

Универзитет у Београду
Математички факултет



Лазар М. Васовић

Скривени Марковљеви модели (НММ) 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

1	Увод	1
2	Мотивација	4
2.1	Погађање фенотипа	4
2.2	Потрага за генима	7
2.3	Коцкање са јакузама	9
2.4	Још неки проблеми	11
3	Моделовање помоћу <i>HMM</i>	13
3.1	Дефиниција <i>HMM</i>	13
3.2	Могућности <i>HMM</i> (1)	16
3.3	Надградња дефиниције	19
3.4	Могућности <i>HMM</i> (2)	21
4	Биолошки значај <i>HMM</i>	23
4.1	Потрага за генима	23
4.2	Профилни <i>HMM</i>	23
5	Учење <i>HMM</i>	24
6	Закључак	25
	Библиографија	27

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[8]

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[23, 20, 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[22].

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[35].



Слика 1.2: Једноставан пример скривеног Марковљевог модела[36]

У претходном пасусу су, наравно, скривени Марковљеви модели представљени малтене само концептуално, на високом нивоу. У наставку су, међутим, они постепено уведени, заједно са мотивацијом за њихову употребу у виду биолошких проблема који се њима решавају. Према идеји електронског уџбеника, излагање прати књигу *Bioinformatics Algorithms: An Active Learning Approach*, а имплементирани су сви пратећи алгоритми. Резултујући уџбеник са *Python* кодовима, у виду *Jupyter* свезака, доступан је на *GitHub*-у[10].

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 поднасловe: *Classifying the HIV Phenotype, Gambling with Yakuza, Two Coins up the Dealer's Sleeve, Finding CG-Islands*, као 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 кандидати 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 незгодну способност: уме да се сакрије процесом гликозилације. Наиме, протеини омотача су махом 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* (~ 345 аминокиселина), док је важнији гликопротеин *gp120* (~ 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 коктељ 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[26]. 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 [22] генерисан помоћу [6]

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: алгоритми 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[30], 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[38]. Левин је, стога, сматрао да *DNA* носи информације једнаке било којој четворословној азбуци, а додатно 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 тачне 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[22]

	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 коју су путујући коцкари организовали у својим импровизованим коцкарницама био је чо-хан (јап. 丁半, *chō-han*), у дословном преводу „пар-непар”[24]. Игра је сасвим једноставна – претеча јакуза (крупје) баца две коцкице, док се играчи кладе да ли ће збир бити паран или непаран. Игра је такође поштена – једнако се остварују 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 крупје, 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 је у наставку разматрана нешто простија игра бацања новчића. У њој крупје 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.

Крупјева превара у овом случају могла би бити употреба отежаног новчића, код кога исходи нису равномерно расподељени. Нека је познато да крупје 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 крупје 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:

$$P\{H|F\} = P\{T|F\} = \frac{1}{2}, P\{H|B\} = \frac{3}{4}, P\{T|B\} = \frac{1}{4}.$$

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

утичу на наредне – вероватноћа да n бацања произведе низ исхода $x = x_1 \dots x_n$, 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:

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

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

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

Уколико је $P\{x|F\} > P\{x|B\}$, онда је вероватније да је крупije бацао праведни новчић, док је у случају $P\{x|F\} < P\{x|B\}$ 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:

$$\log_2 \left(\frac{P\{x|F\}}{P\{x|B\}} \right) = \log_2 \left(\frac{2^n}{3^k} \right) = n - k \log_2 3.$$

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 крупije 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 крупije 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 крупije наизменично баца два претходно описана новчића, а добијени низ исхода је $x = \text{НННННТТНННТТТТТ}$, онда прозор $x_1...x_{10} = \text{НННННТТННН}$ има негативан логаритамски однос, док је однос преклапајућег прозора $x_6...x_{15} = \text{ТТНННТТТТТ}$ позитиван. Није јасно како одлучити који је новчић бачан у пресеку $x_6...x_{10} = \text{ТТННН}$, односно у ком тренутку је тачно дошло до замене новчића, те да ли је замене уопште и било или је крупije поштен.

