Универзитет у Београду
Математички факултет

Лазар М. Васовић

Скривени Марковљеви модели (HMM) у биоинформатици

мастер рад

Београд, 2021.

Ментор:

др Јована Ковачевић, доцент
Универзитет у Београду, Математички факултет

Чланови комисије:

... ..., ...
..., ...

... ..., ...
..., ...

Датум одбране: септембар 2021.

Наслов мастер рада: Скривени Марковљеви модели (*HMM*) у биоинформатици

Резиме: ...

Кључне речи: биоинформатика, скривени Марковљеви модели (*HMM*)

Садржај

Глава 1

Увод

Биоинформатика је интердисциплинарна област која се бави применом рачунарских технологија у области биологије и сродних наука, са нагласком на разумевању биолошких података. Кључна особина јој је управо поменута мултидисциплинарност, која се може представити дијаграмом са слике 1.1.

Слика 1.1: Венов дијаграм интердисциплинарности[6]

Овако представљена, биоинформатика је заправо спој статистике, рачу-

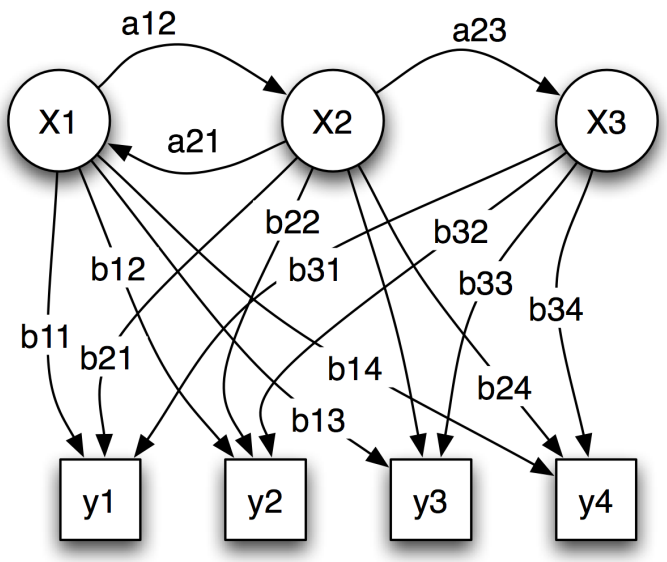
нарства и биологије – сва три истовремено – по чему надилази појединачне спојеве: биостатистику, науку о подацима и рачунарску биологију. Конкретно, статистички (математички) апаратат служи за рад са подацима, рачунарске технологије тај апарат чине употребљивијим, док биологија даје потребно доменско знање (разумевање) за рад са биолошким и сродним подацима. Иако се може рећи да је биоинформатика, у савременом смислу представљеном приказаним дијаграмом, релативно млада наука, брзо је постала популарна и многи су јој посветили пажњу или се њоме баве[13, 10, 1].

Међу познатим личностима из овога домена издвајају се научници Филип Компо (*Phillip Compeau*) и Павел Певзнер (*Pavel Pevzner*), аутори књиге *Bioinformatics Algorithms: An Active Learning Approach*. Прво издање књиге изашло је 2014. године, а друго већ наредне, у два тома. Актуелно, треће издање, издато је 2018. године, у једном тому. Захваљујући динамичном и активном приступу биолошким проблемима и њиховим информатичким решењима, као и многим додатним материјалима за учење, књига се користи као уџбеник на више од сто светских факултета[2]. Међу њима је и Математички факултет Универзитета у Београду, односно на њему доступни мастер курс Увод у биоинформатику, а делови књиге користе се и у настави повезаног мастер и докторског курса Истраживање података у биоинформатици[3].

Актуелна иницијатива на нивоу курса Увод у биоинформатику јесте израда електронског уџбеника, заснованог на поменутој књизи. Идеја је да заинтересовани студенти као мастер рад обраде по једно поглавље књиге, при чему обрада укључује писање текста на српском језику, али и имплементацију и евентуалну визуелизацију свих или макар већине пратећих алгоритама. Овај рад настао је управо у склопу представљене иницијативе, међу првима.

Уџбеник кроз једанаест глава обрађује разне теме које су занимљиве у оквиру биоинформатике: почетак репликације (алгоритамско загревање), генске мотиве (рандомизовани алгоритми), асемблирање генома (графовски алгоритми), секвенцирање антибиотика/пептида (алгоритми грубе силе), поређење и поравнање геномских секвенци (динамичко програмирање), блокове синтеније (комбинаторни алгоритми), филогенију (еволутивна стабла), груписање гена (кластеровање), проналажење шаблона (префиксна и суфиксна стабла), откривање гена и мутација секвенце (скривени Марковљеви модели), напредно секвенцирање пептида (рачунарска протеомика). Циљ овог рада је обрада десетог поглавља, заснованог на скривеним Марковљевим моделима[12].

Скривени Марковљев модел (у наставку углавном скраћено *HMM*, према енгл. *Hidden Markov Model*), укратко, представља статистички модел који се састоји из следећих елемената: скривених стања (x_i), опсервација (y_i), вероватноћа прелаза (a_{ij}), полазних (π_i) и излазних вероватноћа (b_{ij}), по примеру са слике 1.2. *HMM* се тако може схватити као коначни аутомат, при чему стања задржавају уобичајено значење, док вероватноће прелаза описују колико се често неки прелаз реализује. Полазне вероватноће одређују почетно стање. Овакав аутомат допуњује се идејом да свако стање са одређеном излазном вероватноћом емитује (приказује) неку опсервацију. Штавише, најчешће су само опсервације и познате у раду са *HMM*, док се позадински низ стања погађа („предвиђа"), па се управо зато стања и модели називају скривеним[20].



