен тако да поседује приказ напретка експеримента у реалном времену, већ након обављеног експеримента истраживач може погледати резултате. Алат омогућује истраживачима да употребе исту секвенцу за предпроцесирање, експеримент и анализу над много проблема, и да се фокусирају само на резултате тих експеримената.
- *KEEL као алат за учење*. Потребе студената су прилично различите од потреба истраживача. Уопштено говорећи, циљ више није да се врши упоредна оцена статистике. Такође нема потребе да се експеримент изводи небројено пута. Ако се алат користи на предавању, извршавање алгоритма мора да буде кратко и потребно је имати увид у рад алгоритма у реалном времену како би студенти научили да сами подешавају параметре за неки алгоритам/експеримент. У том смислу, алат за учење је у ствари поједностављена верзија алата за истраживање, где су само најважнији алгоритми доступни за рад. Извршавање експеримента се врши у реалном времену, и корисник има визуелни преглед напретка алгоритма. Такође на крају алгоритма студент може приступити резултатима експеримента и прегледати их.

7.1 Увод

Технолошки напредак омогућава биолозима да прикупе велику количину геномских података помоћу аутоматизованих ДНК секвенцера, микро низова који генеришу генетске изразе информација за цео организам. Ови подаци садрже важне информације које могу до проналажења третмана за смртоносне болести, као и да нам побољшају квалитет живота. Иако ове технике могу да нам послуже као вредни и корисни алати за анализу генома, резултати су далеко од идеалних. У великом броју геномских апликација, суочени смо са изазовом небалансираних скупова података, где можемо видети један позитиван

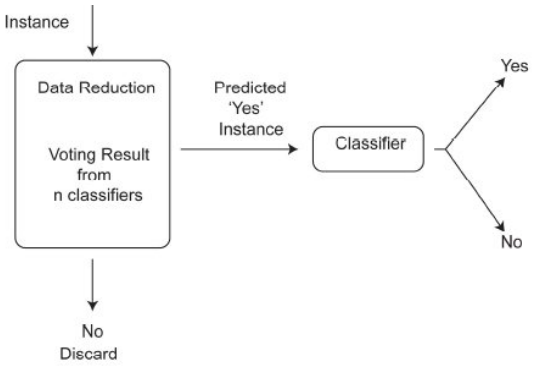
пример након стотина или хиљада негативних. А ова класа, која је у мањини, представља важнију класу, односно класу коју желимо да посматрамо. Такође у области рачунарске (интернет) безбедности, већина трагова у системским логовима представљају не злонамерне конекције и отуда је скуп тренинг података за изградњу система за аутоматску детекцију злонамерних упада веома небалансиран.

Овде ћемо представити три технике које се баве проблемом небалансираних скупова података, када је број података једне класе значајно превазилази број података друге, али битно значајније класе. Иако је коефицијенти дисбаланса у многим функционалним геномских апликације изузетно висок, срећом сложеност већинске класе тежи да буде умерено ниска. Ово омогућава да поделимо већинску класу на кластере. Прави се основни класификатор за сваки кластер да се издвоји већинска од мањинске класе. Још једна предност овог метода је да, за разлику од традиционалних undersampling метода, овај метод користи све податке инстанце из већинске класе тако да не постоји никакав губитак информација.