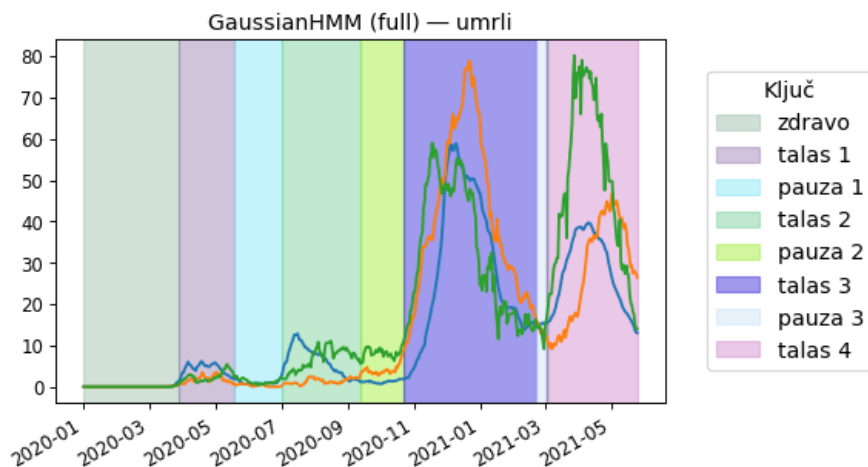
Још једном је јасно да би најбоље било осмислити статистички потковано решење за све досад изложене проблеме. То је и учињено у следећем поглављу, баш са претходно изложеним бацањем новчића као прилично једноставним, али ипак сасвим интуитивним мотивационим примером.

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[28].

Овакво решење могло би се аналогно користити за додељивање новооткривених протеина некој постојећој фамилији[32] (класификација), моделовање и препознавање људског понашања, гестова, рукописа и говора[25] (класификација), обраду звука и сигнала[13] (класификација и кластеровање), одређивање врсте речи у тексту[31] или чак моделовање тока пандемије *COVID-19* у Републици Србији засновано на најосновнијим подацима, као на слици 2.2.



Слика 2.2: Моделовање епидемије *COVID-19* у Србији[9]

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[33] (енгл. *Conditional Random Field*, *CRF*), или комбинацијом са другим техникама као што су вештачке неуронске мреже[21] (енгл. *Artificial Neural Network*, *ANN*).

Глава 3

Моделовање помоћу *НММ*

Након мотивације, дошло је време за дефиницију скривених Марковљевих модела, као предложеног решења свих досад изложених проблема. Поред дефиниције, на примеру бацања новчића (непоштене коцкарнице) приказано је како се тачно проблеми моделују помоћу *НММ*, 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 поднаслов: *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 *НММ*

Како би се лакше дошло до општег модела свих досадашњих проблема, а посебно бацања новчића, крупније 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 надале 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. НММ 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 $x = \{x_1, \dots, x_n\}$, $\pi = \{\pi_1, \dots, \pi_n\}$ и $a = \{a_{ij}\}_{1 \leq i, j \leq n}$) и број могућих опсервација m (тако заправо $y = \{y_1, \dots, y_m\}$ и $b = \{b_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq m}$) као помоћни елементи сваког НММ. Како су сви скупови коначни, прецизније се говори о дискретним (мултиномијалним) НММ, мада је иначе могуће моделовати разне непрекидне расподеле[27].

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

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

Постоје, међутим, изузеци који су детаљније обрађени у наставку, када се говори о важним надградњама појма 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 $transition_{i,k}$, а матрицу b_{ij} као $emission_i(b)$. Такође, потребу за скупом полазних вероватноћа – који је заправо опционалан, о чему

ће бити речи касније – уводе тек касније, па је *НММ* 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, *НММ* 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 *НММ* може илустровати *НММ* 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, *НММ* једнозначно одређује структуру свога дијаграма, а важи и обрнуто.

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

- скривена стања $x = \{F, B\}$ – нпр. $x_1 = F$ и $x_2 = B$,
- опсервације $y = \{H, T\}$ – нпр. $y_1 = H$ и $y_2 = T$,
- 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. $a_{12} = P\{x_1 \mapsto x_2\} = P\{F \mapsto B\} = \frac{1}{10}$,
- емисије $b = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}$ – нпр. $b_{21} = P\{y_1|x_2\} = P\{H|B\} = \frac{3}{4}$.