Слика 1.2: Једноставан пример скривеног Марковљевог модела[21]

У претходном пасусу су, наравно, скривени Марковљеви модели представљени малтене само концептуално, на високом нивоу. У наставку ће, међутим, они бити постепено уведени, заједно са мотивацијом за њихову употребу у виду биолошких проблема који се њима решавају. Према идеји електронског уџбеника, излагање ће пратити књигу *Bioinformatics Algorithms: An Active Learning Approach*, а биће имплементирани и сви пратећи алгоритми. Сви кодови доступни су на *GitHub* репозиторијуму у власништву аутора[8].

Глава 2

Мотивација

За почетак, изложена је мотивација за употребу скривених Марковљевих модела у биоинформатици. Конкретно, представљена су два важна биолошка проблема која се њима могу решити и пратећи појмови из домена, као и једна историјски мотивисана вероватносна мозгалица. Ова глава, дакле, покрива прву петину обрађеног поглавља *Chapter 10: Why Have Biologists Still Not Developed an HIV Vaccine? – Hidden Markov Models*, и то тачно следеће поднаслове: *Classifying the HIV Phenotype*, *Gambling with Yakuza*, *Two Coins up the Dealer's Sleeve*, *Finding CG-Islands*, и највећи део додатка из *Detours*.

2.1 Погађање фенотипа

HIV је вирус хумане имунодефицијенције, један од најпознатијих вируса, који заражава људе широм света. Својим дугорочним деловањем доводи до смртоносног синдрома стечене имунодефицијенције, познатијег као сида или ејдс. Мада поједини аутори распрострањеност *HIV*-а називају пандемијом, Светска здравствена организација означава је као „глобалну епидемију"[4].

Постојање *HIV*-а званично је потврђено почетком осамдесетих година двадесетог века, мада се претпоставља да је са примата на људе прешао знатно раније. Недуго по овом открићу, тачније 1984, из америчког Министарства здравља и услуга становништву најављено је да ће вакцина бити доступна кроз наредне две године. Иако до тога није дошло, председник Бил Клинтон је 1997. потврдио да „није питање *да ли* можемо да произведемо вакцину против сиде, већ је просто питање *када* ће до тога доћи". Вакцина, међутим, ни данас није доступна, а многи покушаји су отказани након што се испоставило

да кандидати чак повећавају ризик од инфекције код појединих испитаника.

Антивирусне вакцине најчешће се праве од површинских протеина вируса на који се циља, у нади да ће имунски систем, након вакцине, у контакту са живим вирусом знатно брже препознати протеине омотача вируса као стране и уништити их пре него што се вирус намножи у телу. *HIV* је, међутим, карактеристичан по томе што врло брзо мутира, па су његови протеини изузетно варијабилни и није могуће научити имунски систем да исправно одреагује на све мутације. Штавише, може се десити да имунитет научи да исправно реагује само на једну варијанту вируса, а да реакција нема никаквог ефекта на остале варијанте. Овакав имунитет је лошији од имунитета који ништа не зна о вирусу, пошто не покушава да научи ништа ново, што је разлог већ поменуте ситуације да су код неких испитаника вакцине кандитати повећали ризик од заразе. Да ствар буде гора, *HIV* брзо мутира и унутар једне особе, тако да је разлика у узорцима узетих од различитих пацијената увек значајна.

Када се све узме у обзир, као обећавајућа замисао за дизајн свеобухватне вакцине намеће се следећа идеја: идентификовати неки пептид који садржи најмање варијабилне делове површинских протеина свих познатих сојева *HIV*-а и искористити га као основу вакцине. Ни то, међутим, није решење, пошто *HIV* има још једну згодну способност: уме да се сакрије процесом гликолизације. Наиме, протеини омотача су махон гликопротеини, што значи да се након превођења за њих могу закачити многобројни гликански (шећерни) ланци. Овим процесом долази до стварања густог гликанског штита, који омета имунски систем у препознавању вируса. Све досад изнето утиче на немогућност прављења прикладне вакцине у скоријем времену.

Чак и ван контекста вакцине, мутације *HIV*-а прилично су занимљиве за разматрање. Конкретно, илустративно је бавити се *env* геном, чија је стопа мутације 1–2 % по нуклеотиду годишње. Овај ген кодира два релативно кратка гликопротеина који заједно граде шиљак (спајк) омотача, део вируса задужен за улазак у људске ћелије. Мање важан део шиљка је гликопротеин *gp41* ($\sim$ 345 аминокиселина), док је важнији гликопротеин *gp120* ($\sim$ 480 аминокиселина). О варијабилности другог говори чињеница да на нивоу једног пацијента, у кратком року, скоро половина аминокиселина буде измењено позадинским мутацијама одговарајућег гена, као да је сасвим други протеин.

Ствари постају још занимљивије када се, поред генотипа вируса, разматра и његов фенотип. Примера ради, сваки вирус *HIV*-а може се означити

као изолат који ствара синцицијум или као изолат који га не ствара. Након уласка у људску ћелију, гликопротеини омотача могу да изазову спајање заражене ћелије са суседним ћелијама. Резултат тога је синцицијум – нефункционална вишеједарна ћелијска (цитоплазматична) маса са заједничком ћелијском мембраном. Овакав изолат *HIV*-а означава се као онај који ствара синцицијум и он се тим процесом знатно брже умножава, што даље значи да је опаснији и агресивнији, јер уласком у само једну ћелију убија многе друге у суседству. Одређивање тачног генотипа и погађање фенотипа важно је како би се пацијенту преписао најприкладнији коктел антивирусних лекова.