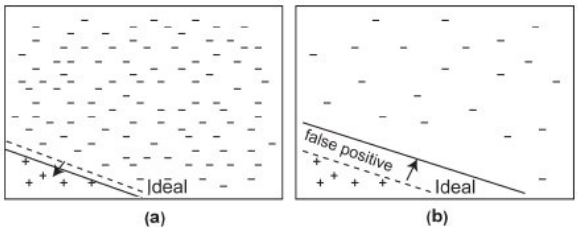
7.2 CPM-I Class Purity Maximization

Главна идеја овог алгоритма је била да се направи филтер за проналажење већинске класе без претераног губитка мањинске. Ово омогућава да се смањи дисбаланс, који чини учење више прилагодљиво. Како је мањинска класа оскудна, веома је важно да било који начин смањивања небалансираности неког скупа података не доведе до елиминације мањинске класе из оригиналног скупа података. Идеја да се ово постигне је да се нађе што већи број кластера који садрже само већинску класу или само мали проценат мањинске класе. У суштини, жеља је да се открију већинске класе које су много далеко од циљне границе (и на тај начин смањити количину небалансираности) тако да се концентришемо на разликовање много тежих граничних случајева. Кључ овог решења је да се пронађу кластери који су „чисти“ или готово „чисти“ од инстанци мањинске класе. Тако да је овај приступ развијен за максимизовање „чистих“ кластера.

Алгоритам бира пар инстанци из мањинске и већинске класе за центре. Остале инстанце се деле у две подгрупе узимајући у обзир раздаљину од ових центара, али тако да макар једна подгрупа остане што чистија. Поступак се понавља рекурзивно за сваки од ове две подгрупе док год више не можемо да формирамо два кластера, од којих барем један има већи проценат чистоће него његов родитељ. Тада се прави колекција узорака додавањем мањинских инстанци у сваки кластер који није чист, и конструише се стабло одлучивања за сваки узорак. Слика 9 илуструје овај процес. Када је потребно класификовати нову инстанцу, прво пролазимо кроз процес смањивања дисбаланса да би смо открили најбољи кластер у којем би ова инстанца припала. Ако нова инстанца припадне чистом кластеру, она се одбацује. Само оне инстанце које припадну кластерима који нису чисти, бивају даље обрађене помоћу стабла одлучивања. Ако добијемо да инстанца припада мањинској класи, она се филтрита помоћу финалног класификатора (који се конструише помоћу неуралних мрежа).



Слика 9. Уопштен приказ прцедуре за редукцију дисбаланса и шема за предикцију



Слика 10. Илустрација небалансираности међу подацима и приказ Undersampling метода: (a) небалансирани скуп података – граница за одлучивање је померена ка мањинској класи; (b) после Undersampling-а – граница је померена ка већинској класи

Овај алгоритам је упоређен са три водећа предиктора; перформансе поређења су извршене над њиховим веб-базираним (енг. web-based) програмима користећи се истим тестовима за фер поређење.

1. NNSplice1 са Berkeley Drosophila Genome Project (BDGP) - NNSplice је подпроцес система за детекцију гена, Genie. Две независне неуралне мреже се користе да предвиде донора и примаоца, засноване на динуклеотидним фреквенцијама
2. GeneSplicer2 са Institute for Genomic Research (TIGR) - GeneSplicer је метод који користи стабло одлучивања базирајући се на Maximal Dependence Decomposition и побољшаном Markov-ом моделу
3. SpliceView3 са Institute of Advanced Biomedical Technologies (ITBA) - SpliceView узима у обзир сигнале из консензуса секвенци из граничних региона

Иако овај метод и NNSplice деле исту грађу засновану на неуралним мрежама, овај приступ је успео да смањи количину дисбаланса. Даље, овај метод користи сасвим другачији начин за добијање резултата од сва три горепоменута метода. И као резултат оваквог начина изградње модела, добијају се много прецизнији резултати предикције. Конструкција овог модела била је вођена следећим циљевима: при конструкцији филтера користе се све инстанце из већинске класе, док год се не добије потврда да резултујући филтер има велику стопу опозивости – што нам даје бољу предикцију; и друго, резултујући филтер мора да елиминише што већи број инстанци већинске класе без да изгуби било коју инстанцу мањинске класе.

7.3 ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning

ADASYN се базира на идеји да се адаптивно генеришу узорци мањинске класе, на основу њихове дистрибуције: више синтетичких података се генерише за мањинске класе за које се испоставља да су теже за учење. Као резултат овај приступ побољшава учење у вези са дистрибуцијом података на два начина: (1) смањује пристрасност коју уводи дисбаланс између класа, и (2) адаптивно помера границу одлучивања за класификацију ка тежим случајевима за учење.