Одговарајући дијаграм приказан је на слици 3.1 и пружа исте информације. Служи се истим стилем 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

Историјски гледано, појам *HMM* увели су Ленард Баум и сарадници кроз низ статистичких радова објављених у другој половини шездесетих година двадесетог века[16]. У питању је надградња појма Марковљевих ланаца (енгл. *Markov Chain*, *MC*), који су у суштини *HMM* без емисија. Ради се, дакле, о уобичајеном стохастичком аутомату, који се састоји из стања и вероватноћа прелаза. *MC* је почетком века формулисао руски статистичар Андреј Марков, по коме су и названи, како би моделовао Марковљеве процесе – стохастичке промене стања такве да тренутно стање зависи искључиво од претходног[7]. Прва практична примена *HMM* била је препознавање говора, док је биолошка примена почела 1986, Бишоповим и Томпсоновим поравнањем *DNA*[18].

3.2 Могућности *HMM* (1)

Могуће је дефинисати појам скривеног пута $p = p_1 \dots p_k$ као низ k стања кроз која *HMM* пролази, а да притом емитује секвенцу опсервација $o = o_1 \dots o_k$. Примера ради, може бити да је низ видљивих исхода $o = THTHTHTHTH$, 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 *НММ* за бацање новчића у непоштеној 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 *НММ* израчунати вероватноће $P\{p\}$ и $P\{o|p\}$ за $k = 1$, и то као $P\{x_i\} = \pi_i$, односно $P\{y_j|x_i\} = b_{ij}$. Према познатој формули условне вероватноће, важи $P\{p, o\} = P\{p\}P\{o|p\}$, па је и та вероватноћа тривијално позната за путеве јединичне дужине, као $P\{x_i, y_j\} = \pi_i b_{ij}$. У питању је заједничка вероватноћа да *НММ* пролази кроз низ стања p , а да притом емитује управо секвенцу опсервација o . Према уобичајеним принципима, могуће је приметити следеће: $\sum_o \sum_p P\{p, o\} = 1$. Наиме, када се саберу вероватноће свих могућих комбинација низа емисија и скривених путева одређене дужине k , добија се јединица, што значи да је покривен цео простор догађаја у *НММ*. Из ове дводимензионалне (заједничке) расподеле путева и емисија могу се без проблема извести маргиналне (појединачне) расподеле путева $P\{p\} = \sum_o P\{p, o\}$ и симбола $P\{o\} = \sum_p P\{p, o\}$.

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

За почетак, добро је формално дефинисати проблем. Већ је закључено да формулација попут 1 није добра нити смислена. Зато је и уведен појам *НММ*.

Проблем 1: Непоштена коцкарница

На основу низа исхода бацања новчића, одредиши када круиш у непоштеној коцкарници корисћи који од два моћућа новчића.

Улаз: низ $o = o_1 \dots o_k$ исхода (H и T) бацања два новчића (F и B).

Излаз: низ $p = p_1 \dots p_k$ новчића такав да је o_i резултат бацања p_i .

Добра формулација преко појма *НММ* дата је кроз проблем 2. Управо је она детаљно обрађена у наставку овог поднаслова, као његов централни део.

Проблем 2: Декодирање приказа

Пронаћи оптимални пут кроз НММ ако је емитована ниска o .

Улаз: ниска $o = o_1 \dots o_k$ и *НММ* $\{a, b, \pi\}$ који ју је емитовао.

Издаз: скривени пут p који максимизује вероватноћу $P\{p, o\}$ над свим могућим путевима, дакле $\arg\max_p P\{p, o\}$ за улазно o .

Прва идеја јесте исцрпна претрага простора догађаја над маргиналном расподелом $P\{p, o\}$ за познато o . Како је $P\{p, o\} = P\{p\}P\{o|p\}$, тако је најзгодније независно израчунати $P\{p\}$ и $P\{o|p\}$ за сваки од n^k скривених путева. Број путева дужине k у *НММ* са n могућих стања је експоненцијалан, јер се одабир своди на варијације – уређене изборе са понављањем.

Први потпроблем је израчунавање вероватноће пута, што се може формализовати проблемом 3. Он је у наставку решен у виду једне формуле.