Испоставља се да је примарна структура гликопротеина *gp120* важан суштински генотипски предиктор фенотипа *HIV*-а. Наиме, узимајући у обзир само низ аминокиселина које чине *gp120*, може се направити једноставан класификатор који погађа да ли проучавани изолат ствара синцицијум или не. Конкретно, научник Жан Жак де Јонг је 1992. анализирао вишеструко поравнање такозване *V3* петље, издвојеног региона у оквиру *gp120*, и формулисао правило 11/25. Према том правилу, сој *HIV*-а највероватније ствара синцицијум уколико му се на 11. или 25. позицији у *V3* петљи налазе аминокиселине аргинин (R) или лизин (K). Пример мотива *V3* петље дат је на слици 2.1. Приметно је да су управо 11. и 25. позиција међу најваријабилнијим, те да удео критичних R и K на њима није претерано велик. Наравно, на фенотип утичу и многе друге позиције унутар *gp120* и других протеина.

Слика 2.1: Мотив *V3* петље из [12] генерисан помоћу [5]

За крај и поенту уводне приче о *HIV*-у, остаје неразрешен још један веома значајан проблем. Како би се уопште разматрало предвиђање фенотипа на основу примарне структуре *gp120*, неопходно је прво доћи до прецизног вишеструког поравнања различитих секвенци аминокиселина. Прво, поравнање мора бити хируршки прецизно, јер нпр. само једна грешка доводи до погрешног податка која вредност је на 11. и 25. позицији *V3* петље. Следеће,

неопходно је адекватно обрадити инсерције и делеције, што су врло честе мутације *HIV*-а у многим регионима генома. На крају, потребно је на прави начин оценити квалитет поравнања, нпр. коришћењем различитих матрица скора за сваку појединачну позицију. Ово је донекле могуће урадити коришћењем техника представљених у петом поглављу (*Chapter 5: How Do We Compare DNA Sequences? – Dynamic Programming*), али уз два главна проблема: алгортми динамичког програмирања су високе сложености и са мање слободе код скорова, а притом не пресликавају најбоље суштину биолошког проблема класификације фенотипа у алгоритамски проблем (фале кораци након поравнања). Постоји, дакле, потреба за новом формулацијом која обухвата све што је потребно за статистички потковано поравнање секвенци.

2.2 Потрага за генима

Познато је да геном чини тек мали део *DNA* секвенце. Другим речима, *DNA* добрим делом не кодира протеине. Стога је један од важних биолошких проблема управо проналажење места на којима се гени налазе. Прецизније, тражи се место где њихово преписивање (транскрипција) започиње.

Почетком двадесетог века, Фибус Левин открио је да *DNA* чине четири нуклеотида, чији су главни део азотне базе: аденин (A), цитозин (C), гуанин (G) и тимин (T). У то време, међутим, није била позната тачна структура наследног материјала, што је двострука завојница, коју су пола века касније открили Вотсон и Крик. Левин је, стога, сматрао да *DNA* носи информације које су једнаке било којој четворословној азбуци, а додатно је сматрао и да је удео сваког од четири нуклеотида једнак. Занимљивост је да овај упрошћени модел одговара стању у савременој биоинформатици – *DNA* се углавном и посматра као секвенца нуклеотида, односно ниска над азбуком $\{A, C, G, T\}$.

Открићем тачке структуре допуњена је теза о једнаком уделу нуклеотида. Како су нуклеотиди на супротним ланцима упарени, њихов удео јесте врло сличан када се посматра целокупна *DNA*. То, међутим, није случај када се посматра само један ланац, што је уобичајено у генетици и биоинформатици. Примера ради, удео гуанина и цитозина, који чине један базни пар, код људи је 42 %, што је ипак статистички значајно мање од пола. На вишем нивоу гранулације, у случају да се посматрају само по две суседне базе, испоставља се да динуклеотиди CC, CG, GC, GG узимају сасвим различите уделе.

Конкретно, иако би се очекивало да, под претпоставком равномерне расподеле, сваки од њих узима удео 4–5 %, динуклетид CG чини само 1 % људског генома. Све ово значи да је DNA секвенца ипак нешто даље од случајне.

Поставља се питање зашто је удео CG тако мали. Одговор, међутим, није комплексан, поготову ако се додатно примети да је удео TG нешто виши од очекиваног, а посебно у регионима у којима је удео CG изразито мали. Разлог томе лежи у метилацији, најчешћој измени која природно настаје унутар DNA. Поједини нуклеотиди, наиме, могу бити нестабилни, па се на њих лако накачи метил група (CH_3). Међу најнестабилнијим управо је цитозин иза ког следи гуанин, дакле C из CG. Метиловани цитозин даље се често спонтано деаминује у тимин, чиме динуклеотид CG лако постаје TG. Свеукупни резултат је да се CG глобално појављује веома ретко, а TG нешто чешће.

Метилација мења експресију суседних гена. Експресија оних гена чији су нуклеотиди у великој мери метиловани често је потиснута. Иако је сам процес метилације важан у току ћелијске диференцијације – доприноси неповратној специјализацији матичних ћелија – она углавном није пожељна у каснијем добу. Хиперметилација гена повезана је са различитим врстама рака. Стога је метилација врло ретка око гена, што значи да је на тим местима CG знатно чешће. Овакви делови DNA називају се CG острвима или CpG местима. Разлика у уделу динуклеотида у некодирајућим и регионима богатим генима дата је кроз табелу 2.1. Разлика у уделу CG наглашена је црвеном бојом.

