

Универзитет у Београду
Математички факултет



Лазар М. Васовић

Скривени Марковљеви модели у
биоинформатици – електронска лекција

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 у биоинформатици – електронска лекција

Резиме: ...

Кључне речи: биоинформатика, скривени Марковљеви модели (*HMM*), електронска лекција, *CG* острва (*CpG* места), профилни *HMM* (*HMM* профили)

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	Додатни проблеми	11
3	Моделовање	13
3.1	Дефиниција модела	13
3.2	Могућности модела	16
3.3	Надградња дефиниције	19
3.4	Витербијев алгоритам	22
3.5	Алгоритам „напред”	27
4	Биолошки значај	31
4.1	Гени – два стања	31
4.2	Гени – више стања	36
4.3	Профилни модели	42
4.4	Рад са профилима	50
5	Учење модела	62
6	Закључак	63
	Библиографија	65

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 представља дијаграмом са слике 1.1.



Слика 1.1: Венов дијаграм интердисциплинарности[10]

Овако представљена, биоинформатика је заправо спој статистике, рачу-

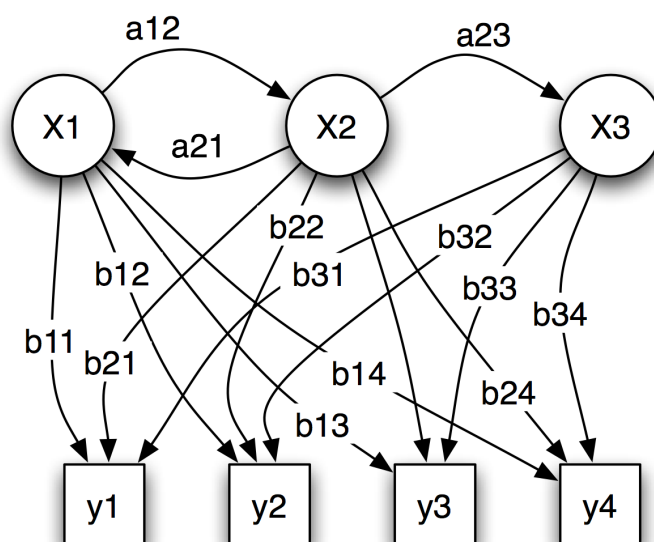
нарства и биологије – сва три истовремено – по чему надилази појединачне спојеве: биостатистику, науку о подацима и рачунарску биологију. Конкретно, статистички (математички) апаратат служи за рад са подацима, рачунарске технологије тај апарат чине употребљивијим, док биологија даје потребно доменско 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[1, 28, 33].

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[30].

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 опажања и позната у раду са *HMM*, док се позадински низ стања погађа („предвиђа”), па се управо зато стања и модели називају скривеним[53].



Слика 1.2: Једноставан пример скривеног Марковљевог модела[54]

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[12].

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[39]. 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 [30] генерисан помоћу [8]

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 велике 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[46], 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[56]. 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[30]

	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 добро дошло 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”[34]. 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:

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

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 крупје 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 крупје 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 крупје 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.

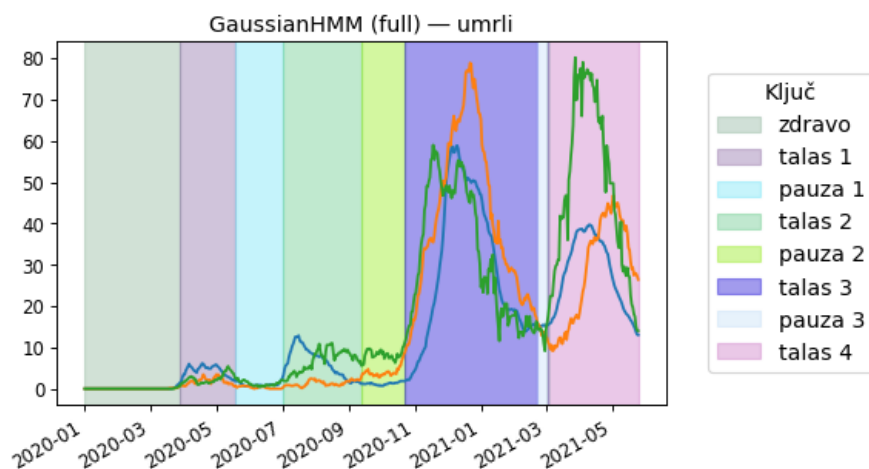
Још једном је јасно да би најбоље било осмислити статистички потковано решење за све досад изложене проблеме. То је и учињено у следећем поглављу, баш са претходно изложеним бацањем новчића као прилично једноставним, али ипак сасвим интуитивним мотивационим примером.

2.4 Додатни 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[42].

Овакво решење могло би се аналогно користити за додељивање новооткривених протеина некој постојећој фамилији[48] (класификација), моделовање и препознавање људског понашања, гестова, рукописа и говора[35] (класификација), обраду звука и сигнала[20] (класификација и кластеровање), одређивање врсте речи у тексту[47] 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[11]

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[49] (енгл. *Conditional Random Field*, *CRF*), или комбинацијом са другим техникама као што су вештачке неуронске мреже[29] (енгл. *Artificial Neural Network*, *ANN*).

Глава 3

Моделовање

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 модела

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 опажања 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 новчића – 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 ,
- опажања, опсервација, емисија, приказа, исхода, симбола y_i ,
- 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[40].

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* _{l, k} , а матрицу b_{ij} као *emission* _{l} (b). Такође, потребу за скупом полазних вероватноћа – који је заправо опционалан, о чему

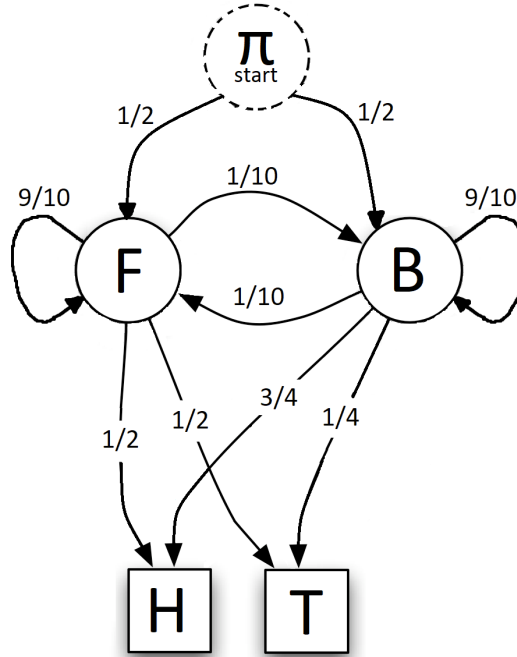
ће бити речи касније – уводе тек касније, па је *HMM* код њих у основи уређена четворка уместо петорка. Овде је ипак одлучено да се користи познатија нотација, како би читаоцима била лакша употреба повезане литературе. Штавише, *HMM* се у литератури често дефинише још простије, као уређена тројка $\{a, b, \pi\}$, 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 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 *НММ* увели су Ленард Баум и сарадници кроз низ статистичких радова објављених у другој половини шездесетих година двадесетог века[23]. У питању је надградња појма Марковљевих ланаца (енгл. *Markov Chain*, *MC*), који су у суштини *НММ* 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[9]. Прва практична примена *НММ* била је препознавање говора, док је биолошка примена почела 1986, Бишоповим и Томпсоновим поравнањем *DNA*[25].

3.2 Могућности модела

Могуће је дефинисати појам скривеног пута $p = p_1 \dots p_k$ као низ k стања кроз која *НММ* 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 *HMM* 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 *HMM* 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 *HMM* пролази кроз низ стања p , а да притом емитује управо секвенцу опсервација o . Према уобичајеним принципима, могуће је приметити следеће: $\sum_p \sum_o P\{p, o\} = 1$. Наиме, када се саберу вероватноће свих могућих комбинација низа опажања и скривених путева одређене дужине k , добија се јединица, што значи да је покривен цео простор догађаја у *HMM*. 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 0 није добра нити смислена. Зато је и уведен појам *HMM*.

Проблем 0: Непоштена коцкарница

На основу низа исхода бацања новчића, одредићи када крућице у непоштеној коцкарници користиле који од два моћућа новчића.

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

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

Добра формулација преко појма *HMM* дата је кроз проблем 1. Управо је

она детаљно обрађена у наставку овог поглавља, као његов централни део.

Проблем 1: Декодирање приказа

Пронаћи оптимални пут кроз НММ ако је емитована ниска o .

Улаз: ниска $o = o_1 \dots o_k$ и $\text{НММ}\{a, b, \pi\}$ 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 . Стога се формулише нови проблем 2.

Проблем 2: Вероватноћа пута и исхода

Израчунајти вероватноћу пута и одајања у 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.

Издаз: заједничка вероватноћа пута и исхода $P\{p, o\}$.

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 (такође и ниски симбола) дужине k у НММ са n могућих стања иначе је експоненцијалан зато што се одабир сваког 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[13]

Израчунајти вероватноћу скривеног пута p кроз НММ.

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

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:

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

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

Проблем 4: Вероватноћа исхода на путу[15]

Израчунајте вероватноћу приказа o на путу p кроз HMM .

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

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

Други елемент $P\{o|p\}$, дакле, представља вероватноћу да HMM емитије 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 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:

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

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 HMM и сродне појмове попут скривеног пута и низа опсервација:

- уводи се експлицитно почетно стање $x_0 = \pi$ уместо одвојених 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 HMM 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:

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

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

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

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 $x_{n+1} = \omega$, што је посебно важно за проблем декодирања,
- уместо нестабилних правих вероватноћа користе се логаритамске вероватноће, што ублажава рачунске грешке, мада усложњава алгоритме.

Пожељно је усвојити и 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 најјефикасније 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 *HMM* сада се може свести на једноставну уређену двојку:

- мапа логаритамских вероватноћа 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, *HMM* и даље званично сматра уређеном тројком $\{a, b, \pi\}$, а не интерно коришћеном трансформисаном двојком $\{a_{\log}, b_{\log}\}$. 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 .

Надграђени *HMM* моделује непоштену коцкарницу на следећи начин:

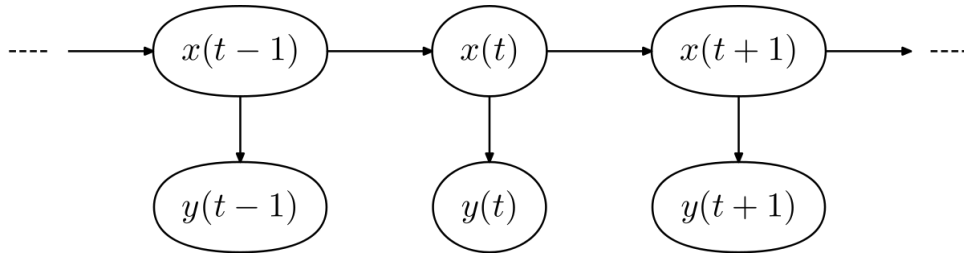
$$\bullet \text{ прелази } a_{\log} = \begin{matrix} & \begin{matrix} F & B \end{matrix} \\ \begin{matrix} \pi \\ F \\ B \end{matrix} & \begin{pmatrix} \log \frac{1}{2} & \log \frac{1}{2} \\ \log \frac{9}{10} & \log \frac{1}{10} \\ \log \frac{1}{10} & \log \frac{9}{10} \end{pmatrix} \end{matrix} - \text{нпр. } a_{\log,F,B} = P_{\log}\{F \mapsto B\},$$

$$\bullet \text{ емисије } b_{\log} = \begin{matrix} & \begin{matrix} H & T \end{matrix} \\ \begin{matrix} F \\ B \end{matrix} & \begin{pmatrix} \log \frac{1}{2} & \log \frac{1}{2} \\ \log \frac{3}{4} & \log \frac{1}{4} \end{pmatrix} \end{matrix} - \text{нпр. } b_{\log,B,H} = P_{\log}\{H|B\} = \log \frac{3}{4}.$$

3.4 Витербијев алгоритам

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 , израчунавања $P\{p, o\}$ за познати низ приказа o , и на крају одабира пута који представља $\arg\max_p P\{p, o\} = \arg\max_p P\{p|o\}$. Овиме се добро моделује условна расподела скривених путева при познатим опажањима. Логаритам је монотона трансформација, тако да се задатак максимизације не мења ни када се посматрају стабилније вредности $P_{\log}\{p, o\}$. 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[55], 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 временског својства сваког Марковљевог модела, а које је представљено на слици 3.2.