Кључна идеја ADASYN алгоритма је да се користи густина дистрибуираности као критеријум да се аутоматски одлучи број синтетичких узорака који су потребни да се генеришу за сваку мањинску класу. Физички, то је мера дистрибуираности тежина за различите мањинске класе према њиховом нивоу тежине при учењу. Резултујући скуп података нако примене ADASYN алгоритма неће само обезбедити балансираност података унутар скупа података, већ ће такође форсирати алгоритам за учење да се фокусира на оне примерке који су тежи за учење.

Циљ је да се обезбеде различите тежине за различите мањинске класе да би надокнадили искривљеност дистрибуције података. Овај алгоритам адаптивно исправља ову искривљеност, базирајући се на карактеристикама дистрибуције података. Отуда није потребна никаква хипотеза за евалуацију за генерисање синтетичких података код овог алгоритма.

Базиран на оригиналној дистрибуцији података, ADASYN може адаптивно да генерише синтетичке податке за мањинске класе да редукује дисбаланс. Штавише, ADASYN такође може аутономно да помера границу одлучивања за класификацију да би се више фокусирао на примерима који се теже уче, и на тај начин побољшава перформансе при учењу. Ово се постиже динамичким регулисањем тежина и адаптивним процедурама за учење на основу дистрибуираности података.

7.4 RUS-I Random under-sampling

Многи алгоритми користе Еуклидско растојање као меру за k-NN. Међутим ова функција за мерење дистанце некад није згодна када домен предтавља квалитативне атрибуте. За такве случајеве, функција за мерење дистанце је обично једноставна: вредност нула (ако оба узорка имају исту вредност атрибута) и вредност 1 (ако је ова вредност различита). Овај алгоритам користи другу функцију за мерење дистанце - Heterogeneous Value Difference Metric (HVDM). Ова функција користи Еуклидско растојање за квантитативне атрибуте и VDM растојање за квалитативне атрибуте. VDM метрика обезбеђује много пригоднију функцију за мерење дистанце за квалитативне атрибуте, ако је упоредимо са преклапајућом метриком, јер VDM метрика узима у обзир сличност код класификације за сваку могућу вредност квалитативног атрибута да би израчунала растојање између ових вредности. Још једно побољшање основног k-NN алгоритма је да се одмери допринос сваког од к суседа према њиховој удаљености упита, дајући већу тежину ближим суседима.

Како методи за балансирање праве превише израчунавања, овде је направљена једна индексна структура названа M-tree да би убрзала извршавање k-NN упита. M-tree узима у обзир само релативна растојања од тренутног узорка, више него њихове апсолутне позиције у мултидимензионалном простору да организује и подели метрички простор. У метричком простору, близина узорка је дефинисана само функцијом растојања која задовољава позитивност, симетрију и постулате троугла неједнакости.

8 Симулација података

8.1 Шта је симулација и зашто је важна?

Студије распострањености удруживања гена су постале реалност при студијама генетике комплексних болести. Ова технологија обезбеђује богате колекције генетских информација о пацијентима, уз помоћ којих се генетичари надају да изуче „нову биологију“ и пронађу важне генетске и факторе из окружења које утичу на процес стварања и/или погоршавања болести. Како стратегије за анализу ових података нису држале корак са лабораторијским методама, мало је вероватно да ће оне (стратегије) одмах довести до побољшања разумевања генетског утицаја на људске болести, као и реакција на лекове. Тренутно не постоји ни један аналитички метод који нам омогућава да извучемо све информације за студије распрострањености удруживања гена. Због тога се нови методи предлажу и развијају. За успех ових метода веома је важно да имају способност да симулирају скупове података који се састоје од полиморфизама генома са реалном повезаношћу ка небалансираним обрасцима. Са овим скуповима података, можемо уградити нове гентске моделе болести чиме можемо оценити способност ових нових модела да пронађу симулиране ефекте.