Табела 2.1: Удео динуклеотида у једном ланцу људског X хромозома – лево у регионима CG острва, а десно ван њих[12]

	A	C	G	T	A	C	G	T
A	0,053	0,079	0,127	0,036	0,087	0,058	0,084	0,061
C	0,037	0,058	0,058	0,041	0,067	0,063	0,017	0,063
G	0,035	0,075	0,081	0,026	0,053	0,053	0,063	0,042
T	0,024	0,105	0,115	0,050	0,051	0,070	0,084	0,084

Закључак је, дакле, да се проблем потраге за генима може свести на проналажење CG острва. Наиван приступ решавању овог проблема јесте употреба клизајућег прозора. Могао би се узети прозор фиксне величине и померати кроз DNA секвенцу. Они прозори са натпросечним уделом CG били би кандидати за CG острва. Остаје, међутим, питање како одредити добру величину прозора, али и шта радити када преклапајући прозори нуде различиту класификацију подниза. И овде би боље било статистички потковано решење.

2.3 Коцкање са јакузама

Јакузе су припадници истоимене криминалне организације, традиционалног синдиката организованог криминала. Савремене јакузе потичу од јапанских путујућих коцкара, који су били распрострањени у осамнаестом веку. Једна од најпознатијих игара које су путујући коцкари организовали у својим импровизованим коцкарницама био је чо-хан (јап. 丁半, *chō-han*), у буквалном преводу „пар-непар". Игра је сасвим једноставна – претеча јакуза (крупије) баца две коцкице, док се играчи кладе да ли ће збир бити паран или непаран. Игра је такође праведна – једнака је вероватноћа оба исхода парности.

До занимљивог тренутка долази када се из било ког разлога осетно више играча опклади на један од два могућа резултата. Тада би имало смисла да похлепни крупије, у жељи да заради (он узима проценат зараде победника), баца отежане коцкице, које ће са већом вероватноћом дати резултат који је добио мање опклада. Једноставности ради, уместо чо-хана ће у наставку бити разматрана једноставнија игра бацања новчића. У њој крупије баца новчић, а играчи се кладе да ли ће пасти писмо или глава. Она је знатно лакша за анализу, а суштина је иста и доводи до статистички поткованог решења у претходним поднасловима изложених биолошких и сродних проблема.

Крупијева превара у овом случају могла би бити употреба отежаног новчића, код кога исходи нису равномерно расподељени. Нека је познато да крупије има два новчића: један праведан и један отежан тако да на главу пада трипут чешће него на писмо. Циљ је за одређени низ исхода одредити да ли је настао бацањем праведног или отежаног новчића. Пажљивијом анализом проблема, испоставља се да је питање вара ли крупије лоше формулисано. Наиме, оба новчића могу да произведу било који низ исхода, па тако нпр. и отежани новчић може константно да пада на писмо. Иако дефинитивно није могуће са сигурношћу утврдити који је новчић коришћен, могуће је нешто слично и често довољно добро – одредити који је вероватније коришћен.

Конкретно, нека је упитни новчић бачен одређени број пута, при чему је добијен низ исхода. Вероватноће исхода (H од енгл. *heads* – глава и T од енгл. *tails* – писмо) код праведног (F од енгл. *fair* – фер) и отежаног (B од енгл. *biased* – пристрасан) новчића могу се исказати следећим формулама:

$$P_F\{H\} = P_F\{T\} = \frac{1}{2}, P_B\{H\} = \frac{3}{4}, P_B\{T\} = \frac{1}{4}.$$

Како су бацања независни догађаји – претходни исходи ни на који начин

не утичу на наредне – вероватноћа да n бацања произведе низ исхода $x = x_1x_2...x_n$, од којих је пало k глава, јесте производ појединачних вероватноћа:

$$P\{x\} = \prod_{i=1}^{n} P\{x_i\} = P\{H\}^k P\{T\}^{n-k}.$$

Због тога вероватноћа сваког низа исхода код праведног новчића износи:

$$P_F\{x\} = \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = \frac{1}{2^n}.$$

С друге стране, вероватноћа низа исхода код отежаног новчића је:

$$P_B\{x\} = \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{n-k} = \frac{3^k}{4^n}.$$

Уколико је $P_F\{x\} > P_B\{x\}$, онда је вероватније да је крупије бацао праведни новчић, док је у случају $P_F\{x\} < P_B\{x\}$ бацао отежани. Занимљиво је напоменути да ипак није лако израчунати бројеве $1/2^n$ и $3^k/4^n$ за велико n. Они су тада изразито мали, па је питање да ли су добро представљени у рачунару, те да ли њихово поређење даје тачан резултат. Стога се израчунава логаритамски однос вероватноћа, који у конкретном случају износи:

$$\log_2\left(\frac{P_F\{x\}}{P_B\{x\}}\right) = \log_2\left(\frac{2^n}{3^k}\right) = n - k\log_2 3.$$

Овај број се већ без проблема израчунава за разне вредности n и k. Конкретно, нека је $n = 100$ (сто бацања), а $k = 63$ (нешто већи удео глава). Тада је логаритамски однос приближно једнак 0,15. Позитивна вредност $\log(x/y)$ значи да је $x/y > 1$, односно $x > y$ у случају ненегативних вероватноћа. Ово значи да је већа вероватноћа да је крупије бацао праведни новчић, иако је $k = 63$ интуитивно и по апсолутној вредности ближе $3/4 \cdot 100 = 75$ него $1/2 \cdot 100 = 50$. Негативан логаритамски однос довео би до супротног закључка. Алтернативно, како је неопходно одредити само знак израза $n - k\log_2 3$, то се може учинити поређењем n и $k\log_2 3$, односно $k/n = 0{,}63$ и $1/\log_2 3 \approx 0{,}6309$ након дељења k са обе стране. Лева страна је мања, па је однос позитиван.