Проблем 3: Вероватноћа скривеног пута[11]

Израчунајти вероватноћу скривеног пута p кроз НММ.

Улаз: скривени пут $p = p_1 \dots p_k$ кроз *НММ* $\{a, b, \pi\}$.

Издаз: вероватноћа улазног пута $P\{p\}$.

Први елемент $P\{p\}$, дакле, представља вероватноћу скривеног пута p , односно вероватноћу да *НММ* прође кроз низ стања p . Већ је показано да за једночлане путеве важи $P\{x_i\} = \pi_i$. Вишечлани путеви заправо почињу једночланим, а онда се проширују користећи стохастичке прелазе. Стога је $P\{p_1 p_2 \dots p_{k-1} p_k\} = P\{p_1\}P\{p_1 \mapsto p_2\} \dots P\{p_{k-1} \mapsto p_k\}$. Објашњено је већ и да је $P\{x_i \mapsto x_j\} = a_{ij}$, па се свеукупно вероватноћа пута може израчунати као:

$$P\{p\} = P\{p_1\} \prod_{i=2}^k P\{p_{i-1} \mapsto p_i\} = \pi_{ind(p_1)} \prod_{i=2}^k a_{ind(p_{i-1}), ind(p_i)}.$$

Други потпроблем је израчунавање вероватноће исхода при познатом путу, што се може формализовати као 4. И то се решава само једном формулом.

Други елемент $P\{o|p\}$, дакле, представља вероватноћу да *НММ* емитује ниску o при проласку кроз низ стања p . Већ је показано да за једночлане путеве важи $P\{y_j|x_i\} = b_{ij}$. Код вишечланих нема разлике, пошто

Проблем 4: Вероватноћа исхода на путу[12]

Израчунајте вероватноћу приказа o на путу p кроз НММ.

Улаз: скривени пут $p = p_1 \dots p_k$ кроз $\text{НММ}\{a, b, \pi\}$ и ниска $o = o_1 \dots o_k$ која је тим проласком емитована.

Ислаз: условна вероватноћа приказа на путу $P\{o|p\}$.

је пут фиксиран и само се прате опсервације. Стога је $P\{o_1 \dots o_k | p_1 \dots p_k\} = P\{o_1 | p_1\} \dots P\{o_k | p_k\}$. Свеукупно се вероватноћа пута може израчунати као:

$$P\{o|p\} = \prod_{i=1}^k P\{o_i | p_i\} = \prod_{i=1}^k b_{\text{ind}(p_i), \text{ind}(o_i)}.$$

3.3 Надградња дефиниције

Пре коначног решавања проблема декодирања, у дигресији која следи дорађена је дефиниција скривених Марковљевих модела, што доприноси једноставнијем раду са њима. Наиме, како би претходне формуле биле лакше за комбиновање и конкретну имплементацију, добро је на следећи начин надградити *НММ* и сродне појмове попут скривеног пута и низа опсервација:

- уводи се експлицитно почетно стање x_0 уместо одвојеног праћења полазних вероватноћа π , чиме свако π_i постаје део матрице прелаза a_{0i} ,
- почетно стање се увек подразумева, као нулти члан скривеног пута, па тако свако $p = p_1 \dots p_k$ постаје $p = p_0 p_1 \dots p_k$, и то тако да је $p_0 = x_0$,
- уводи се нулта емисија y_0 , што је заправо празан карактер, чиме се дозвољава да стања буду тиха и не емитују ништа, као почетно стање,
- матрице a_{ij} и b_{ij} постају мапе a_{x_i, x_j} и b_{x_i, y_j} , што знатно олакшава рад, а исто важи и за низ π_i , ако се чува (прослеђује), који постаје мапа π_{x_i} ; у вези са тим, из мапа се може прочитати скуп скривених стања и опсервација, чиме се *НММ* дефинитивно своди на тројку $\{a, b, \pi\}$.