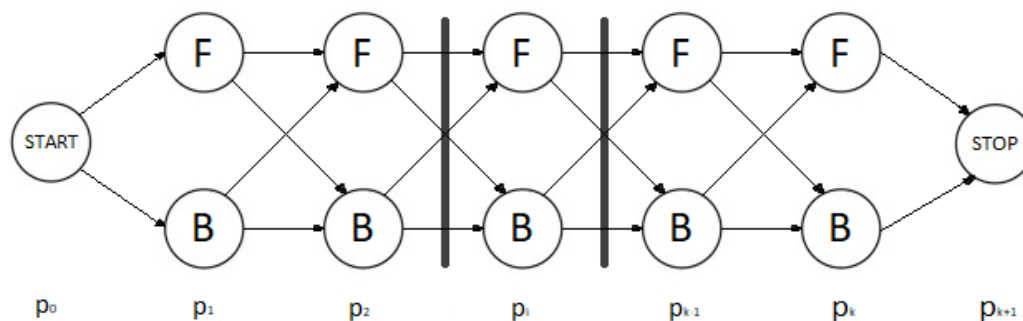
Слика 3.2: Ток времена код скривених Марковљевих модела[50]

Сваки *НММ*, наиме, моделује један Марковљев процес, што је поменуто при дефиницији. Последица је да тренутно стање зависи искључиво од претходног на путу и ниједног другог – мапа a моделује $p_{t-1} \mapsto p_t$, у ознакама са слике $x(t-1) \mapsto x(t)$. Исто тако, опсервација зависи искључиво од текућег стања – мапа b моделује $p_t \mapsto o_t$, у ознакама са слике $x(t) \mapsto y(t)$. Стога се

HMM понекад дефинише и нешто другачије, као уређени пар $\{X, Y\}$, где је X систем који се моделује, а Y процес чије понашање директно зависи од X .

Прецизније, X је Марковљев процес са неопсервабилним („скривеним“) стањима (x из дефиниције), а циљ модела је да се нешто о том процесу сазна на основу опажања (y из дефиниције) процеса Y , чије је понашање видљиво. Притом условна расподела Y (на слици конкретна вредност $y(t)$, а у низу опсервација приказ o_t) у неком временском тренутку t (индекс низа) зависи искључиво од стања X у том истом тренутку (на слици конкретна вредност $x(t)$, а на скривеном путу стање p_t). Приметно је да је ова дефиниција у суштини једнака претходно изложеним, с тим што је математички напреднија (захтевнија) – углавном је теже разумети торку апстрактних статистичких процеса него једноставних структура попут скупова, низова, матрица и мапа. На конкретном примеру непоштене коцкарнице, X је процес одабира (замене) новчића, а Y процес бацања новчића, односно добијања исхода тог бацања.

Све у свему, описано временско својство оправдава употребу Витербијевог графа, чији је пример за проблем непоштене коцкарнице дат на слици 3.3. Граф се састоји из мреже (матрице) чворова чија основа има n редова и k колона. Свака колона састоји се од низа чворова који представљају сва скривена стања у тренутку t . Из сваког чвора у колони $t - 1$ усмерена је по једна грана у сваки чвор из колоне t , на основу чињенице да се из сваког стања у тренутку $t - 1$ може прећи у било које стање у тренутку t . Поред ове основе, мрежа има и два посебна чвора – извор (експлицитно почетно стање) и понор (експлицитно завршно стање). Замисао овакве мреже је да истовремено моделује све скривене путеве дужине k кроз *HMM* са n скривених стања.



Слика 3.3: Витербијев граф непоштене коцкарнице

Стварно, различитих путева од извора до понора има тачно n^k , и сваки одговара једном скривеном путу у *НММ*. Остаје још питање како отежати гране Витербијевог графа, након чега се он може искористити за проблем максимизације кумулативне тежине у понору. То је заправо основни проблем над сваким Менхетном, који се може решити алгоритмима из петог поглавља.

За хватање интуиције у вези са моћи Витербијевог графа, није лоше увести проблем 5. Задатак је пронаћи највероватнији скривени пут дужине k .

Проблем 5: Највероватнији скривени пут

*Израчунајте највероватнији скривени пут p кроз *НММ*.*

Улаз: дужина k скривеног пута кроз $\text{НММ}\{a, b, \pi\}$.

Изаз: највероватнији скривени пут $p_{\text{opt}} = p_1 \dots p_k$.

Наивно решење проблема своди се на већ разматрану исцрпну претрагу простора скривених путева, којих је n^k . Вероватноћа сваког пута рачуна се у $O(k)$ корака, па је временска сложеност експоненцијална $O(n^k k)$. Ипак, могуће је искористити Витербијев граф како би се постигло знатно побољшање.

Нека је мрежа чворова представљена мапом s , таквом да $s_{x_i, t}$ складишти неки податак о чвору (стању) x_i у тренутку t . Оваква структура погодна је за свођење полазног проблема на проблем динамичког програмирања. Како је крајњи циљ максимизација вероватноће пута, нека $s_{x_i, t}$ заправо складишти вероватноћу оптималног пута дужине t који се завршава у стању x_i . Очигледно, за путеве јединичне дужине, односно у тренутку $t = 1$, важи:

$$s_{x_i, 1} = P\{x_i\} = \pi_{x_i} = a_{\pi, x_i}.$$

Испоставља се да се и остале тежине могу узети из матрице прелаза, што важи због темпоралног својства Марковљевих процеса. Како свако стање зависи искључиво од првог претходног, тако се и вероватноћа нејединичног пута може максимизовати тако што се размотре сва могућа претходна стања, односно за једно стање краћи путеви. Тако је рекурентна формула:

$$s_{x_i, t} = \max_j \{s_{x_j, t-1} \cdot a_{x_j, x_i}\},$$

$$P\{p_{\text{opt}}\} = \max_p P\{p\} = \max_j \{s_{x_j, k}\}.$$

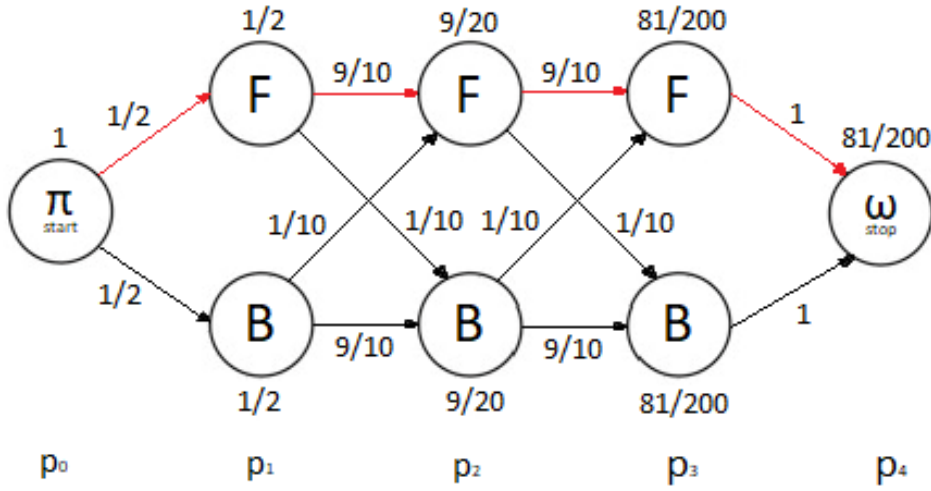
Наравно, проблеми са рачуном се решавају логаритамском трансформацијом:

$$s_{\log, x_i, 1} = P_{\log}\{x_i\} = \pi_{\log, x_i} = a_{\log, \pi, x_i},$$

$$s_{\log, x_i, t} = \max_j \{s_{\log, x_j, t-1} + a_{\log, x_j, x_i}\},$$

$$P_{\log}\{p_{opt}\} = \max_p P_{\log}\{p\} = \max_j \{s_{\log, x_j, k}\}.$$

Ова верзија је боља и због тога што су Менхетн алгоритми адитивни по природи, односно засновани су на сабирању, а не множењу вредности. Слика 3.4 приказује како Витербијев граф моделује три бацања у непоштену казину.



Слика 3.4: Максимизација $P\{p\}$ са три бацања

Општи облик ове рекурентне релације заправо је заснован на тежинама грана τ , где мапа облика $\tau_{x_i, x_j, t}$ означава тежину гране из чвора x_i ка x_j у тренутку t , укључујући експлицитни почетни извор π и завршни понор ω . Оне су у конкретном случају биле логаритми вероватноћа преласка или саме вероватноће, као на слици, у ком случају се множи уместо сабира:

$$s_{x_i, 1} = \tau_{\pi, x_i}, s_{x_i, t} = \max_j \{s_{x_j, t-1} + \tau_{x_j, x_i, t}\},$$

$$P_{opt} = \max_j \{s_{x_j, k} + \tau_{x_j, \omega}\}.$$

Како је моделовано $P\{p\}$, тако се може моделовати и $P\{p, o\}$ за фиксирано o . У првом случају, важило је $\tau_{x_i, x_j, t} = \tau_{x_i, x_j} = a_{x_i, x_j}$, док је у другом нешто сложеније $\tau_{x_i, x_j, t} = a_{x_i, x_j} \cdot b_{x_j, o_t}$, дакле вероватноћа догађаја да *НММ* пређе из стања x_i у стање x_j , након чега емитује симбол o_t . Формуле су сада:

$$s_{x_i, 1} = \pi_{x_i} \cdot b_{x_i, o_1} = a_{\pi, x_i} \cdot b_{x_i, o_1},$$

$$s_{x_i,t} = \max_j \{s_{x_j,t-1} \cdot a_{x_j,x_i} \cdot b_{x_i,o_t}\},$$

$$P\{p_{opt}, o\} = \max_p P\{p, o\} = \max_j \{s_{x_j,k}\}.$$

Једнаке су и логаритамске верзије, па се оне не наводе као можда сувишне.

Када је у питању проблем максимизације, могуће је моделовати и $P\{o|p\}$, с тим што за то није неопходан Витербијев граф. Поставка 6 је у наставку.

Проблем 6: Највероватније опсервације на путу

Израчунајти највероватнији низ емисија на путу p кроз НММ.

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

Излаз: највероватнија опажања $o_{opt} = o_1 \dots o_k$ на путу p .

Формула максималне вероватноће је једноставна, по свакој опсервацији:

$$P\{o_{opt}|p\} = \prod_{i=1}^k \max_j b_{p_i, y_j}.$$

Претходно изложени систем рада са *НММ*, заснован на Витербијевог графу и динамичком програмирању назива се Витербијев алгоритам[18], посебно када се примењује на декодирање – проблем 1. Изведеним рекурентним формулама једино треба додати систем путоказа, како би поред вероватноће оптималног пута могао бити добијен (реконструисан) и сам највероватнији пут.

Важна предност Витербијевог алгоритма је његова сложеност. Основна мрежа графа има nk чворова и $n^2(k-1)$ грана (из свих n стања ка свим n стањима у $k-1$ временском прелазу), чему се додају још два додатна чвора и $2n$ грана повезаних са тим чворовима. Израчунавање иде по чворовима, користећи гране, тако да је укупна временска и просторна сложеност $O(n^2k)$ уколико би се користио експлицитни граф. Ово је временски знатно боље од наивних $O(n^k)$, али је просторно захтевније, јер наивни приступ захтева само $O(k)$ помоћног простора. У питању је уобичајени компромис између времена и простора, када алгоритам за бржи рад захтева више меморије.

У многим случајевима је, међутим, граф довољно само замислити, а у раду користити искључиво мапу s и путоказе, а не и тежине τ , што за собом повлачи нешто бољу просторну сложеност $O(nk)$. Ово важи код декодирања, јер су тежине (гране) већ похрањене у мапама a и b . Други начин побољшања је ако се τ представи као функција уместо мапа, што је такође могуће код проблема декодирања, јер тежине не зависе од временског тренутка. Додатно,

уколико је довољно добити само максималну вероватноћу, а не и сам пут, не треба чувати путоказе, а мапу s могуће је свести на два низа који се наизменично попуњавају, чиме се просторна сложеност своди на $O(n)$.