Изложени вероватносни модел игре пада у воду када се узме у обзир могућност да крупије наизменично баца праведни и отежани новчић. Наиме, искусни преварант могао би да смањи сумњу да користи отежани новчић тако што би га понекад – додуше, ретко, како не би био ухваћен – заменио са праведним, и тако укруг. Поставља се питање како само на основу низа исхода и

евентуално познате вероватноће промене новчића након сваког бацања одредити када је бачен праведни, а када отежани новчић. И овога пута, одговор може бити само несигурног типа – који новчић је када вероватније коришћен.

Слично као код проблема проналажења *CG* острва, потребно је на неки начин различите секвенце новчића упоредити и одредити која је бољи одговор на постављено питање. И овде би наивно решење подразумевало употребу клизајућег прозора који би пролазио кроз све поднизове бацања. На нивоу прозора могли би се рачунати логаритамски односи, према којима би се даље одредило порекло прозора – позитиван однос сугерише да је прозор настао бацањем праведног новчића и супротно. Овакав приступ занемарује тачну вероватноћу замене новчића, мада имплицитно узима у обзир да је она мала.

Остају, међутим, већ поменути проблеми са прозорским приступом: како одредити добру величину прозора, као и шта радити када преклапајући прозори нуде различиту класификацију подниза. Примера ради, ако крупије наизменично баца два претходно описана новчића, а добијени низ исхода је $x = HHHHHTTHHHTTTTT$, онда прозор $x_1...x_{10} = HHHHHTTHHH$ има негативан логаритамски однос, док је однос преклапајућег прозора $x_6...x_{15} = TTHHHTTTTT$ позитиван. Није јасно како одлучити који је новчић бацан у пресеку $x_6...x_{10} = TTHHH$, односно у ком тренутку је тачно дошло до замене новчића, те да ли је замене уопште и било или је крупије праведан.

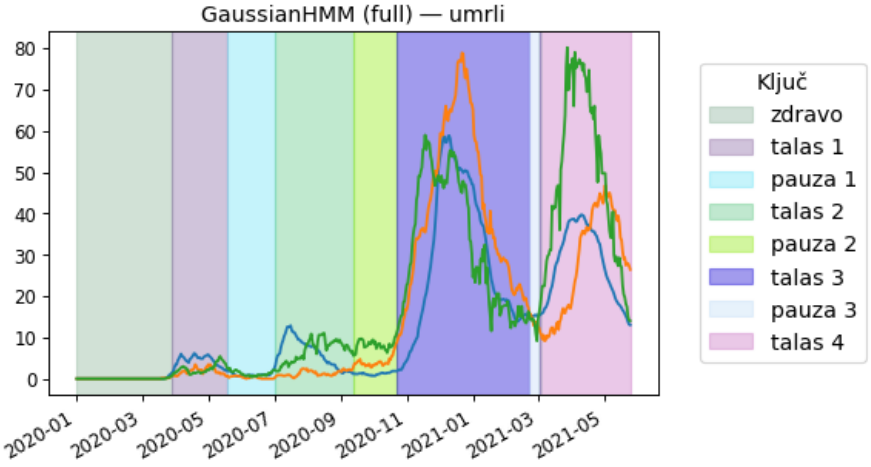
Још једном је јасно да би најбоље било осмислити статистички потковано решење за све досад изложене проблеме. То ће и бити учињено у следећем поглављу, баш са претходно изложеним бацањем новчића као прилично једноставним, али ипак сасвим интуитивним мотивационим примером.

2.4 Још неки проблеми

Досад су изложена два биолошка проблема за која је закључено да би добро било осмислити статистички потковано решење: погађање фенотипа и потрага за генима. Први се своди на класификацију геномске секвенце (нпр. *HIV*-а) на основу познатих могућих исхода и њихових примера. Други се своди на откривање *CG* острва, региона *DNA* са високим уделом динуклеотида *CG*. Иако су ово два конкретна проблема из домена биологије, јасно је да би се жељено решење могло применити и на мноштво других сличних проблема, што укључује последњи мотивациони пример са бацањем новчића.

Приметно је да је секвенцијалност главна особина података са којима се ради при решавању претходно описаних проблема. Први проблем стога се заправо лако уопштава на проблем класификације било каквих секвенцијалних података, под условом да се сличност мери на основу измена које одговарају мутацијама које настају у геному, што су супституције, инсерције и делеције. Други проблем му је сличан, с тим што класификује (заправо групише – кластерује) поднизове једне секвенце. Кад се све узме у обзир, испоставља се да би жељено решење истовремено било корисно како за проблеме надгледаног, тако и ненадгледаног машинског учења над секвенцијалним подацима[16].

Овакво решење могло би се аналогно користити за додељивање новооткривених протеина некој постојећој фамилији[18] (класификација), моделовање и препознавање људског понашања, гестова, рукописа и говора[14] (класификација), обраду звука и сигнала[9] (класификација и кластеровање), одређивање врсте речи у тексту[17] или чак моделовање тока пандемије *COVID-19* у Републици Србији засновано на најосновнијим подацима, као на слици 2.2.



Слика 2.2: Моделовање тока епидемије *COVID-19* у Републици Србији[7]

Досад је увелико наговештено да су добар избор скривени Марковљеви модели (енгл. *Hidden Markov Model, HMM*), па ће надаље бити речи о њима. Ипак, ваља напоменути да се наведени проблеми још ефектније решавају својеврсним проширењима *HMM*-а, попут условних случајних поља[19] (енгл. *Conditional Random Field, CRF*), или комбинацијом са другим техникама као што су вештачке неуронске мреже[11] (енгл. *Artificial Neural Network, ANN*).