Један начин да на бољи начин изградимо аналитички протокол је да имамо скуп података са унапред познатим одговором, али ово је немогуће ако користимо стварне податке. Када се стварни подаци користе да се тестирају нови методи, и добијени резултат је значајан, немогуће је да се открије да ли је тај резултат стваран или не. Слично томе, ако не пронађемо ни један значајан резултат, не можемо рећи да он не постоји стварно, као ни да је метод лош и недовољно снажан. Због тога, за успех ових метода веома је важно да имају способност да симулирају скупове података који се састоје од полиморфизама генома са што већом техничком изводљивошћу. Када имамо симулиране податке, то нам омогућава да проценимо да ли методологија може да пронађе добро познате образце, и да ли су симулације добро осмишљене за детекцију генетичких модела болести, на нивоу генетичке структуре.

Данас симулација података је једана од најбољих парадигми за моделовање понашања комплексних система, иако има неких недостатака. Симулациони модел је само груба апроксимација стварног ситема за студије случаја; свака произведена апроксимација неће покрити све детаље које које можемо тражити у реалном систему. Раскорак између модела и реалности је добро познат проблем у рачунарској индустрији и математици, али ситуација је далеко од тога да је заиста очајна. Јаз може бити намерно произведен, зато што домен интереса може бити само мали део целокупне стварности. Штавише, способност да се створе вештачки светови чији се односи могу мењати дозвољава нам да истражимо све могућности стварног система. Чињеница да можемо симулирати систем под неприродним околностима, може нам помоћи код незамисливо флексибилних сценарија.

Симулација је у ствари имитација операција из стварног света или система, укључујући и фактор време. Да би се спровела „уметност симулације“ потребно је да модел већ постоји,

односно да је развијен; овај модел представља кључну карактеристику понашања изабраног физичког или апстрактног система или процеса. Модел представља сам систем, а симулација представља операције система током неког периода времена.

Рачунарска симулација је покушај да се моделује стварни свет или хипотетичка ситуација на рачунару, тако да се може направити студија да би се разумело како и шта тај систем ради. Мењајући промењиве током симулације, могу се направити предвиђања о исходу понашања система. То је алат да се практично истражи понашање система за који се прави студија случаја.

Рачунарска симулација је такође постала користан део моделовања многих природних система у физици, хемији и биологији, као и у економији, рачунарству... Добар пример је симулација протока и комуникације путем светске мреже (www). У овим симулацијама, модел мења понашање током симулације, према сету иницијалних параметара који представљају стање окружења.

Симулација такође ослобађа извршиоца анализе од великог броја понављања при замењивању бројева у формулама и табелама и омогућује му да се концентрише само на резултате. Још једна предност је увид у перформансе система, како код процеса моделовања, тако и код искуства добијено кроз симулационе експерименте. Опет, да поновимо, рачунари нам могу помоћи: може бити веома тешко или скупо да се сакупе подаци за мерење, тако да подршка рачунара при анализи може квантификовати значај сакупљања података и аналитичких процедура.

Коришћење моделовања и симулирања у инжењерству је веома добро познато. Симулациона технологија припада скупу инжењерских алата у апликација свих домена. Моделовање и симулација су већ помогли да се смање трошкови и повећа квалитет производа и система, а научене лекције су документоване и архивиране.

Моделовање и симулација представљају посебну дисциплину за себе. Због тога што има широк домен примене може се погрешно претпоставити да је моделовање и симулација апликација сама за себе. Али то није случај и потребно је да се она препозна од стране стручњака који хоће да користе овај метод. Да би осигурали да резултати симулатора буду примењиви у стварном свету инжењери морају разумети предпоставке, концепте и имплементациона ограничења овог поља.