У стварности је могуће добити још бољу сложеност. Наиме, многи *НММ* имају забрањене прелазе између неких стања. Таква ситуација веома је честа, а приказана је још на уводној слици 1.2. Могуће је без проблема уклонити гране Витербијевог графа које одговарају таквим прелазима, што знатно смањује време извршавања алгоритма. Посебно занимљиви могу бити недозвољени прелази који укључују извор и понор. На тај начин се може онемогућити да неко стање буде полазно или завршно, што често има биолошки смисао, о чему ће бити речи на познатом примеру профилних модела протеина.

3.5 Алгоритам „напред”

Сваки *НММ*, подсећања ради, може се схватити као уређени пар два процеса – скривеног Марковљевог који се читава скривеним путем p и опсервабилног зависног који се читава низом емисија o . Цела идеја *НММ* јесте детаљно статистички потковано моделовање тих процеса и њиховог односа.

Досад је било речи о појединачној расподели $P\{p\}$, условној $P\{o|p\}$ и заједничкој $P\{p, o\}$. Како је код последњег подразумевано да је позната ниска o , тиме је заправо моделована и условна расподела $P\{p|o\}$. Могуће је моделовати и појединачну расподелу $P\{o\}$, која је једина преостала како би модел био комплетиран. Основни задатак из овог домена дат је кроз проблем 7. Потребно је израчунати вероватноћу да *НММ* емитује неку ниску дужине k .

Проблем 7: Вероватноћа опсервација[14]

*Израчунајте вероватноћу приказа o у *НММ*.*

Улаз: низ опажања $o = o_1 \dots o_k$ у $\text{НММ}\{a, b, \pi\}$.

Изаз: вероватноћа улазног низа опажања $P\{o\}$.

Још једном, наивни приступ састоји се од генерисања свих n^k путева и сумирања вероватноћа на њима, према раније изложеној маргинализацији $P\{o\} = \sum_p P\{p, o\}$. Занимљиво је, међутим, приметити да је ова маргинализација врло слична садржају мапе s код Витербијевог алгоритма, која у понору израчунава $P\{p_{opt}, o\} = \max_p P\{p, o\}$. Једина разлика је у примење-

ном оператору – да ли је сума или максимум. Ово није случајно, јер је идеја обе формуле обилазак свих скривених путева кроз *НММ* истовремено.

Све у свему, сасвим је оправдано увести нову мапу f (од енгл. *forward* – напред), надахнуту претходном s (од енгл. *score* – скор), такву да елемент $f_{x_i,t}$ складишти вероватноћу префикса опажања дужине t (подниз $o_1...o_t$), насталог на скривеном путу који завршава стањем x_i . Одатле су формуле:

$$f_{x_i,1} = \pi_{x_i} \cdot b_{x_i,o_1} = a_{\pi,x_i} \cdot b_{x_i,o_1},$$

$$f_{x_i,t} = \sum_j f_{x_j,t-1} \cdot a_{x_j,x_i} \cdot b_{x_i,o_t},$$

$$P\{o\} = \sum_j f_{x_j,k}.$$

Као и досад, логаритамске верзије производе мењају збировима. Овога пута има и један додатак: сума се мења посебним оператором $\text{logsumexp}_j f(j)$, који моделује сабирање у логаритамском домену – апроксимира $\log \sum_j e^{f(j)}$.

У наставку, ваља поменути и сродан проблем одређивања највероватнијег исхода, односно $o_{opt} = \arg\max_o P\{o\}$. Формулација је дата проблемом 8.

Проблем 8: Највероватније опсервације

Израчунајте највероватнији низ емисија у НММ.

Улаз: дужина k пута кроз $\text{НММ}\{a, b, \pi\}$.

Излаз: највероватнија опажања $o_{opt} = o_1...o_k$.

И овде је наивно решење сувише неефикасно. Штавише, лошије је сложености од досадашњих $O(n^k k)$, јер је сада потребно генерисати и сваки могући низ опсервација. Сложеност исцрпне претраге стога је $O(n^k m^k k)$. Нешто боља сложеност добија се ако се не генеришу сви путеви, јер се свако од укупно m^k опажања може оценити већ изложеним алгоритмом „напред”. Тада је свеукупна сложеност $O(n^2 m^k k)$. Напредно решење може се конструисати помоћу тродимензионог Витербијевог графа (предлог аутора Певзнера), који применом оба оператора успешно максимизује суму $\max_o \sum_p P\{p, o\} = P\{o_{opt}\}$.

Алтернативни поглед на ствари подразумева остајање у дводимензионом простору – замисао је да из сваког од n стања (такође и почетног π) ка свим n стањима у $k - 1$ временском прелазу иде по m грана, по једна за сваку могућу опсервацију. Тиме гране Витербијевог (мулти)графа више не моделују само прелазе из једног стања у друго, већ успут и емисије. Тежине се

одабирају тако да осликавају вероватноћу промене стања, а затим емитовања симбола представљеног граном. Сваки пут од извора до понора сада није само скривени пут p , већ пут p (чворови) са придруженим опажањима o (гране). Максимални збир добија се максимизирањем сума између нивоа. Сложеност је у том случају $O(n^2mk)$, а формуле (логаритамске су аналогне):

$$f_{x_i,1} = \max_k \{\pi_{x_i} \cdot b_{x_i,y_k}\} = \max_k \{a_{\pi,x_i} \cdot b_{x_i,y_k}\},$$

$$f_{x_i,t} = \max_k \sum_j f_{x_j,t-1} \cdot a_{x_j,x_i} \cdot b_{x_i,y_k},$$

$$P\{o_{opt}\} = \max_o P\{o\} = \sum_j f_{x_j,k}.$$

Заменом суме максимумом у претходним формулама, добија се највероватнији пар скривеног пута дужине k и на њему емитоване ниске симбола. У питању је решење проблема 9, сличног досад разматранима.

Проблем 9: Највероватнији скривени пут и опсервације

Израчунајте највероватнији пут и опажања у НММ.

Улаз: дужина k пута кроз НММ $\{a, b, \pi\}$.

Излаз: највероватнија комбинација пута p и опажања o .

То је $\max P\{p, o\}$, што је исплативије од $n^k m^k$ или $n^2 m^k$ наивних покушаја (и овде су логаритамске верзије аналогне, па се не наводе као сувишне):

$$f_{x_i,1} = \max_k \{\pi_{x_i} \cdot b_{x_i,y_k}\} = \max_k \{a_{\pi,x_i} \cdot b_{x_i,y_k}\},$$

$$f_{x_i,t} = \max_{j,k} \{f_{x_j,t-1} \cdot a_{x_j,x_i} \cdot b_{x_i,y_k}\},$$

$$\max P\{p, o\} = \max_j \{f_{x_j,k}\}.$$

Комплетности ради, могу се формално представити и проблеми 10 и 11, којима се експлицитно израчунава $P\{p|o\}$ и $\max_p P\{p|o\}$ за познато o .

Проблем 10: Вероватноћа пута при исходу

Израчунајте вероватноћу пута p кроз НММ ако је опажено o .

Улаз: ниска $o = o_1 \dots o_k$ коју је емитовао НММ $\{a, b, \pi\}$ и скривени пут $p = p_1 \dots p_k$ кроз који је прошао.

Излаз: условна вероватноћа пута при приказу $P\{p|o\}$.

Сама вероватноћа пута ако је опажена нека секвенца емисија може се израчунати преко формуле условне вероватноће. Решење је, дакле:

$$P\{p|o\} = \frac{P\{p, o\}}{P\{o\}}.$$

Главнина оваквог приступа је израчунавање вероватноће исхода, тако да је сложеност $O(n^2k)$. Наивни приступ би, као и досад, узео $O(n^k k)$ времена.

Проблем 11: Највероватнији пут при исходу

Израчунајте највероватнији пут p кроз НММ ако је опажено o .

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

Изаз: највероватнију пут $p_{opt} = p_1 \dots p_k$ ако је опажено o .

Максимизација је још једноставнија, када се примети раније поменуто $\arg\max_p P\{p, o\} = \arg\max_p P\{p|o\}$ за познато o . Ово значи да је довољно искористити решење проблема 1, уз прикладно скалирање вероватноће.

Сада је познато како Витербијевим графом моделовати и максимизовати сваку од критичних вероватноћа $P\{p\}$, $P\{o\}$, $P\{p, o\}$, $P\{p|o\}$ и $P\{o|p\}$, чиме је модел комплетиран, барем што се тиче његове описне стране (остаје учење). Досадашња постигнућа модела сумирана су табелом 3.1, која следи.

Табела 3.1: Могућности скривених Марковљевих модела

Број	Проблем – алгоритам			Сложеност	
	Улаз	Циљ	Вредност	Наивни	Напредни
1	o_k	$(\arg)\max_p$	$P\{p, o\}$	$O(n^k k)$	$O(n^2 k)$
2	p_k, o_k	—	$P\{p, o\}$	$O(k)$	—
3	p_k	—	$P\{p\}$	$O(k)$	—
4	p_k, o_k	—	$P\{o p\}$	$O(k)$	—
5	k	$(\arg)\max_p$	$P\{p\}$	$O(n^k k)$	$O(n^2 k)$
6	p_k	$(\arg)\max_o$	$P\{o p\}$	$O(m^k k)$	$O(k)$
7	o_k	—	$P\{o\}$	$O(n^k k)$	$O(n^2 k)$
8	k	$(\arg)\max_o$	$P\{o\}$	$O(n^k m^k k)$ $O(n^2 m^k k)$	$O(n^2 m k)$
9	k	$(\arg)\max_{p,o}$	$P\{p, o\}$	$O(n^k m^k k)$ $O(n^2 m^k k)$	$O(n^2 m k)$
10	p_k, o_k	—	$P\{p o\}$	$O(n^k k)$	$O(n^2 k)$
11	o_k	$(\arg)\max_p$	$P\{p o\}$	$O(n^{2k} k)$ $O(n^k k)$	$O(n^2 k)$

Глава 4

Биолошки значај

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 Гени — два стања

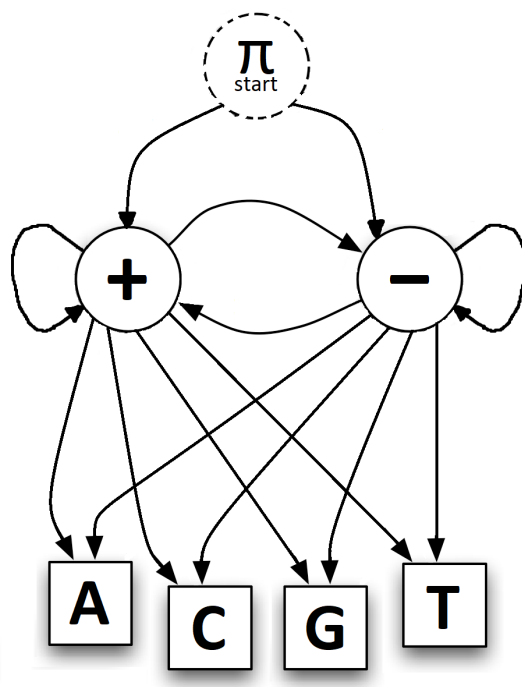
У уводном делу, посвећеном мотивацији, дискутовано је о проналажењу места на којима се гени налазе, односно где њихово преписивање (транскрипција) започиње. Објашњено је зашто је удео динуклеотида *CG* мали у некодирајућим регионима *DNA*, а нешто већи у кодирајућим, те како има смисла ту чињеницу искористити за откривање такозваних *CG* острва (*CpG* места), што су региони богати генима. Имплементиран је и наиван приступ решавању овог проблема, заснован на клизећем прозору, али су уз њега остале неразјашњене важне недоумице: како одредити добру величину прозора, као и шта тачно радити када преклапајући прозори нуде различиту класификацију подниза.

Сада је циљ доћи до прецизног, једнозначног и статистички поткованог решења употребом одговарајућег скривеног Марковљевог модела. Замисао

је у суштини једноставна – улазни низ нуклеотида посматра се као секвенца опажања коју треба декодирати. Другим речима, за сваки карактер ниске са улаза потребно је одредити да ли је вероватније настао као емисија CG острва или не, што је заправо позадински скривени процес. Стога важи следеће:

- скривена стања $x = \{+, -\}$ – јесте CG острво или није,
- опсервације $y = \{A, C, G, T\}$ – азбука DNA нуклеотида.

Скупови скривених стања и могућих опажања се, дакле, лако одређују, па чак и веома личе на разматрани мотивациони проблем непоштене коцкарнице. Наиме, такође су присутна два стања, мада опсервација има нешто више. Све у свему, проблем се може апстраховати неком врстом коцкарнице, у којој крупније мења две различито отежане четворостране коцкице. Уопштени дијаграм (без вероватноћа) оваквог модела приказан је на слици 4.1.