Оваква дорада свој пун потенцијал показује у напреднијим применама, мада је и њен почетни допринос незанемарљив. Формуле сада постају:

$$P\{p\} = \pi_{p_1} \prod_{i=2}^k a_{p_{i-1}, p_i} = \prod_{i=1}^k a_{p_{i-1}, p_i}, \quad P\{o|p\} = \prod_{i=1}^k b_{p_i, o_i}.$$

Заједничка формула вероватноће проласка кроз пут p и приказа o јесте:

$$P\{p, o\} = P\{p\}P\{o|p\} = \prod_{i=1}^k a_{p_{i-1}, p_i} \prod_{i=1}^k b_{p_i, o_i} = \prod_{i=1}^k a_{p_{i-1}, p_i} b_{p_i, o_i}.$$

Интуитивно, заједнички догађај заправо представља низ независних догађаја прелаза и емисија, па је зато $P\{p, o\} = a_{p_0, p_1} b_{p_1, o_1} \dots a_{p_{k-1}, p_k} b_{p_k, o_k}$, дакле прелаз из почетног стања у p_1 , па емисија o_1 у p_1 , затим прелаз из p_1 у p_2 , и тако даље. Све ове формуле дају елегантан начин рачунања само уз помоћ a и b .

Ваља искористити прилику и поменути још неке важне надградње НММ из литературе, које су у стварности применљивије од основне верзије:

- опсервације y могу бити бесконачан скуп, извучене из неке непрекидне расподеле; тада се мапа вероватноћа b_{ij} посматра као мапа расподела b_i , која складишти расподеле (густине расподела) емисија стања x_i ,
- само нека стања се означавају као завршна или се уводи експлицитно завршно стање, нпр. у последњем бацању је обавезан праведни новчић,
- уместо нестабилних правих вероватноћа користе се логаритамске вероватноће, што ублажава рачунске грешке, мада усложњава алгоритме.

Пожељно је усвојити последњу надградњу, након које формуле постају (подсетник на правило – логаритам производа је збир логаритама):

$$P_{\log}\{p\} = \log P\{p\} = \log \pi_{p_1} + \sum_{i=2}^k \log a_{p_{i-1}, p_i} = \sum_{i=1}^k \log a_{p_{i-1}, p_i},$$

$$P_{\log}\{o|p\} = \log P\{o|p\} = \sum_{i=1}^k \log b_{p_i, o_i},$$

$$P_{\log}\{p, o\} = \log P\{p, o\} = \sum_{i=1}^k (\log a_{p_{i-1}, p_i} + \log b_{p_i, o_i}).$$

Заправо је најефикасније директно радити са логаритамским вероватноћама, односно све вероватноће одмах логаритмовати, укључујући улазне из мапа a и b . Под овом претпоставком, формуле су лакше за запис и рачун:

$$P_{\log}\{p\} = \pi_{\log, p_1} + \sum_{i=2}^k a_{\log, p_{i-1}, p_i} = \sum_{i=1}^k a_{\log, p_{i-1}, p_i},$$

$$P_{\log}\{o|p\} = \sum_{i=1}^k b_{\log,p_i,o_i}, P_{\log}\{p,o\} = \sum_{i=1}^k (a_{\log,p_{i-1},p_i} + b_{\log,p_i,o_i}).$$

Надграђени НММ сада се може свести на једноставну уређену двојку:

- мапа логаритамских вероватоћа прелаза a_{\log,x_i,x_j} ,
- мапа логаритамских излазних вероватноћа b_{\log,x_i,y_j} .

За конструкцију оваквог објекта треба имати оригинално a и b , као и π , па се зато ипак, интуиције ради, НММ и даље званично сматра уређеном тројком $\{a, b, \pi\}$, а не интерно коришћеном трансформисаном двојком $\{a_{\log}, b_{\log}\}$.

Надграђени НММ моделује непоштену коцкарницу на следећи начин:

- прелази $a_{\log} = \begin{matrix} & F & B \\ \pi & \begin{pmatrix} \log \frac{1}{2} & \log \frac{1}{2} \\ \log \frac{9}{10} & \log \frac{1}{10} \\ \log \frac{1}{10} & \log \frac{9}{10} \end{pmatrix} \end{matrix}$ – нпр. $a_{\log,F,B} = P_{\log}\{F \mapsto B\}$,
- емисије $b_{\log} = \begin{matrix} & H & T \\ F & \begin{pmatrix} \log \frac{1}{2} & \log \frac{1}{2} \\ \log \frac{3}{4} & \log \frac{1}{4} \end{pmatrix} \\ B & \end{matrix}$ – нпр. $b_{\log,B,H} = P_{\log}\{H|B\} = \log \frac{3}{4}$.