Међу разлозима за стално повећање интересовања за симулативним апликацијама су:

- Коришћење симулација је, по правилу, јефтиније и безбедније од спровођења експеримената са прототипом реалних ствари. Један од највећих рачунара на свету је дизајниран да симулира детонацију нуклеарних направа и њихових ефеката са циљем да подржи бољу припремљеност у случају нуклеарне експлозије. Слични напори се користе да се симулирају урагани и друге природне катастрофе.
- Симулације су често много реалније од традиционалних експеримената, јер дозвољавају слободну конфигурацију параметара из средине која се налази у оперативном апликационом пољу крајњег производа. Примери су подржавање операције „Америчких Фока“ (енг. US Navy) на великим дубинама или симулирање површина суседних планета за које НАСА планира подухвате.

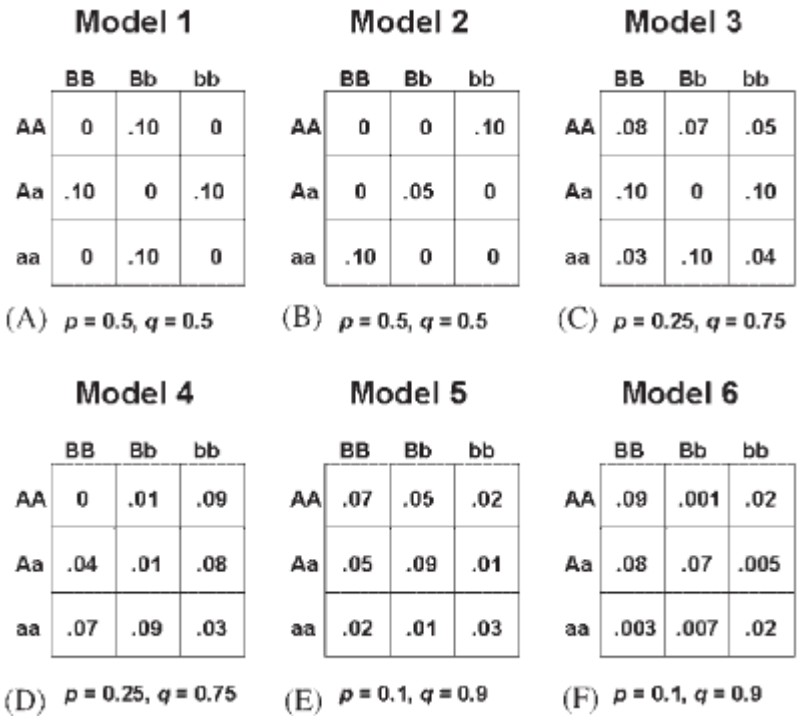
- Симулације се често могу спроводити брже од реалног времена. То дозвољава њихово коришћење за анализе различитих алтернатива, нарочито када се потребни подаци за симулацију могу лако бити добијени из оперативних података.
- Симулације дозвољавају стварање кохерентног синтетичког окружења које дозвољава интеграцију симулираних система, у раним аналитичким фазама, уз помоћ виртуелних система са првим прототипским компонентама, у виртуалну тест средину за коначни систем. Ако се користи правилно, средина може мигрирати из домена развоја и тестова у домен тренинга и едукације у цикличним фазама система које следе (укључујући и опцију за тренирање и оптимизацију виртуелних близанаца из реалног света под реалним ограничењима чак и пре него што се направе прве компоненте).

8.2 Како ради симулација?