Слика 4.1: Скривени модел CG острва са два стања

Остаје још одредити све битне вероватноће. За овај део задатка погодује применити прави биоинформатички приступ. Подсећања ради, биоинформатика је у уводу дефинисана као интердисциплинарна област која се бави применом рачунарских технологија у области биологије и сродних наука, са

нагласком на разумевању биолошких података. Наведено је да статистички (математички) апарат 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. Управо је то овде и примењено – статистика (математика) дефинише појам *HMM*, а рачунарске технологије (конкретно *Python* и *Jupyter*) ефикасно га имплементирају.

Потребно је још консултовати се са биологијом, а овде заправо и генетиком, како би се адекватно одредили параметри модела. За почетак, треба приметити да се може добити фактички било какав исечак *DNA* секвенце. Другим речима, не постоји гаранција да ће почетни регион бити кодирајући или не, тако да је најсигурније равномерно расподелити почетна стања:

- полазне вероватноће $\pi = \begin{matrix} + \\ - \end{matrix} \begin{pmatrix} 0,5 \\ 0,5 \end{pmatrix}$ – равномеран почетак.

Даље, питање је колико често долази до промене стања. Одговор је да се то дешава веома ретко, с тим што мањи део секвенце представља *CG* острво, тако да је нешто већа шанса да дође до напуштања *CG* острва, него уласка у њега. Свеукупно, могла би се одабрати мапа прелаза попут следеће:

- прелази $a = \begin{matrix} + & - \\ + & - \end{matrix} \begin{pmatrix} 0,98 & 0,02 \\ 0,01 & 0,99 \end{pmatrix}$ – мала могућност промене.

За крај, остаје најтеже питање: како моделовати вероватноће опажања. Постоје разни приступи, а један од њих заснован је на емпиријским подацима. Примера ради, уколико се трага за генима у људском *X* хромозому, могу се апроксимирати вероватноће нуклеотида на основу вероватноћа динуклеотида из таблице 2.1. Резултат тога је следећа мапа вероватноћа опсервација:

- емисије $b = \begin{matrix} & A & C & G & T \\ + & - \end{matrix} \begin{pmatrix} 0,222 & 0,2555 & 0,299 & 0,2235 \\ 0,274 & 0,227 & 0,2295 & 0,2695 \end{pmatrix}$ – емпиријски.

Очекивано, нешто је већи удео цитозина и гуанина у кодирајућим регионима. Важи и супротно – нешто је већи удео аденина и тимина у некодирајућим регионима *X* хромозома, што је такође очекивана повезана појава.

Овакав модел, међутим, није успешан јер су вероватноће тако постављене да није могуће препознати мала *CG* острва. Приликом декодирања, закључак

ће за сваку малу секвенцу бити да је највероватније цела (не)кодирајућа, из једноставног разлога што је свака промена стања веома скупа, а удео нуклеотида није толико различит. Стога се може добити побољшање уколико се вероватноће прелаза мало приближе, а емисија мало више удаље.

То се може учинити тако што се, за почетак, вероватноће промене стања поставе на нешто већу једну десетину. Ако је *НММ* у некодирајућем стању, може се претпоставити да је расподела нуклеотида равномерна – сваки се емитује са могућношћу једне четвртине. У супротном, сматра се да се цитозин и гуанин приказују четири пута чешће. Резултујуће вероватноће сада су:

- прелази $a = \begin{matrix} + & - \\ + & (0,9 & 0,1) \\ - & (0,1 & 0,9) \end{matrix}$ – већа могућност промене,
- емисије $b = \begin{matrix} & A & C & G & T \\ + & (0,1 & 0,4 & 0,4 & 0,1) \\ - & (0,25 & 0,25 & 0,25 & 0,25) \end{matrix}$ – поправљено.

Нажалост, ни овај модел није ништа бољи. Параметри су, иначе, преузети са примера употребе библиотеке *pomegranate*[51], па је добар тренутак за кратку дигресију о њој. У питању је модул програмског језика *Python* који омогућује рад са многим пробабилистичким моделима, што поред скривених Марковљевих модела укључује и Марковљеве ланце, Бајесове и Марковљеве мреже (случајна поља), графове фактора и уопштене мешовите моделе.

Када је у питању рад са *НММ*, од досад обрађених проблема решени су само декодирање (Витербијев алгоритам) и вероватноћа опажања (алгоритам „напред”), те је модул с те стране минималистички. Ипак, добра страна је што су имплементирана сва изнесена проширења дефиниције (логаритамске вероватноће, непрекидне расподеле, експлицитно почетно и завршно стање итд.), а омогућено је и нешто прилично оригинално – итеративно моделовање. Наиме, модел се у *pomegranate* не прави прослеђивањем готових низова, матрица или мапа, већ део по део, тако што се прво направе жељене расподеле емисија, затим стања са тим расподелама, затим прелази између стања, након чега се финализује топологија модела. Имплементирано је и учење модела, о чему ће у раду бити касније речи. У коду електронског уџбеника решена је потрага за генима и помоћу овог модула, а резултати су, наравно, једнаки.

Још један познати модул језика *Python* за рад са *HMM* јесте *hmmlearn*[45]. Имплементира дискретне (мултиномијалне), Гаусове (расподеле емисија су нормалне) и мешовите (емисије потичу из мешавине нормалних расподела) скривене Марковљеве моделе. Попут претходне, и ова библиотека имплементира само Витербијев и алгоритам „напред”, као решења најважнијих проблема код *HMM*. Такође подржава учење параметара модела. С друге стране, ради искључиво са матрицама параметара, па тако мултиномијални модел сасвим одговара основној дефиницији *HMM*, без икаквих надградњи. Чак су стања и емисије у потпуности апстраховани индексима. Стога је једноставна за рад и брзо добијање резултата. И она је у коду електронског уџбеника примењена у потрази за генима, још једном са истим резултатима.

Није лош тренутак да се помене досад занемарена чињеница да су *HMM* генеративни модели. Не само што служе за опис појава које моделују, већ могу у потпуности да их опонашају. Оваква могућност је важна карактеристика сваког модела који је има. Већ је разматрано како се могу добити скривени путеви, опажања или комбинација који максимизују неку вероватноћу. Сада се испоставља да је још једноставније могуће добити произвољан скривени пут и опсервацију на њему. Скривени пут жељене дужине генерише се на основу почетних вероватноћа и мапе преласка, а исход на путу помоћу мапе емисија. Претходни модел за откривање гена тако се може искористити за генерисање вештачке *DNA* секвенце, која задовољава тај статистички опис *DNA*. Она, наиме, у појединим деловима садржи *CG* острва и тиме се чини природнијом од случајно генерисане, што је додатна корист од *HMM*.

Што се тиче самог модела, остаје проблем што су досадашња оба покушаја била неуспела. Трећа идеја могло би бити додатно повећавање вероватноће промене стања. Мала повећања не би променила резултат, док би већа изврнула смисао *CrG* места – острвом би се прогласио сваки цитозин и гуанин. Ово није необично и иде уз чињеницу да ова употреба *HMM* спада под домен ненадгледаног учења, где модел по самосталној процени групише поднизове улазне секвенце. То значи да се лако могу добити неочекивани или незадовољавајући резултати, попут једноставне поделе према текућем карактеру.

Следећа идеја настоји да ово превазиђе тако што укључује већи број опажања. Наиме, могуће је *DNA* секвенцу схватити као низ динуклеотида уместо самих нуклеотида, као код прозорског приступа. Сада важи следеће:

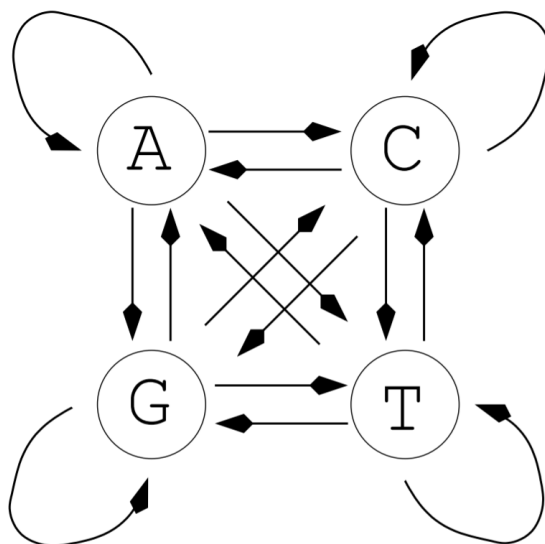
- опсервације $y = \{A, C, G, T\} \times \{A, C, G, T\}$ – азбука динуклеотида.

Може се дорадити и мапа прелаза, док се мапа емисија узима према табели 2.1, која управо непосредно табелира вероватноће динуклеотида. Овај приступ даје засад најбоље резултате, који се могу видети у коду лекције.

4.2 Гени – више стања

Како је већ напоменуто, сви досадашњи модели су по структури подсећали на непоштени казино – имали су два стања која се ретко мењају и углавном четири опсервације, мада је најбољи резултат добијен при последњем покушају, са чак шеснаест различитих динуклеотидних емисија. Алтернативна идеја увођењу већег броја исхода јесте увођење већег броја стања, што се најчешће реализује кроз два приступа, који су представљени у наставку.

Први приступ је врло популаран и основни је пример на многим универзитетским курсевима који обрађују скривене Марковљеве моделе[37, 38, 41, 52]. Разматран је ранијих година и на вежбама из Увода у биоинформатику, у оквиру којег је овај рад настао. Полазна идеја је да се *CG* острва и региони ван њих могу моделовати као два одвојена Марковљева ланца (подсећања ради, у питању су *HMM* без емисија, а скраћено се називају *MC*). Стања ланца су јавна (то јест, нису скривена), пошто верно прате *DNA* секвенцу коју моделују, и одговарају азбуци нуклеотида, па се могу представити сликом 4.2.



Слика 4.2: Марковљев ланац за моделовање *DNA* секвенце[38]

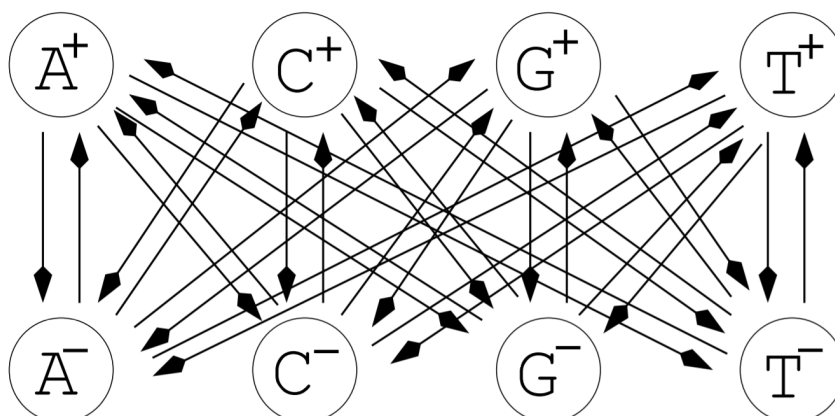
Одговарајуће матрице прелаза могу се одредити емпиријским путем, дакле обрадом секвенци за које је познато јесу ли CG острва или не. Уобичајено се узимају вредности из табеле 4.1, које су унапред припремљене (израчунате).

Табела 4.1: Вероватноћа прелаза између нуклеотида једне секвенце – лево у регионима CG острва, а десно ван њих[52]

	A	C	G	T	A	C	G	T
A	0,180	0,274	0,426	0,120	0,300	0,205	0,285	0,210
C	0,171	0,367	0,274	0,188	0,322	0,298	0,078	0,302
G	0,161	0,339	0,375	0,125	0,248	0,246	0,298	0,208
T	0,079	0,355	0,384	0,182	0,177	0,239	0,292	0,292

Аналогно прозорском приступу заснованом на динуклеотидном садржају секвенце и табели 2.1, могуће је помоћу MC за сваки подниз одредити да ли је већа вероватноћа да јесте CG острво или да није. Бројчана сагласност се за оба MC може израчунати већ имплементираним алгоритмом за одређивање вероватноће пута кроз HMM , као решење проблема 3. Одабир припадности пада на ланац са већом вероватноћом. Резултати су једнаки као у првом покушају, а остају нерешени проблеми прозорског приступа: како одредити добру величину прозора и како разрешити сукобе настале преклапањем прозора.