Глава 3

Моделовање помоћу *HMM*

Након мотивације, дошло је време за дефиницију скривених Марковљевих модела, као предложеног решења свих досад изложених проблема. Поред дефиниције, на примеру бацања новчића (неправедне коцкарнице) приказано је како се тачно проблеми моделују помоћу *HMM*, те како се на основу тог модела може одговорити на нека важна питања. Ова глава, дакле, покрива другу петину обрађеног поглавља *Chapter 10: Why Have Biologists Still Not Developed an HIV Vaccine? – Hidden Markov Models*, и то тачно следеће поднаслове: *Hidden Markov Models*, *The Decoding Problem*, *Finding the Most Likely Outcome of an HMM*, као и преостали део теоријског додатка из *Detours*.

3.1 Дефиниција *HMM*

Како би се лакше дошло до општег модела свих досадашњих проблема, а посебно бацања новчића, крупије се, уместо као особа, може схватити као примитивна машина – аутомат. Структура аутомата за почетак није важна, али његово деловање јесте. Аутомат је секвенцијалне природе, те оперише кроз низ корака. У сваком кораку је у неком приватном стању, које означава који новчић је заправо бачен (конкретно F и B), при чему јавно приказује исход бацања тог новчића (конкретно H и T). Стање је, дакле, непознато, па се другачије назива скривеним стањем. И стања и приказе (емисије) згодно је апстраховати симболима, нпр. баш карактерима, како је и учињено.

У сваком кораку, аутомат доноси две одлуке: у које скривено стање прећи (да ли га променити) и који симбол емитовати у том новом стању. Испоставља се да се обе одлуке могу донети у потпуности стохастички, што би значило

да је добијен жељени статистички потковани модел проблема. Заиста, прва одлука може се донети тако што се случајно одабере F или B као почетно стање (нпр. баш равномерно, са једнаким вероватноћама $1/2$), а надаље се у сваком кораку стање мења са неком малом вероватноћом (нпр. $1/10$), док се са знатно већом преосталом (нпр. $9/10$) остаје у истом стању. Друга одлука доноси се на основу прве и већ познатих вероватносних особина коцкице – нпр. вероватноћа емитовања H једнака је $1/2$ у стању F, а $3/4$ у стању B.

Претходно изложени аутомат заправо одговара дуго најављиваном појму скривених Марковљевих модела. *HMM* се традиционално представља као статистички модел који се састоји из следећих основних елемената:

- скривених стања (x_i) – свако стање из скупа x има индекс i,
- опсервација, емисија, приказа, симбола (y_i) – такође индексирано,
- полазних вероватноћа (π_i) – колико је често x_i почетно стање,
- вероватноћа прелаза (a_{ij}) – колико се често из x_i прелази у x_j,
- излазних вероватноћа (b_{ij}) – колико се често у стању x_i емитује y_j.

Пример који одговара оваквој дефиницији дат је на слици 1.2. Наравно, подразумева се да су познати број стања n (тако заправо $x = \{x_1, ..., x_n\}$, $\pi = \{\pi_1, ..., \pi_n\}$ и $a = \{a_{ij}\}_{1 \leq i,j \leq n}$) и број могућих опсервација m (тако заправо $y = \{y_1, ..., y_m\}$ и $b = \{b_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq m}$) као помоћни елементи сваког *HMM*. Како су сви скупови коначни, прецизније се говори о дискретним (мултиномијалним) *HMM*, мада је иначе могуће моделовати разне непрекидне расподеле[15].

Како би овакав модел био у потпуности статистички заснован и смислен, обично се захтева да се све појединачне вероватноће сабирају у јединицу:

$$\sum_{i=1}^{n} \pi_i = 1, (\forall i \in \{1, ..., n\}) \sum_{j=1}^{m} a_{ij} = 1, (\forall i \in \{1, ..., n\}) \sum_{j=1}^{m} b_{ij} = 1.$$

Постоје, међутим, изузеци који ће детаљније бити обрађени касније, када се буде говорило о важним надградњама скривених Марковљевих модела.

У овом тренутку је такође значајно нагласити да аутори Компо и Певзнер у уџбенику *Bioinformatics Algorithms* користе нешто другачију нотацију, верну енглеском језику. Наиме, они скуп x означавају као *States*, скуп y као Σ, матрицу a_{ij} као $transition_{l,k}$, а матрицу b_{ij} као $emission_l(b)$. Такође, потребу за скупом полазних вероватноћа – који је заправо опционалан, о чему

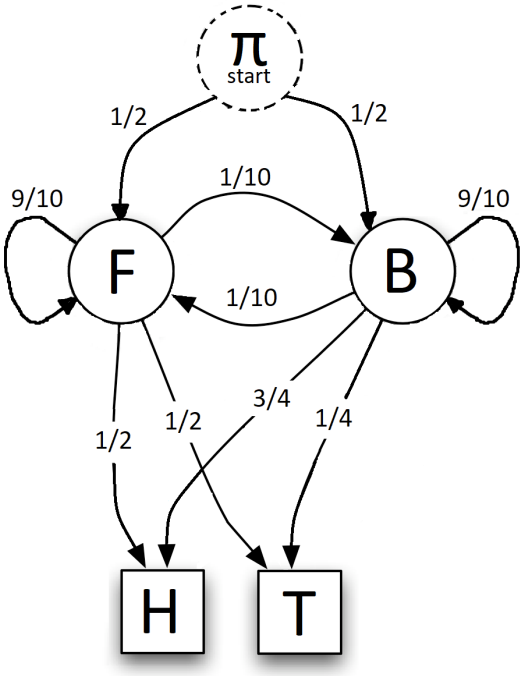
ће бити речи касније – уводе тек касније, па је *HMM* код њих у основи уређена четворка уместо петорка. Овде је ипак одлучено да се користи познатија нотација, како би читаоцима била лакша употреба повезане литературе. Штавише, *HMM* се у литератури често дефинише још простије, као уређена тројка $\{a, b, \pi\}$, односно $\{A, B, \pi\}$ ако се користе велика слова. Стварно, скупови x и y просто се могу заменити индексима, познатим из наведене тројке.