Да би се проценила моћ МДР-а за детектовање генских интеракција, симулирали смо case-control податке користећи 6 различитих епистазних модела са 2 места у којима су функционална места single-nucleotide Polymorphisms (SNPs). Први модел је базиран на нелинеарној XOR функцији која генерише интерактивне ефекте у којима је висок ризик од болести зависан од наслеђивања генотипа из једног места (Aa) или генотипа из другог места (Bb), али не и оба. Комбинације генотипа високог ризика су AaBB, Aabb, AABb и aaBb. У другом моделу болест високог ризика је зависна од наслеђивања тачно две високо рискантне алеле са два различита места. За овај модел комбинације генотипа високог ризика су AAbb, AaBb и aaBB (Слика 11B). Остала четири модела су била генерисана коришћењем Муровог метода откривања епистаза, коришћењем алелских фреквенција од p=0.25 и q=0.75 за моделе 3 (Слика 11C) и 4 (Слика 11D), и алелских фреквенција од p=0.1 и q=0.9 за моделе 5 (Слика 11E) и 6 (Слика 11F). Сви ови модели су изабрани јер показују ефекте интеракције у одсуству главних ефеката када су генотипи били генерисани по узору на Hardy-Weinberg пропорције. Интеракција без главних ефеката је пожељна јер даје велики степен комплексности том изазову идентификовања генетских интеракција. Да су главни ефекти били присутни било би тешко проценити да ли су одређена места била детектована због главних ефеката или због интеракције, или због оба.

Сваки скуп података се састојао од 200 случајева и 200 контрола, сваки са 10 СНП-ова, од кога су 2 била функционална. Сваки СНП је имао две алела које су са заједничким алаелама имале фреквенцију од 0.5, 0.75, или 0.9 као што је већ описано у 6 различитих модела. Генотипи су били генерисани према пропорцијама Hardy-Weinberg-а.

Такође постоји могућност прављења и небалансираних скупова података на тај начин што ће унутар функције generate() која је задужена за прављење случајних података, променити однос case-controls. У овом експерименту смо користили следеће врсте дисбаланса: 10 – 90%, 20 – 80% и 30 – 70%.



Слика 11

```
public void generate() throws FileNotFoundException{
          PrintStream fout=new PrintStream("Model"+model+".csv");
          String  head= "X1" + "," + "X2" + "," + "X3" + "," + "X4" +
"," + "X5" + "," + "X6" + "," + "X7" + "," + "X8" + "," + "X9" + "," +
"X10," + "C";
          fout.println(head);
          int cases=0;
          int controls = 0;

          while(controls<200){
               String s="";
               for(int j=0; j<10; j++){
                    s += generateOneNp(mafs[j])+',';
               }
               int thisCase = getStatus(s, penetrances);
               if(thisCase ==1) cases++;
               else controls++;
               s+=thisCase;
               System.out.println(s);
               fout.println(s);
          }

          while(cases<100){
               String s="";
               for(int j=0; j<10; j++){
                    s += generateOneNp(mafs[j])+',';
               }
```

```
                int thisCase = getStatus(s, penetrances);
                if(thisCase ==1){
                      cases++;
                      s+=thisCase;
                      System.out.println(s);
                      fout.println(s);
                }
          }
          controls = 0;
          while(controls<100){
                String s="";
                for(int j=0; j<10; j++){
                      s += generateOneNp(mafs[j])+',';
                }
                int thisCase = 0;
                if(thisCase ==0){
                      controls++;
                      s+=1;
                      System.out.println(s);
                      fout.println(s);
                }
          }
    }
```

Након извршене симулације података небалансираних скупова, прешло се на балансирање ових скупова података помоћу алгоритама описаних у седмом поглављу. Сваки од резултата, како оригиналних (небалансираних), тако и новодобијених (балансираних) пропуштен је кроз МДР, и на крају је направљен упоредни приказ резултата и извршена је њихова анализа.

9 Резултати

У наставку је дат упоредни приказ резултата. За сваки модел је приказано следеће: у свакој од колона је дата вредност тачности тренинга (енг. Training Accuracy) за сваки од скупова података (најпре небалансирани, а затим балансирани уз назнаку који је алгоритам за балансирање примењен). У првој врсти налазе се проценти дисбаланса података.