Као решење, предлаже се спајање ова два ланца у један. Резултујући MC дат је на слици 4.3. Он сада има осам стања, за сваки пар нуклеотида и припадности CG острву. Једноставности ради, пошто укупно има $8^2 = 64$ прелаза, приказани су само нови, док се стари (слика 4.2) подразумевају.



Слика 4.3: Спојени ланци за моделовање CG острва[38]

Пре свега, неопходно је одредити нову, заједничку матрицу преласка. За то се треба подсетити већ поменутог доменског биолошког знања, према коме је мало вероватан прелазак из кодирајућег у не кодирајуће стање (нпр. само 2 %), а још мање вероватно обрнуто (нпр. тек 1 %). Нове вероватноће могу се добити скалирањем старих, нпр. на начин представљен табелом 4.2.

Табела 4.2: Вероватноћа прелаза унутар CG острва

	A^+	C^+	G^+	T^+
A^+	0,180p	0,274p	0,426p	0,120p
C^+	0,171p	0,367p	0,274p	0,188p
G^+	0,161p	0,339p	0,375p	0,125p
T^+	0,079p	0,355p	0,384p	0,182p
	A^-	C^-	G^-	T^-
A^+	0,180(1-p)	0,274(1-p)	0,426(1-p)	0,120(1-p)
C^+	0,171(1-p)	0,367(1-p)	0,274(1-p)	0,188(1-p)
G^+	0,161(1-p)	0,339(1-p)	0,375(1-p)	0,125(1-p)
T^+	0,079(1-p)	0,355(1-p)	0,384(1-p)	0,182(1-p)

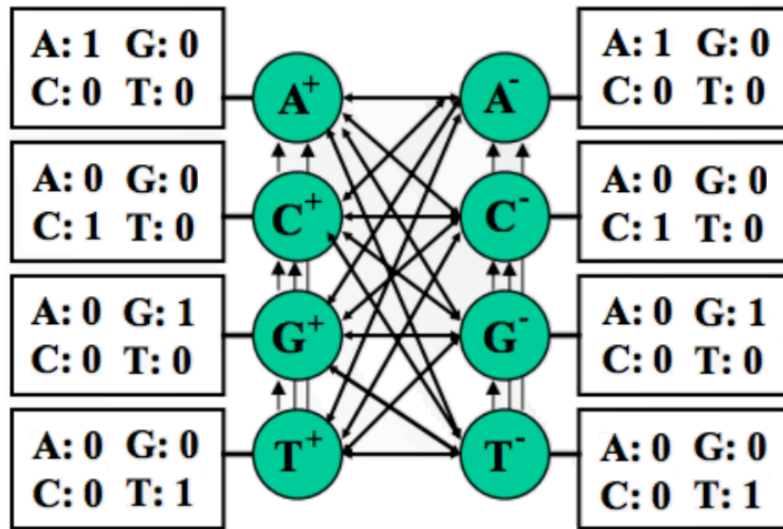
Идеја је, дакле, расподелити вероватноће претходних прелаза унутар CG острва типа $P\{N \mapsto M\}$ на две нове типа $P\{N^+ \mapsto M^+\} = pP\{N \mapsto M\}$ и $P\{N^+ \mapsto M^-\} = (1 - p)P\{N \mapsto M\}$, где p представља вероватноћу останка унутар острва, нпр. $p = 98\%$. Аналогно томе, вероватноће претходних прелаза ван CG острва типа $P\{N \mapsto M\}$ могу се расподелити на две нове типа $P\{N^- \mapsto M^-\} = qP\{N \mapsto M\}$ и $P\{N^- \mapsto M^+\} = (1 - q)P\{N \mapsto M\}$, где q представља вероватноћу останка ван острва, нпр. $q = 99\%$. Алтернативна је равномерно расподелити вероватноће промене стања, као у табели 4.3.

Табела 4.3: Вероватноћа прелаза ван CG острва

	A^-	C^-	G^-	T^-
A^-	0,300q	0,205q	0,285q	0,210q
C^-	0,322q	0,298q	0,078q	0,302q
G^-	0,248q	0,246q	0,298q	0,208q
T^-	0,177q	0,239q	0,292q	0,292q
	A^+	C^+	G^+	T^+
A^-	(1-q)/4	(1-q)/4	(1-q)/4	(1-q)/4
C^-	(1-q)/4	(1-q)/4	(1-q)/4	(1-q)/4
G^-	(1-q)/4	(1-q)/4	(1-q)/4	(1-q)/4
T^-	(1-q)/4	(1-q)/4	(1-q)/4	(1-q)/4

Сада се спојени Марковљев ланац може и мора надградити у скривени Марковљев модел. Наиме, два одвојена MC имала су смисла без унапређења, јер су имала само четири суштински јавна стања, која одговарају нуклеотидима из опажене секвенце. Нових осам стања је, међутим, по природи скривено, јер опажањем секвенце није познато у ком је модел стању, као у непоштеној коцкарници и другим моделима. Примера ради, већ за једночлани исход G није јасно да ли је настао у стању G^+ или G^- . Ниска дужине k може настати на 2^k различитих непознатих путева, који су стога скривени.

Већ је имплицирано да симбол N може приказати само стање типа N^+ и N^- . Важи и обрнуто, јер новоуведено стање N^σ управо и означава појаву симбола N у старом стању σ . Ово се може схватити и као својеврсни образац пројектовања (узорак, шаблон) када је у питању рад са MC и HMM . Свеукупна последица је да је мапа емисија врло једноставна (такорећи дегенерисана) – свако стање са јединичном вероватноћом емитује одговарајући симбол. Слика 4.4 приказује коначан дијаграм овог модела, са изостављеним многобројним вероватноћама прелаза и наглашеним вероватноћама емисија.



Слика 4.4: Нова структура модела CG острва, према [41]

Што се тиче почетних вероватноћа, могу се узети емпиријске вредности засноване на узорку или пак равномерне могућности $1/8$. Кад се све сабере:

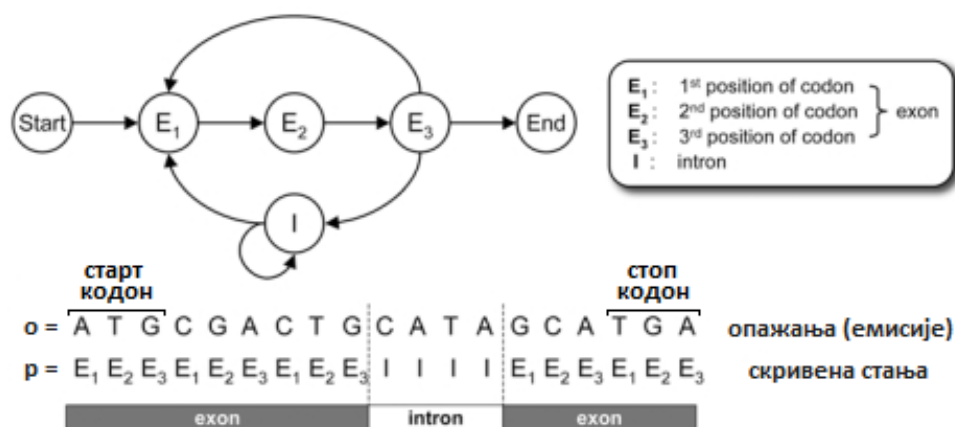
- опсервације $y = \{A, C, G, T\}$ – азбука DNA нуклеотида,
- скривена стања $x = y \times \{+, -\}$ – Декартов производ симбола,

- полазне вероватноће $\pi = \frac{1}{8}$ или емпиријске, према узорку,
- прелази $a =$ припремљене вредности према табели 4.2 или 4.3,
- емисије $b = 1$ ако стање одговара, иначе 0, према слици 4.4.

И овакав модел даје врло добре резултате. Други приступ потрази за генима који укључује већи број стања и није заснован на проналажењу *CG* острва. Алтернативна идеја заправо моделује сложенију структуру еукариотске *DNA*, с циљем да још детаљније и прецизније аотира улазне секвенце.

Иако *CG* острва јесу добар показатељ да се у близини налази неки промотер, који би могао да покрене преписивање (транскрипцију) гена, још би боље било када би се могло тачно одредити који нуклеотиди представљају ген, а који не. Познато је да се *DNA* може поделити на више поднизова који имају двојаку природу – или су интрони или егзони[43, 58]. Интрони су интрагенски региони, па тако представљају некодирајуће делове наследног материјала, који су уметнути између гена и који се уклањају у процесу сплајсовања. Егзони (ексони, због експресије), с друге стране, кодирају протеине, и увек су дужине дељиве са бројем три. Они се заправо састоје из триплета нуклеотида (кодона) који појединачно кодирају аминокиселине, које касније граде протеине. Постоји и неколико специјалних кодона – почетни *ATG* и завршни *TAA*, *TAG*, *TGA* – који означавају места на којима преписивање почиње и завршава се, мада стартни кодон на другим местима кодира метионин.

Последица је да се *DNA* може моделовати и аутоматом са слике 4.5. На слици су представљени пример, стања и прелази, са већим бројем забрањених.



Слика 4.5: Структура модела еукариотске *DNA*, према [57]

Ово је, међутим, врло упрошћен модел, а у стварности је организација *DNA* знатно комплекснија – посебну структуру имају подланци на почетку и крају *DNA* ланца, посебан удео нуклеотида имају делови на прелазу између егзона и интрона, а посебно се издвајају и такозвани *ORF*-ови, који су целом дужином кодирајући, без интронских прекида[36, 38]. Стога је очекивано да успешан модел ипак мора имати већи број стања, што и јесте случај.

Један од успешних модела за тачно предвиђање гена који кодирају протеине јесте *GENSCAN*, алат који су 1997. године осмислили Берц и Карлин[7, 26]. Модел је карактеристичан по томе што предвиђа гене на оба ланца *DNA* истовремено, па тако за већину елемената секвенце има дуплирана стања. Примера ради, делове егзона не моделује кроз три стања, као на слици 4.5, већ кроз шест. Укупно има 27 стања. Надограђен је појмом трајања (ново темпорално својство), по чему је такође карактеристичан, тако да није у питању сасвим обичан *HMM*. Принцип рада је, међутим, исти: улаз је *DNA* секвенца, а излаз декодирана стања, израчуната Витербијевим алгоритмом. Алат је прилично успешан, са стопом погодака од преко 90 % по нуклеотиду и око 80 % по егзону, о чему се детаљније може прочитати у цитираном раду.

За крај, није лоше сумирати успех *HMM* у потрази за генима. Када су у питању модели са само два стања (јесте или није *CG* острво), проблематично је уколико се постави мала вероватноћа промене стања. Такав модел мале секвенце по правилу проглашава за целе (не)кодирајуће. С друге стране, повећањем вероватноће прелаза долази до извртања идеје, и тада само текући карактер постаје битан. Знатно побољшање добија се посматрањем динуклеотидног састава ниске, уместо скенирањем једног по једног карактера. Једнако добро се понаша модел са више стања, иако су му емисије дегенерисане.

Од свих разматраних модела је, међутим, најбоља надградња *HMM* реализована кроз сложени алат *GENSCAN*. Она можда не проналази *CG* острва као таква, али зато прецизно лоцира кодирајуће егзоне. Неки општи закључак могао би бити да се боље показују модели са више стања, који конкретније хватају зависности. Штавише, није лоше напоменути да овај проблем спада у оне поменуте у мотивационом уводу, који се ефикасније могу решити помоћу многобројних надградњи *HMM*. Једна од општијих успешних дорада јесу условна случајна поља, која умањују број погрешних предвиђања[31, 32].

4.3 Профилни модели

Крунско и можда најпознатије постигнуће скривених Марковљевих модела управо је њихова употреба у статистички поткованој класификацији секвенцијалних података. Конкретно, у наставку је размотрена примена у класификацији протеина, мада је објашњено и како се резултат уопштава.

Протеини су, наиме, организовани у разнолике протеинске фамилије, а чест биолошки задатак јесте додељивање новооткривеног полипептида некој од познатих фамилија. Замисао је да се на неки начин оцени припадност новог аминокиселинског ланца неким познатим породицама, а затим протеин додели оној са највећим скором. Наивни приступ оцењивању подразумева поређење улазног полипептида са сваким чланом породице појединачно, те напоследку сабирање тако добијених скорова или пак узимање максималног.