Згодно је запамтити и следеће вредности као помоћне елементе модела:

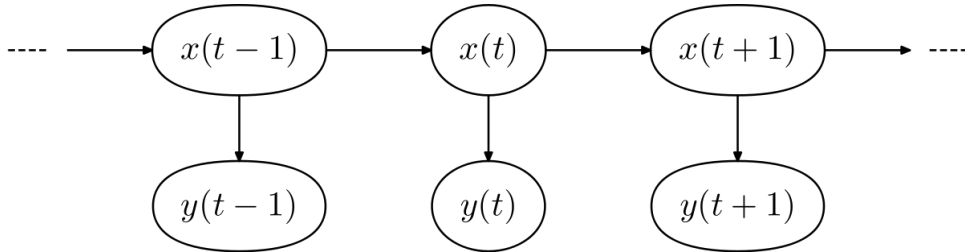
- скуп скривених стања x и њихов број n ,
- скуп могућих емисија y и њихов број m ,
- мапу логаритамских полазних вероватноћа π_{\log} ,
- оригиналне вредности у мапама a, b, π .

3.4 Могућности НММ (2)

Одређивањем формуле $P\{p,o\}$ за путеве произвољне дужине, могуће је приступити проблему максимизације. Како је већ предложено, наивна идеја исцрпне претраге састоји се од генерисања сваког од n^k скривених путева p ,

израчувања $P\{p, o\}$ за познати низ приказа o , и на крају одабира пута који представља $\operatorname{argmax}_p P\{p, o\}$. Логаритам је монотона трансформација, тако да се задатак не мења ни када се посматрају стабилније вредности $P_{\log}\{p, o\}$, односно важи $\operatorname{argmax}_p P_{\log}\{p, o\} = \operatorname{argmax}_p P\{p, o\}$. Из изведене формуле је очигледно да је за свако израчунавање заједничке вероватноће потребно $O(k)$ корака, па је укупна временска сложеност наивног приступа $O(n^k k)$, што је релативно прихватљиво за кратке скривене путеве и мали број стања.

Путеви су, међутим, често врло дугачки, а НММ имају велики број скривених стања, те наивни приступ није прихватљив у општем случају. Стога је амерички инжењер електротехнике Ендрју Витерби 1967. предложио ефикасније решење[37], засновано на посебном графу, који се може схватити као врста Менхетн графа, појма који је представљен у петом поглављу уџбеника (*Chapter 5: How Do We Compare DNA Sequences? – Dynamic Programming*). У питању је Витербијев граф, осмишљен на основу основног својства сваког Марковљевог модела, представљеног на слици 3.2. Наиме, сваки НММ моделује један Марковљев процес, што је поменуто при дефиницији. Последица је да је тренутно стање зависи искључиво од претходног и ниједног другог.



Слика 3.2: Ток времена код скривених Марковљевих модела[34]

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 непосредно опише биолошки значај *HMM*, односно њихова примена у досад 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 поднасловe *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 поднасловe: *Learning the Parameters of an HMM*, *Soft Decisions in Parameter Estimation* и *Baum-Welch Learning*.

Глава 6

Закључак

Досад је изложен појам скривених Марковљевих модела, као и њихов био-информатички значај. Дата је детаљна мотивација за увођење статистички потованог аутомата, након чега је појам *НММ* разрађен на мотивационом примеру непоштене коцкарнице (бацање два новчића). Затим је и примењен на решавање важних биолошких проблема, попут проналажења *CG* острва (места са генима) и напредног бављења генским и протеинским профилима.

У последњој глави овог рада су надаље сумиране информације из закључних подналова обрађеног поглавља *Chapter 10: Why Have Biologists Still Not Developed an HIV Vaccine? – Hidden Markov Models*, и то тачно *The Many Faces of HMMs* и *Epilogue: Nature is a Tinkerer and not an Inventor*, мада су поменути и додатни подаци из помоћног подналова *Bibliography Notes*.