На основу већ разматране слике 1.2, познато је да се *HMM* може илустровати *HMM* дијаграмом. У питању је граф чији су чворови стања и опсервације, а гране вероватноће преласка и емисије. Стил је у суштини произвољан, мада се на слици примећује разлика у значењу графичких елемената. Стања су приказана кружним, а емисије квадратним чворовима. Вероватноће преласка исписане су изнад грана, а излазне вероватноће на самим гранама. Прелази и емисије нулте вероватноће (нпр. прелаз са x_1 на x_3 или на самог себе) нису ни приказани. Други стилови могу приказати све гране, а емисије и вероватноће емисија означити испрекиданим линијама. Независно од стила, *HMM* једнозначно одређује структуру свога дијаграма, а важи и обрнуто.

Сада је могуће искористити *HMM* за прецизно моделовање мотивационог проблема бацања коцкице у неправедној коцкарници. У конкретном случају, изложеном на почетку поднаслова, уређена петорка изгледа овако:

- скривена стања $x = \{F, B\}$ – нпр. $x_1 = F$,
- опсервације $y = \{H, T\}$ – нпр. $y_1 = H$,
- полазне вероватноће $\pi = \left\{\frac{1}{2}, \frac{1}{2}\right\}$ – нпр. $\pi_1 = P\{x_1\} = P\{F\} = \frac{1}{2}$,
- преласци $a = \begin{pmatrix} \frac{9}{10} & \frac{1}{10} \\ \frac{1}{10} & \frac{9}{10} \end{pmatrix}$ – нпр. $a_{11} = P\{x_1 \mapsto x_1\} = P\{F \mapsto F\} = \frac{9}{10}$,
- емисије $b = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}$ – нпр. $b_{11} = P\{y_1|x_1\} = P\{H|F\} = \frac{1}{2}$.

Одговарајући дијаграм приказан је на слици 3.1 и пружа исте информације. Служи се истим стилом као претходно описани граф, с тим што додатно испрекидано приказује замишљено полазно стање, што је новина на слици.



Слика 3.1: Скривени Марковљев модел бацања новчића

3.2 Могућности *HMM*

Могуће је дефинисати појам скривеног пута $p = p_1...p_k$ као низ k стања кроз која *HMM* пролази, а да притом емитује секвенцу опсервација $o = o_1...o_k$. Примера ради, може бити да је низ исхода $o = THTHHHTHTTH$, а позадински низ скривених стања $p = FFFBBBBBFFF$. Главна идеја је анализирати у ком су односу p и o, те са којом се вероватноћом реализују.

Уз излагање *HMM* за бацање новчића у неправедној коцкарници, дати су примери значења чланова петорке, који донекле наговештавају могућности скривених Марковљевих модела. Прво, напоменуто је да полазне вероватноће заправо представљају вероватноћу да се у првом кораку ушло у неко стање. Другим речима, то су заправо вероватноће $P\{p\}$ свих могућих једночланих низова скривених стања. Друго, имплицирано је да матрица емисија складишти маргиналну расподелу емисија при познатом стању. У питању су условне вероватноће $P\{o|p\}$ исхода при једночланом низу скривених стања.

Могуће је, дакле, директно из дефиниције *HMM* израчунати вероватноће $P\{p\}$ и $P\{o|p\}$ за $k = 1$. Према познатој формули условне вероватноће, важи $P\{o, p\} = P\{o|p\}P\{p\}$, па је и та вероватноћа тривијално позната за путеве

јединичне дужине. У питању је заједничка вероватноћа да *HMM* пролази кроз низ стања p, а да притом емитује секвенцу опсервација o.

Подсећања ради, оригинални циљ код неправедне коцкарнице био је пронаћи највероватнији низ стања (бачених новчића) за познати низ опсервација (исхода), што је управо максимална вредност $P\{o, p\}$ по свим p за познато o. Претходно опште постављен задатак проналаска највероватнијег низа скривених стања на основу анализе емитованих симбола постаје сасвим конкретан статистички проблем – на основу ниске симбола o одредити највероватнију секвенцу стања p. У наставку ће бити показано како је све то могуће урадити.

Глава 4

Биолошки значај *HMM*

Након дефинисања скривених Марковљевих модела, описа њихове примене и алгоритама који дају одговоре на важна питања у вези са моделованим проблемом, ред је да се конкретно опише биолошки значај *HMM*, односно њихова примена у изложеним биоинформатичким проблемима. Конкретно, глава која следи бави се потрагом за генима, односно откривањем *CG* острва помоћу *HMM*, као и употребом профилних *HMM* за решавање проблема попут откривања фенотипа *HIV*-а. Она, дакле, покрива трећу и четврту петину обрађеног поглавља *Chapter 10: Why Have Biologists Still Not Developed an HIV Vaccine? – Hidden Markov Models*, и то тачно поднаслове *Profile HMMs for Sequence Alignment* и *Classifying proteins with profile HMMs*.

4.1 Потрага за генима

...

4.2 Профилни *HMM*

...

Глава 5

Учење *HMM*

За крај, прича о скривеним Марковљевим моделима допуњује се још једном важном особином *HMM* – способношћу (машинског) учења поткрепљивањем. Досад је било речи о већ готовим моделима, али прави потенцијал *HMM* показују тек онда када се сви параметри модела науче, уместо да се хардкодирају. Ова глава, дакле, покрива последњу петину обрађеног поглавља *Chapter 10: Why Have Biologists Still Not Developed an HIV Vaccine? – Hidden Markov Models*, и то тачно следеће поднаслове: *Learning the Parameters of an HMM*, *Soft Decisions in Parameter Estimation* и *Baum-Welch Learning*.