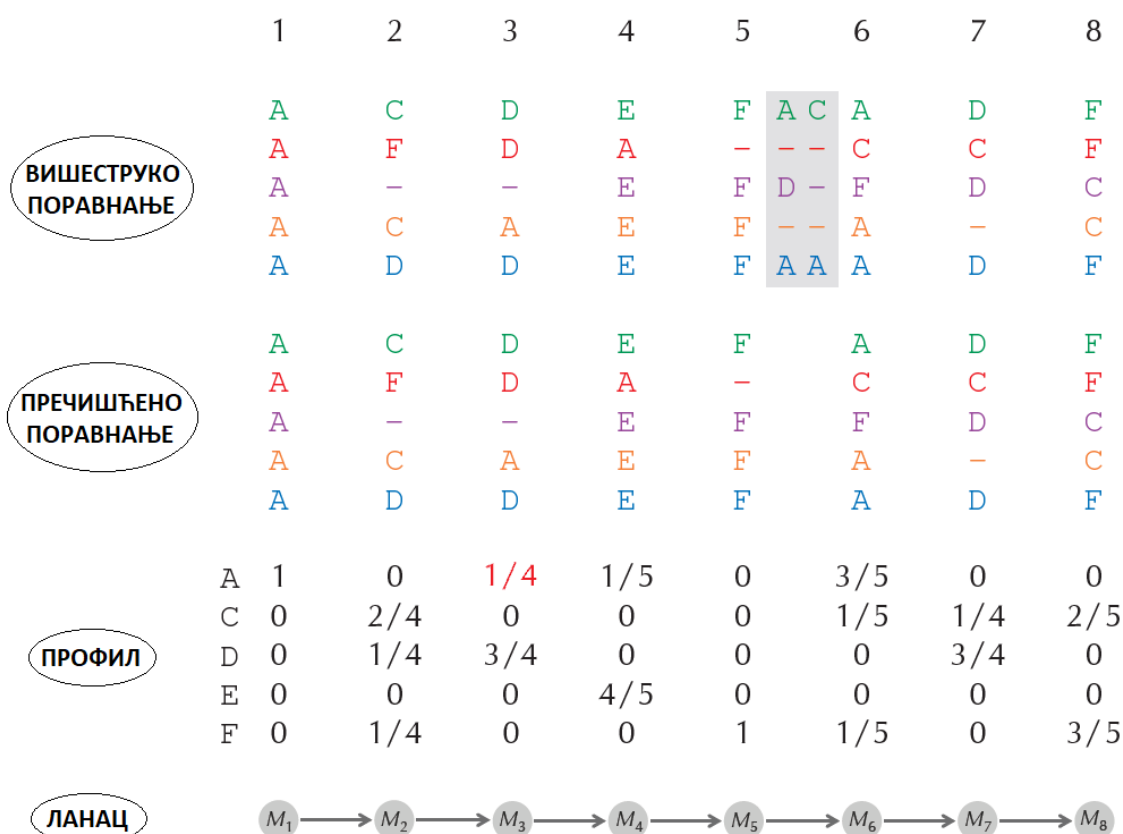
Протеини се у биоинформатици представљају својом примарном структуром, као ниска аминокиселина (азбука од двадесетак слова), па се међусобно могу лако поредити неким алгоритмом за рад са нискама. Често се, међутим, дешава да између појединих чланова фамилије постоје веће разлике, као нпр. код изузетно варијабилног гликопротеина *gp120* код *HIV*-а, како је размотрено у мотивационом уводу. То резултује нестварно малим скоровима, па стога поређење по паровима у општем случају не даје добре резултате.

Последица је да се протеин који се класификује мора поредити са целом фамилијом одједном. За те потребе, фамилије се најчешће представљају као вишеструка поравнања и изведене профилне матрице, како је представљено у другом (*Chapter 2: Which DNA Patterns Play the Role of Molecular Clocks? – Randomized Algorithms*) и петом (*Chapter 5: How Do We Compare DNA Sequences? – Dynamic Programming*) поглављу проучаваног уџбеника.

Подсећања ради, вишеструко поравнање је матрица карактера (низ ниски), чији редови представљају ниске које се поравнавају, а колоне карактере тих ниски на позицији одређеној том колоном. Како су при поравнању дозвољене инсерције (убацивање) и делеције (брисање слова), у нискама се налази и специјални карактер „-“, који означава празнину. Није необично да постоје ретке колоне, у којима је велики удео празнина. Биолошки гледано, та аминокиселина вероватно није битна карактеристика породице која се моделује, па се занемарује. Прецизније, уклањају се све колоне у којима је удео празнина већи од унапред одређене границе θ . Резултат је пречишћено поравнање.

Напослетку се пречишћено поравнање трансформише у профилну матрицу, чији редови представљају (све) карактере из азбуке поравнатих ниски, док колоне складиште удео сваког карактера на тој позицији, не рачунајући празнине. Приметно је да профил, како говори о вероватносној расподели по колони (позицији) карактера из азбуке y величине m , веома личи на мапу емисија неког *НММ*. Испоставља се да се стварно може тако посматрати.

Најједноставнији *НММ* за моделовање фамилија протеина могао би бити у суштини дегенерисан ланац скривених стања x , такав да свако стање представља једну позицију, којих је укупно n . Од сваког стања x_i постоји само један прелаз са јединичном вероватноћом на стање x_{i+1} , док се вероватноће емисија узимају из профила. Прво стање x_1 обавезно је почетно, док је последње x_n обавезно завршно. На слици 4.6 приказан је претходно изнети ток догађаја од вишеструког поравнања до ланчаног *НММ*-а. Полазно има десет позиција, које се границом одсецања $\theta = 0,35$ свде на коначних осам.



Слика 4.6: Мотивациони пример *НММ* профила, према [30]

Занимљиво је напоменути да се чини да је ово поглавље у књизи написано у журбано, па су се провукле чак три материјалне грешке. Прва се налази на уџбеничкој верзији претходне слике, где је расподела у трећој колони $\{0, 0, 3/4, 0, 0\}$. Фали, дакле, вероватноћа приказа A , која износи $1/4$, и која је на овдашњој слици додата и наглашена црвеном бојом. Грешка се налази и на Певзнеровој презентацији, а није познато да ли је исправљена у најновијем издању. О осталим пропустима биће речи у наставку, када се стигне до њих.

Што се тиче малопре приказаног ланчаног HMM -а, треба напоменути да му се стања најчешће означавају као M_i , како је и на слици, а не као x_i , како је уобичајено код HMM . Разлог томе је што она заправо представљају поклапања (енгл. *match*) на тој позицији. Напоменуто је већ и да је овакав модел дегенерисан. Како има обавезно почетно и завршно стање, као и обавезне прелазе, кроз њега постоји само један скривени пут – тачно $M_1 M_2 \dots M_{n-1} M_n$.

Остаје још питање употребне вредности оваквог модела, односно оваквих модела, јер би постојао по један HMM за сваку породицу. Сада би се скор могао добити као вероватноћа емитовања ниске (новог полипептида) o у осмишљеном моделу. И овога пута, протеин би био додељен оној породици са највећим $P\{o\}$. Одређивање те вероватноће већ је разматрано као решење проблема 7. Ипак, како је модел дегенерисан, нема потребе примењивати сложени алгоритам „напред”, па чак ни рачунати једнаку вероватноћу при обавезном путу $P\{o|M_1 \dots M_n\}$, као решење проблема 4. Довољно је само помножити одговарајуће вредности из профилне матрице. Примера ради, вероватноћа да HMM са слике 4.6 емитује $ADDAFFDF$ износи (слика 4.7):

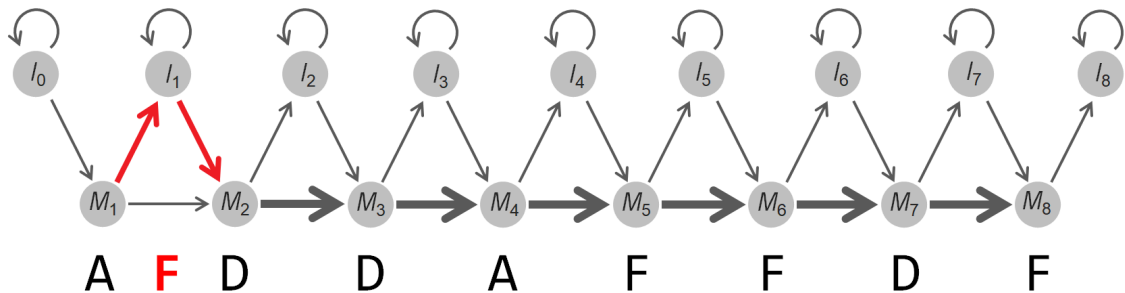
$$P\{ADDAFFDF\} = 1 \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{5} \cdot 1 \cdot \frac{1}{5} \cdot \frac{3}{4} \cdot \frac{3}{5} = 0,003375.$$

ПРОФИЛ	A	1	0	1/4	1/5	0	3/5	0	0
	C	0	2/4	0	0	0	1/5	1/4	2/5
	D	0	1/4	3/4	0	0	0	3/4	0
	E	0	0	0	4/5	0	0	0	0
	F	0	1/4	0	0	1	1/5	0	3/5
	F	0	1/4	0	0	1	1/5	0	3/5
ОПАЖАЊЕ	A	D	D	A	F	F	D	F	

Слика 4.7: Вероватноћа опсервације као скор фамилије, према [30]

Свеукупно, овакав модел није лош утолико што добро осликава сличност протеина са породицом – што је улазни полипептид сличнији фамилији, то је његов скор (вероватноћа) већи. Такође, различито оцењује сваку појединачну колону, што је циљ постављен у мотивационом уводу. Ипак, он због дегенерисаности баш и није прави *HMM*. Штавише, није ништа бољи од саме профилне матрице. Иако све лепо ради за опсервације (полипептиде) дужине $k = n$, главно ограничење је што се лоше моделују исходи других дужина. Њихова вероватноћа је подразумевано нулта, тј. сматрају се немогућим. Овај наивни модел, ипак, није толико бескористан, јер се може употребити као основа за прављење напредног, свеобухватног *HMM*-а за класификацију свакојаких секвенцијалних података, а не само протеина, о чему ће бити речи у наставку.

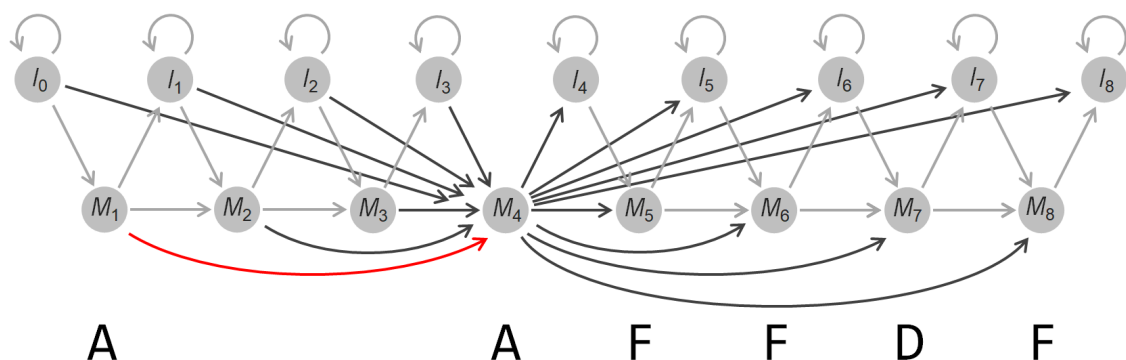
Прво, могуће је суочити се са проблемом опажања дужине $k > n$. Једноставно решење јесте додавање нових стања, која би дозволила додатне приказе. Прецизније, најбоље је додати $n + 1$ стање инсерције I_i , за индекс $i \in \{0, \dots, n\}$. Замисао је да посета стању I_i дозволи емитовање додатних симбола између колона i и $i + 1$ поравнања. Стога се додају гране од M_i ка I_i , као и од I_i ка M_{i+1} . Наравно, како би било могуће емитовати више додатних симбола, додаје се и петља – грана од I_i ка самом себи. Специјално, стање I_0 моделује инсерције пре првог карактера поравнања, док I_n дозвољава уметања након последњег слова. Слика 4.8 приказује употребу нових стања у одређивању пута кроз дорађени ланац са слике 4.6 за досад необјашњиво опажање *A***F***DDA**FFDF*. Ово је, дакле, опажање *ADDAFFDF* са слике 4.7, али уз уметнуто *F* на другом месту. Оптимални пут је подебљан.