Значајна напредна примена *НММ* која превазилази оквире уџбеника јесте моделовање отпорности *HIV*-а на лекове. У уводној мотивацији поменуто је да се заражени пацијенти лече коктемом антивирусних лекова, који је због високе стопе мутација често посебно осмишљен за сваког појединца, како би терапија била успешна. Мутације могу да онеспособе дејство неког лека који је раније имао ефекта. Стога је разумевање отпорности од високог значаја. Нико Беренвинкел и Матијас Дртон су 2006. предложили модел реактивности соја на лекове заснован баш на *НММ*, додуше изразито комплексном[17].

Када су протеини у питању, ваља напоменути да се они у суштини састоје из више повезаних целина које се називају доменима. Домени могу бити различитих структура и функција, и управо се они чешће анализирају него цели протеини. Године 2002. Бејтман и сарадници описали су употребу профилних *НММ*, на основу чега је осмишљена позната база података Пфам[15]. Она се

данас састоји од скоро 20.000 вишеструких поравнања разних протеинских домена и рутински се користи у анализи нових протеинских секвенци[5].

Све у свему, скривени Марковљеви модели прешли су дуг пут од својих првих употреба у рачунарској биологији (нпр. Черчил 1989[19], Круг и сарадници 1994[29], Балди и сарадници 1994[14]) до данашње широке биоинформатичке примене. Поменуто је употреба *НММ* за моделовање и препознавање људског понашања, гестова, рукописа и говора, обраду звука и сигнала, одређивање врсте речи у тексту или чак моделовање тока пандемије *COVID-19* у Републици Србији засновано на најосновнијим подацима, као на слици 2.2. Објашњен је значај *НММ* како код проблема надгледаног, тако и код проблема ненадгледаног машинског учења. Наведене су многе могућности *НММ*, укључујући способност учења свих параметара модела поткрепљивањем.

Паралелно са писањем овог текста, направљен је електронски уџбеник, као суштински најзначајнији допринос рада. Уџбеник је реализован у виду *Jupyter* свезака, које су заједно са свим осталим материјалима доступне на *GitHub*-у[10]. Концепт је такав да свеске садрже једнак текст као у писаном раду, али успут складиште и *Python* кодове који имплементирају у тексту изложене алгоритме. Имплементирана су сва предложена решења из књиге *Bioinformatics Algorithms*, али и многа друга. Како се кодови интерпретирају, они су у потпуности интерактивни и могу лако послужити за самосталан студентски рад и детаљније упознавање са имплементацијама. За случај да читаоцу нису доступни *Python* интерпретатор и/или *Jupyter* сервер, направљена је и *HTML* верзија материјала, која, додуше, није интерактивна.

Свеукупно, обрађена лекција електронског уџбеника доприноси усвајању знања о скривеним Марковљевим моделима и њиховој примени у биоинформатици, независно од тога да ли читалац слуша мастер курс Увод у биоинформатику на Математичком факултету Универзитета у Београду. За разумевање је неопходно само основно предзнање из математике (углавном вероватноће) и биологије (углавном генетике), што је ниво средње школе. Било би добро да иницијатива у оквиру које уџбеник настаје заживи, те да у најскоријем року свака лекција буде доступна у потпуности у електронском облику.

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 биоинформатику и уопштено 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] Pfam 34.0 (March 2021, 19179 entries). <http://pfam.xfam.org/>.
- [6] 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/>.
- [7] Андрей Андреевич Марков. Распространение закона больших чисел на величины, зависящие друг от друга. *Известия физико-математической общесѣва при Казанском университете*, 2(15):135–6, 1906.
- [8] Лазар Васовић. Биоинформатика, 07 2021. Classtools.net, ауторски Венов дијаграм: <https://www.classtools.net/Venn/202107-QTgda5>.
- [9] Лазар Васовић. COVID u Srbiji, 05 2021. GitHub, репозиторијум доступан на интернет адреси: <https://github.com/matfija/COVID-u-Srbiji>.