9.1 Модел 1

	10-90	20-80	30-70
Без балансирања	0,6139	0,6500	0,6571
Са CPM-I	0,5566	0,6943	0,6100
Са RUS-I	0,7812	0,6953	0,6927
Са ADASYN	0,6701	0,6504	0,6629

9.2 Модел 2

	10-90	20-80	30-70
Без балансирања	0,6375	0,6328	0,6226
Са CPM-I	0,6401	0,6533	0,5622
Са RUS-I	0,6875	0,6562	0,6719
Са ADASYN	0,6875	0,6387	0,6094

9.3 Модел 3

	10-90	20-80	30-70
Без балансирања	0,5279	0,5594	0,5369
Са CPM-I	0,5803	0,6020	0,6657
Са RUS-I	0,6548	0,5859	0,6771
Са ADASYN	0,6823	0,6250	0,6027

9.4 Модел 4

	10-90	20-80	30-70
Без балансирања	0,5993	0,6109	0,5804
Са CPM-I	0,6932	0,7137	0,6195
Са RUS-I	0,6447	0,6797	0,6406
Са ADASYN	0,6198	0,6543	0,5915

9.5 Модел 5

	10-90	20-80	30-70
Без балансирања	0,6181	0,5891	0,4708
Са CPM-I	0,5989	0,6186	0,6133
Са RUS-I	0,7500	0,6712	0,5627
Са ADASYN	0,6875	0,6230	0,5893

9.6 Модел 6

	10-90	20-80	30-70
Без балансирања	0,5958	0,5375	0,5440
Са CPM-I	0,5385	0,6035	0,5016
Са RUS-I	0,6094	0,5938	0,6458
Са ADASYN	0,6927	0,6152	0,6406

10 Закључак

Технолошки напредак омогућава биолозима да прикупе велику количину геномских података помоћу аутоматизованих ДНК секвенцера, микро низова који генеришу генетске изразе информација за цео организам. Ови подаци садрже важне информације које могу довести до проналажења третмана за смртоносне болести, као и да нам побољшају квалитет живота. Иако ове технике могу да нам послуже као вредни и корисни алати за анализу генома, резултати су далеко од идеалних.

Многи системи за учење углавном претпостављају да су скупови података који се користе за тренинг - уравнотежени. Међутим то није увек случај у реалном свету. Висок степен дисбаланса се јавља у стварним доменима где систем који доноси одлуке треба да детектује ретку, али веома битну класу. На проблем дисбаланса протеклих година ставља се велики акценат. Учење из небалансираних скупова података се проналази веома често у функционалним геномским апликацијама. У научним апликацијама је уобичајено да постоји само један позитивни примерак над хиљаду негативних инстанци.

Циљ овог рада је био испитати понашање МДР алгоритма у присуству небалансираних скупова података. МДР је data mining стратегија за детекцију и карактеризацију комбинација атрибута или независних променљивих (нпр. СНП, пушење, пол...) који интерагују, а које утичу на неку зависну променљиву. МДР је направљен специјално за идентификацију интеракција између дискретних променљивих које утичу на бинарни исход и сматра се непараметарском алтернативом за традиционалне методе.

Хипотеза коју проверавамо у овом раду: метод ће имати боље резултате уколико се користе уравнотежени (балансирани) скупови података, него небалансирани. Резултати које смо добили, а који су приказани у претходном поглављу доказују да је наша почетна хипотеза оправдана, односно да МДР метод има боље резултате уколико користимо уравнотежене (балансиране) скупове података. Стога употреба алгоритама за балансирање небалансираних скупова података може битно утицати на побољшавање алгоритама за учење.

11 Литература

[1] http://www.multifactordimensionalityreduction.org/
[2] http://www.epistasis.org/
[3] Epistasis Blog
[4] http://en.wikipedia.org/wiki/Multifactor_dimensionality_reduction
[5] http://keel.es/
[6] Nacional Cancer Institute
[7] Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC: **A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility.** *J Theor Biol* (2006)
[8] J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera: **KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework**. *Journal of Multiple-Valued Logic and Soft Computing* (2011)
[9] J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera: **KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems**. *Soft Computing* (2009)
[10] Ritchie MD et al. **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer**. *Am J Hum Genet* (2001)
[11] Ritchie MD, Hahn LW, Moore JH: **Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity**. *Genet Epidemiol*. (2003)