Слика 4.8: Увођење стања инсерције, према [30]

Даље, могуће је ухватити се укоштац и са проблемом опажања дужине $k < n$. Сада је задатак моделовати делеције, односно омогућити прескакање неких

колона поравнања. Наиван приступ овом решењу било би баш прескакање неких поклапања додавањем великог броја грана. Наиме, могла би се додати по грана ка сваком M_i од сваког M_j и I_j за $j < n$, као и по једна од сваког M_i ка сваком M_j и I_j за $j > n$. Другим речима, тада би се у неко стање поклапања могло непосредно доћи из било ког стања лево, као и непосредно отићи у било које стање десно. Како би то изгледало на дорађеном моделу са слике 4.8, може се видети на слици 4.9, која следи. Црвеном бојом наглашена је грана која омогућује рад са досад необјашњивим опажањем $AAFFDF$, дакле са уклоњена два карактера DD . Како слика не би била претерано хаотична, додате су само нове гране чији је један крај поклапање M_4 .

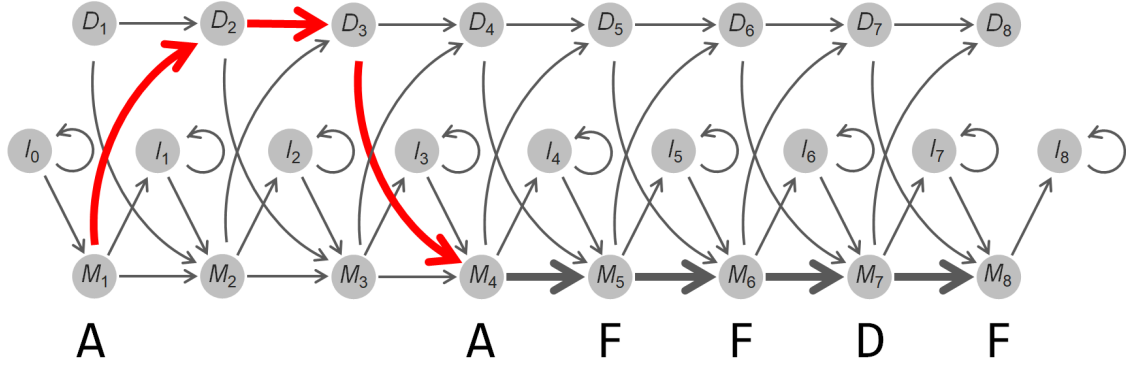


Слика 4.9: Наивна обрада делеција, према [30]

Овакав приступ, међутим, није пожељан, управо због великог броја додатих грана. Подсећања ради, сви важни алгоритми за рад са *HMM* засновани су на Витербијевом графу. Када се говорило о томе, напоменуто је да је сложеност Витербијевог и повезаних алгоритама директно сразмерна броју грана у графу, који, с друге стране, зависи од броја дозвољених прелаза у самом моделу. Претходном дорадом, број прелаза је са линеарног скочио на квадратни – реда $O(n^2)$, где n , као и досад, означава број колона у поравнању.

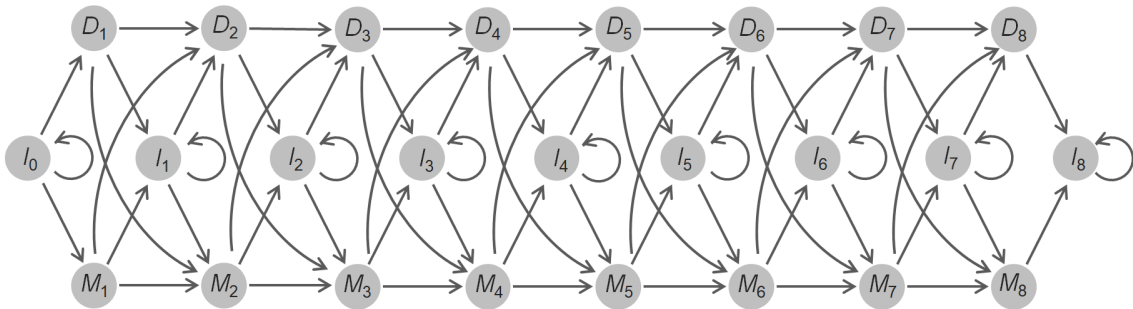
Како би рад са моделом био ефикасан, неопходно је задржати линеаран број грана. То се чини додавањем нових, тихих стања, која не приказују ништа (понекад се замшља да емитују празнину „-“), а представљају алтернативан пут у односу на главни, који иде преко стања поклапања. Прецизније, додаје се n стања делеције D_i , таквих да се од сваког D_i може доћи до M_{i+1} и D_{i+1} , као и од сваког M_i до D_{i+1} . Сада је колону поравнања i , односно стање M_i , могуће прескочити проласком кроз стање D_i . Слика 4.10 приказује

употребу нових стања у одређивању пута кроз дорађени модел са слике 4.8 за већ разматрано опажање $AAFFDF$. Оптимални пут је и сада подебљан.



Слика 4.10: Увођење стања делеције, према [30]

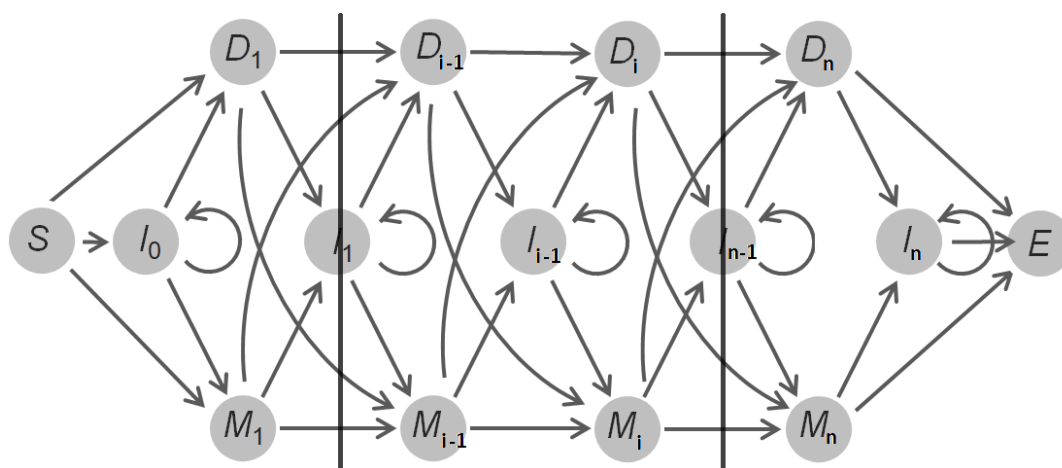
Сада су могући прелази између стања поклапања и инсерције, као и између стања поклапања и делеције. Како би модел био комплетан, фале још прелази између стања инсерције и делеције. Конкретно, треба омогућити прелаз из сваког D_i на I_i , као и од сваког I_i ка D_{i+1} . Сада је могуће произвољно мењати тип стања: од инсерције $I_i \mapsto I_i$, $I_i \mapsto D_{i+1}$, $I_i \mapsto M_{i+1}$, од делеције $D_i \mapsto I_i$, $D_i \mapsto D_{i+1}$, $D_i \mapsto M_{i+1}$, од поклапања $M_i \mapsto I_i$, $M_i \mapsto D_{i+1}$, $M_i \mapsto M_{i+1}$. Коначна структура модела са осам колона пречишћеног вишеструког поравнања, који одговара мотивационом примеру, дата је на слици 4.11.



Слика 4.11: Коначна структура модела, према [30]

Већ је примећено да се код профилних модела са n не означава број скривених стања, већ број колона пречишћеног вишеструког поравнања. Стања је трипут више. Још једна њихова специфичност је да је уобичајен рад са

експлицитним почетним стањем S (од енгл. *Start*) и завршним E (од енгл. *End*), тако да се и она додају. Од полазног стања могуће је ући у I_0 , D_1 или M_1 , док се у терминално долази из I_n , D_n или M_n . Кад се све сабере, општи профилни HMM са n колона могао би се приказати дијаграмом са слике 4.12.



Слика 4.12: Општи профилни HMM , према [30]

Ово је, дакле, такозвани профилни HMM или HMM профил. Како се гради на основу поравнања P и границе одсецања θ , означава се и као $HMM(P, \theta)$. Његово одређивање формално се представља кроз проблем 12.

Проблем 12: Одређивање профилног модела[16]

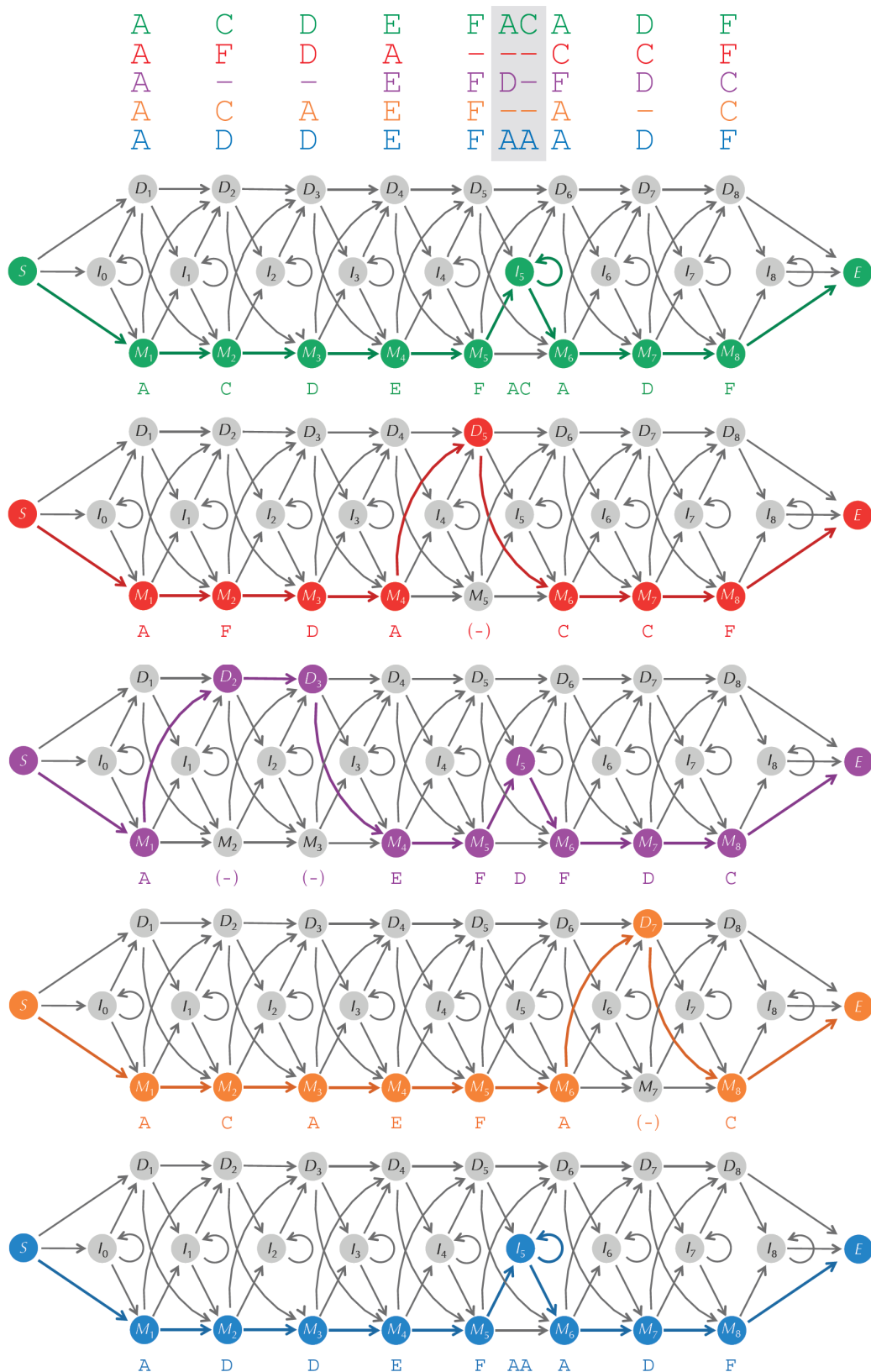
Најправилни профилни HMM на основу вишеструког поравнања.

Улаз: вишеструко поравнање P и граница одсецања θ .

Издаз: профилни модел $HMM(P, \theta)$.