- [10] Лазар Васовић. HMM u bioinformatici, 08 2021. GitHub, репозиторијум доступан на: <https://github.com/matfija/HMM-u-bioinformatici>.
- [11] Compute the Probability of a Hidden Path, 09 2015. ROSALIND, задатак из уџбеника доступан на: <http://rosalind.info/problems/ba10a/>.
- [12] Compute the Probability of an Outcome Given a Hidden Path, 09 2015. ROSALIND, доступно на: <http://rosalind.info/problems/ba10b/>.
- [13] 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.
- [14] Pierre Baldi, Yves Chauvin, Tim Hunkapiller, and Marcella McClure. Hidden Markov Models of Biological Primary Sequence Information. *Proceedings of the National Academy of Sciences of the United States of America*, 91:1059–63, 03 1994. <https://www.pnas.org/content/pnas/91/3/1059.full.pdf>.
- [15] Alex Bateman, Ewan Birney, Lorenzo Cerruti, Laurence Ettwiller, Sean Eddy, Sam Griffiths-Jones, Kevin Howe, Mhairi Marshall, and Erik Sonnhammer. The Pfam Protein Families Database. *Nucleic acids research*, 30:276–80, 02 2002. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC99071/>.
- [16] L. E. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–63, 12 1966. преузимање на: <https://projecteuclid.org/journalArticle/Download?urlId=10.1214%2Faoms%2F1177699147>.
- [17] Niko Beerenwinkel and Mathias Drton. A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data. *Biostatistics*, 8(1):53–71, 03 2006. чланак доступан на интернет адреси: <https://academic.oup.com/biostatistics/article-pdf/8/1/53/697249/kxj033.pdf>.
- [18] M. Bishop and E. Thompson. Maximum Likelihood Alignment of DNA Sequences. *Journal of Molecular Biology*, 190(2):159–65, 07 1986. апстракт доступан на: <https://pubmed.ncbi.nlm.nih.gov/3641921/>.

- [19] Gary A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51(1):79–94, 1989. апстракт доступан на интернет адреси: <https://pubmed.ncbi.nlm.nih.gov/2706403/>.
- [20] 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/>.
- [21] 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 бесплатно доступан на интернет адреси: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.50.1857&rep=rep1&type=pdf>.
- [22] 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.
- [23] 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/>.
- [24] JC Fletcher. Learning Japanese board game culture from Yakuza 0, 03 2017. Polygon, чланак блога доступан на интернет адреси: <https://www.polygon.com/2017/3/10/14848222/learning-japanese-board-game-culture-from-yakuza-0>.
- [25] 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.
- [26] J Jong, A Ronde, Wilco Keulen, Matthijs Tersmette, and Jaap Goudsmit. Minimal requirements for the human immunodeficiency virus type 1 v3 domain to support the syncytium-inducing phenotype: Analysis by single amino acid substitution. *Journal of virology*, 66:6777–80, 12 1992. чланак бесплатно доступан на интернет адреси: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC240176/pdf/jvirol100042-0547.pdf>.

- [27] 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>.
- [28] 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>.
- [29] Anders Krogh, Michael Brown, Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–31, 1994. апстракт доступан на: <https://pubmed.ncbi.nlm.nih.gov/8107089/>.
- [30] P. A. Levene. On the biochemistry of nucleic acids. *Journal of the American Chemical Society*, 32(2):231–240, 1910. делимично доступно на: <https://pubs.acs.org/doi/10.1021/ja01920a010>.
- [31] 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/>.
- [32] 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://bmcbgenomics.biomedcentral.com/track/pdf/10.1186/s12864-016-3097-0.pdf>.
- [33] 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.
- [34] Qef. File:Hmm_temporal_bayesian_net.svg. Wikimedia Commons, илустрација скривеног Марковљевог модела са Викимедије: https://commons.wikimedia.org/wiki/File:Hmm_temporal_bayesian_net.svg.

- [35] Mark Stamp. A Revealing Introduction to Hidden Markov Models. 2021. доступно на: <http://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf>.
- [36] Tdunning. File:HiddenMarkovModel.png. Wikimedia Commons, илустрација скривеног Марковљевог модела са Викимедије: <https://commons.wikimedia.org/wiki/File:HiddenMarkovModel.png>.
- [37] Andrew James Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–9, 04 1967. доступно на: <https://www.asc.ohio-state.edu/goel.1//STAT825/PAPERS/viterbiErrBnds.pdf>.
- [38] James D. Watson and Francis H. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 04 1953. чланак бесплатно доступан на: <http://dosequis.colorado.edu/Courses/MethodsLogic/papers/WatsonCrick1953.pdf>.