Глава 6

Закључак

...

Библиографија

[1] A guide for students. Programa de Pós-Graduação em Bioinformática, Universidade Federal do Paraná (UFPR), Curitiba, уводна реч једног бразилског програма дипломских студија из биоинформатике доступна на: `http://www.bioinfo.ufpr.br/en/a-guide-for-students.html`.

[2] Bioinformatics Algorithms. званични сајт књиге/уџбеника из биоинформатике: `https://www.bioinformaticsalgorithms.org/`.

[3] Bioinformatika. званични сајт курса Увод у биоиформатику и уопшено биоинформатике: `http://www.bioinformatika.matf.bg.ac.rs/`.

[4] Global HIV Programme. World Health Organization, доступно на: `https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics`.

[5] WebLogo, Version 2.8.2 (2005-09-08). Department of Plant and Microbial Biology, University of California, Berkeley, онлајн апликација за илустрацију мотива бесплатно доступна на: `https://weblogo.berkeley.edu/`.

[6] Лазар Васовић. Биоинформатика, 07 2021. Classtools.net, ауторски Венов дијаграм: `https://www.classtools.net/Venn/202107-QTgda5`.

[7] Лазар Васовић. COVID u Srbiji, 05 2021. GitHub, репозиторијум доступан на интернет адреси: `https://github.com/matfija/COVID-u-Srbiji`.

[8] Лазар Васовић. HMM u bioinformatici, 08 2021. GitHub, репозиторијум доступан на: `https://github.com/matfija/HMM-u-bioinformatici`.

[9] Rodrigo Andreão, Bernadette Dorizzi, and Jérôme Boudy. ECG Signal Analysis through Hidden Markov Models. *IEEE*

transactions on bio-medical engineering, 53:1541–9, 09 2006. чланак доступан на: `https://www.researchgate.net/profile/Bernadette-Dorizzi/publication/6872005_ECG_Signal_Analysis_through_Hidden_Markov_Models/links/54aab7730cf25c4c472f4941/ECG-Signal-Analysis-through-Hidden-Markov-Models.pdf`.

[10] Marek Cmero. Frequently Asked Questions about a Career in Bioinformatics, 09 2015. Genome Jigsaw, чланак блога бесплатно доступан на интернет адреси: `https://genomejigsaw.wordpress.com/2015/09/27/faq/`.

[11] Ichael Cohen, David Rumelhart, Nelson Morgan, Horacio Franco, Victor Abrash, and Yochai Konig. Combining Neural Networks And Hidden Markov Models For Continuous Speech Recognition. 06 1999. чланак доступан на интернет адреси: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.50.1857&rep=rep1&type=pdf`.

[12] Phillip Compeau and Pavel Pevzner. *Bioinformatics Algorithms: An Active Learning Approach, 2nd Edition, Vol. II.* Active Learning Publishers, LLC, 2015. претпоследње поглавље *Chapter 10: Why Have Biologists Still Not Developed an HIV Vaccine? – Hidden Markov Models*, стране 178–233.

[13] Nabiilah Ardini Fauziyyah. Bioinformatics: Decoding Nature's Code of Life, 12 2019. Algoritma Technical Blog, чланак блога бесплатно доступан на интернет адреси: `https://algotech.netlify.app/blog/bio-intro/`.

[14] M.J.F. Gales and Steve Young. The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, 1:195–304, 01 2007. доступно на: `https://mi.eng.cam.ac.uk/~mjfg/mjfg_NOW.pdf`.

[15] Michael I. Jordan. Hidden Markov Models & The Multivariate Gaussian, 10 2004. Department of Electrical Engineering and Computer Sciences, UC Berkeley, белешке са предавања бесплатно доступне на интернет адреси: `https://people.eecs.berkeley.edu/~jordan/courses/281A-fall04/lectures/lec-10-26.pdf`.

[16] Ghazaleh Khodabandelou, Charlotte Hug, Rébecca Deneckère, and Camille Salinesi. Supervised vs. Unsupervised Learning for Intentional Process Model Discovery. 06 2014. Business Process Modeling, Development, and

Support (BPMDS), Thessalonique, Greece, чланак доступан на: https://hal-paris1.archives-ouvertes.fr/hal-00994165/document.

[17] Hussain Mutjaba. Frequently Asked Questions about a Career in Bioinformatics, 05 2020. Great Learning, чланак блога доступан на интернет адреси: https://www.mygreatlearning.com/blog/pos-tagging/.

[18] Nam-Phuong Nguyen, Michael Nute, Siavash Mirarab, and Tandy Warnow. HIPPI: highly accurate protein family classification with ensembles of HMMs. *BMC Genomics*, 17:89–100, 11 2016. чланак бесплатно доступан на интернет страници: https://bmcgenomics.biomedcentral.com/track/pdf/10.1186/s12864-016-3097-0.pdf.

[19] Natalia Ponomareva, Paolo Rosso, Ferran Pla, and Antonio Molina Marco. Conditional random fields vs. hidden markov models in a biomedical named entity recognition task. 09 2007. чланак доступан на: http://personales.upv.es/prosso/resources/PonomarevaEtAl_RANLP07.pdf.

[20] Mark Stamp. A Revealing Introduction to Hidden Markov Models. 2021. доступно на: http://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf.

[21] Tdunning. File:HiddenMarkovModel.png. Wikimedia Commons, илустрација скривеног Марковљевог модела са Викимедије: https://commons.wikimedia.org/wiki/File:HiddenMarkovModel.png.