Подсећања ради, сваки HMM је петорка, а засад је одређен само први члан – скривена стања x , којих је укупно $3n + 3$. Одређен је и скуп дозвољених прелаза, па тако из сваког стања излазе највише по три гране, што значи да је укупан број прелаза линеаран – реда $O(n)$, како је и захтевано. Могуће опсервације, као други члан петорке, познате су из улазног поравнања, а зависе од примене. У случају рада са протеинима, скуп y гради двадесетак аминокиселина. Из профилне матрице су познате и емисије у стањима поклапања, а познати су и обавезно почетно (трећи члан петорке π) и завршно стање.

Фале, међутим, све вероватноће прелаза a , као и расподела опажања b у непклапајућим стањима. Њихово одређивање илустровано је сликом 4.13.



Слика 4.13: Одређивање параметара профилног *HMM*, према [30]

4.4 Рад са профилима

Вероватноће се рачунају емпиријски, на основу улазног вишеструког поравнања. Претходна слика сваку од обојених ниски из поравнања (исти мотивациони пример као и досад, са прве слике 4.6) приказује као оптимални пут исте боје кроз профилни *НММ*. Веза између ниске као реда вишеструког поравнања и оптималног пута кроз профилни *НММ* једнозначна је. Уколико се у задржаној колони (чиста позадина) i налази емитовани симбол, пролази се кроз стање поклапања M_i . Ако задржана колона i складишти празнину „-“, пролази се кроз стање делеције D_i . Празнина се у пречишћеној колони (осечена позадина) која се налази између задржаних колона i и $i + 1$ занемарује, док емитовани симболи означавају пролазак кроз стање инсерције I_i .

На основу одређених скривених путева, могуће је простим бројањем одредити колико често долази до неког прелаза, те расподелу емисија по сваком стању. Нека се разматра пример са слике 4.13, који садржи пет путева. Четири пута – сваки сем црвеног – пролазе кроз стање M_5 . Од њих, један пут – наранџасти – наставља у стање M_6 , док преостала три одлазе у стање I_5 . Овиме је одређена фреквенција дозвољених прелаза, па су вероватноће:

$$a_{M_5, I_5} = \frac{3}{4}, a_{M_5, D_6} = 0, a_{M_5, M_6} = \frac{1}{4}.$$

Аналогно се одређују вероватноће прелаза између осталих стања, укључујући специјално полазно и завршно. Ако емпиријски путеви не покривају све прелазе, као између M_5 и D_6 , што је сасвим уобичајено, вероватноћа прелаза сматра се непознатом, односно у питању је имплицитна нула (недозвољени прелаз). Ово је један од примера када мапе не покривају цео простор вероватноћа, односно оне се не сумирају у јединицу, иако би то било очекивано.

И вероватноће опажања одређују се фреквенцијски. На примеру стања M_5 , у сва четири пролаза емитован је симбол F . Стога је мапа следећа:

$$b_{M_5, A} = b_{M_5, B} = b_{M_5, C} = b_{M_5, D} = b_{M_5, E} = 0, b_{M_5, F} = \frac{4}{4} = 1.$$

Такође постоји доста непознатих емисија, па је удео имплицитних нула велик.

Свеукупно, главни утисак је да је значајан број непознатих вредности. Ово онемогућава добро оцењивање великог броја секвенци чије би декодирање ишло тим путевима. Њихова вероватноћа била би нула, што је проблем по ком тренутни модел личи на полазни ланац. Да је стварно тако, показује резултујућа мапа прелаза мотивационог примера, приказана на слици 4.14.

	S	I ₀	M ₁	D ₁	I ₁	M ₂	D ₂	I ₂	M ₃	D ₃	I ₃	M ₄	D ₄	I ₄	M ₅	D ₅	I ₅	M ₆	D ₆	I ₆	M ₇	D ₇	I ₇	M ₈	D ₈	I ₈	E
S			1																								
I ₀																											
M ₁						.8	.2																				
D ₁																											
I ₁																											
M ₂									1																		
D ₂										1																	
I ₂																											
M ₃												1															
D ₃												1															
I ₃																											
M ₄														.8	.2												
D ₄																											
I ₄																											
M ₅																.75	.25										
D ₅																	1										
I ₅																	.4	.6									
M ₆																				.8	.2						
D ₆																											
I ₆																											
M ₇																							1				
D ₇																							1				
I ₇																											
M ₈																										1	
D ₈																											
I ₈																											
E																											

Слика 4.14: Вероватноће прелаза мотивационог примера, према [30]

Моделом недозвољени прелази, попут $M_1 \mapsto M_5$, илустровани су ћелијама чисте позадине, док су могући прелази осенчени. Приметно је да је од укупно 75 дозвољених прелаза, само 19 евалуирано, док је преостала већина непозната – у питању су имплицитне нуле. Како ово није пожељна особина, у наставку је превазиђено употребом малих вероватноћа за непознате прелазе.

Пре тога, међутим, ваља указати на други најављени пропуст у књизи, који се баш налази на уџбеничкој верзији претходне слике. Конкретно, аутори су пермутовали прелази $D_7 \mapsto M_8$ и $I_7 \mapsto M_8$, па је тако првом додељена нулта вероватноћа уместо јединична, а другом обрнуто. Место јединичне вероватноће је на овдашњој слици исправљено и наглашено црвеном бојом. Грешка се налази и на Певзнеровој презентацији, на којој је чак већа. Наиме,

ту је дошло до вишеструких пермутација, па је тако свака ненулта вероватноћа у петој колони поравнања (прелази са I_5 , D_5 и M_5) погрешна. Уместо овдашњих исправних $\{.75, .25, 0\}$, $\{0, 1, 0\}$, $\{.4, .6, 0\}$, на презентацији су вредности у петом квадрату погрешних $\{.25, .75, 0\}$, $\{.33, .67, 0\}$, $\{0, 1, 0\}$. Да ствар буде гора, пермутоване су вредности и у другом и трећем квадрату.

Било како било, треба се вратити на проблем великог удела нула у емпиријски одређеним мапама преласка и емисија. Поменуто је већ да је главни проблем који то изазива лоше оцењивање многих секвенци и путева. Конкретан пример дат је касније. Други проблем је што овакве мапе не покривају цео простор вероватноћа, односно не сумирају се све излазне вероватноће у јединицу. Све ово се, међутим, лако превазилази увођењем псеудовредности σ , што су мале вероватноће дозвољених, али непознатих прелаза и емисија. Оне се, дакле, постављају у мапе тамо где би биле имплицитне нуле.

Наравно, приликом додавања псеудовредности не треба заборавити на нормализацију, како би збир вероватноћа стварно био један. Примера ради, за $\sigma = 1/100$, ред $\{.75, .25, 0\}$ не постаје $\{.75, .25, .01\}$, већ $\{.7426, .2475, .0099\}$. Исто тако, сасвим непознати ред $\{0, 0, 0\}$ за свако σ постаје $\{1/3, 1/3, 1/3\}$, уместо $\{1/100, 1/100, 1/100\}$ у конкретном случају. Такође, не треба заборавити да се псеудовредности додају искључиво дозвољеним прелазима, док недозвољени остају нулте вероватноће. Одређивање овако дорађеног профилног модела $HMM(P, \theta, \sigma)$ формално се представља кроз проблем 13.

Проблем 13: Одређивање дорађеног профилног модела[17]

Најправилни профилни HMM на основу вишеструког поравнања.

Улаз: поравнање P , граница θ , псеудовредност σ .

Излаз: дорађени профилни модел $HMM(P, \theta, \sigma)$.

Овиме је појам профилног скривеног Марковљевог модела комплетиран и чини се да је коначно могуће бацити се у рад са њим. Идеја оваквог модела заправо је двојака. Први циљ је већ поменута класификација, која улазу додељује скор припадности одређивањем вероватноће исхода (алгоритам „напред“). Додатно, могуће је одредити и само поравнање ниски са представљеном фамилијом, које се добија декодирањем улаза (Витербијев алгоритам).

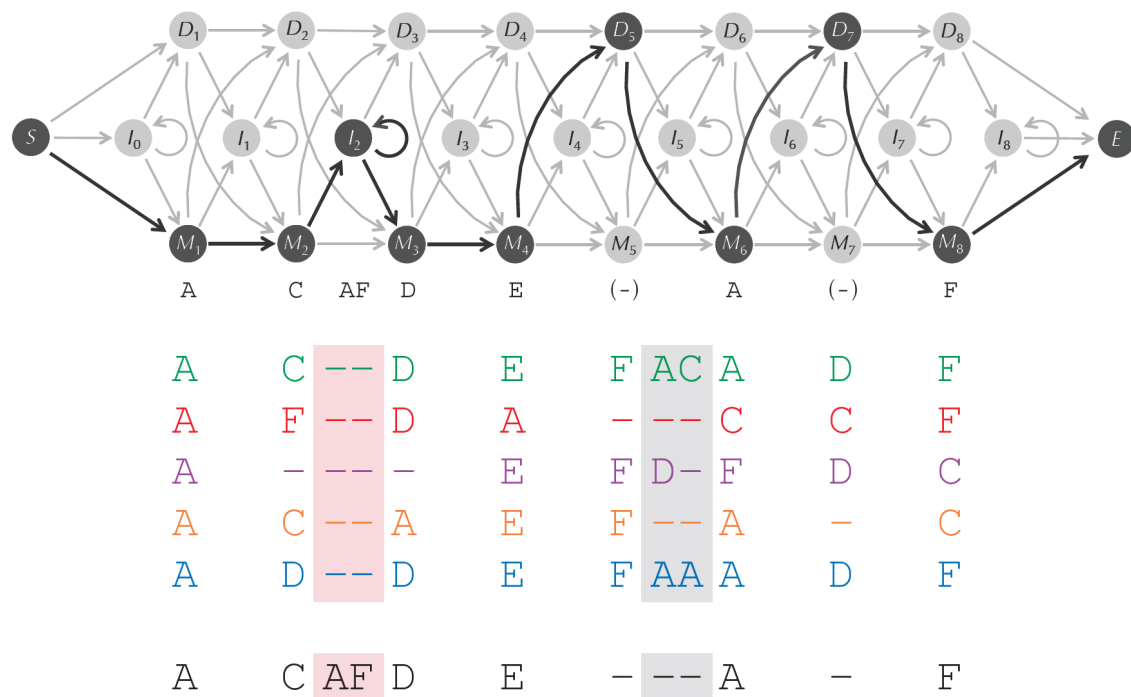
О класификацији је већ било речи – одредити вероватноћу да протеин припада неким породицама, а затим га доделити оној са највећим скором или макар скором који прелази постављену границу. На конкретном примеру

предвиђања фенотипа *HIV*-а из мотивационог увода, могла би постојати два *HMM* профила – један изграђен на основу изолата који стварају синцицијум, а други према оним који га не стварају. Нов изолат за који је упитно ствара ли синцицијум био би улазно опажање за алгоритам „напред” над та два профила, а одговор на питање добио би се одабиром профила са већом вероватноћом.

Одабрани профил се, штавише, опционо може проширити додавањем новог изолата у поравнање, а затим поновним израчунавањем параметара тог *HMM*-а. Ажурирањем профила, он временом све боље описује класу коју моделује, те постаје још прецизнији и употребљивији при класификацији.

Овакав принцип може се уопштити на све друге секвенцијалне податке. Прави се, дакле, по један *HMM* профил за сваку могућу класу, а затим се профили пореде са улазном секвенцом, која се класификује. Инстанци се додељује она класа са највећом вероватноћом опажања. Када су у питању гени и протеини, на интернету је бесплатно доступан претраживач *HMMER*[5], управо заснован на припремљеним скривеним Марковљевим моделима.

Када је други циљ у питању, поравнање ниске са профилем добија се непосредно као резултат декодирања. Пример тога дат је на слици 4.15.



Слика 4.15: Пример поравнања мотивационог примера, према [30]

Приказан је оптимални пут $SM_1M_2I_2I_2M_3M_4D_5M_6D_7M_8E$ кроз досад разматрани профилни HMM за ниску (опажање) $ACAFDEAF$. Ваља приметити да је овај пут иначе немогућ (нулте вероватноће) без употребе псеудовредности, јер је нпр. вероватноћа прелаза $M_2 \mapsto I_2$ непозната, па самим тим имплицитно нулта без неког σ . У бојама је, као и досад, приказана петорка из почетног вишеструког поравнања, док је ниска која се поравнава црна.

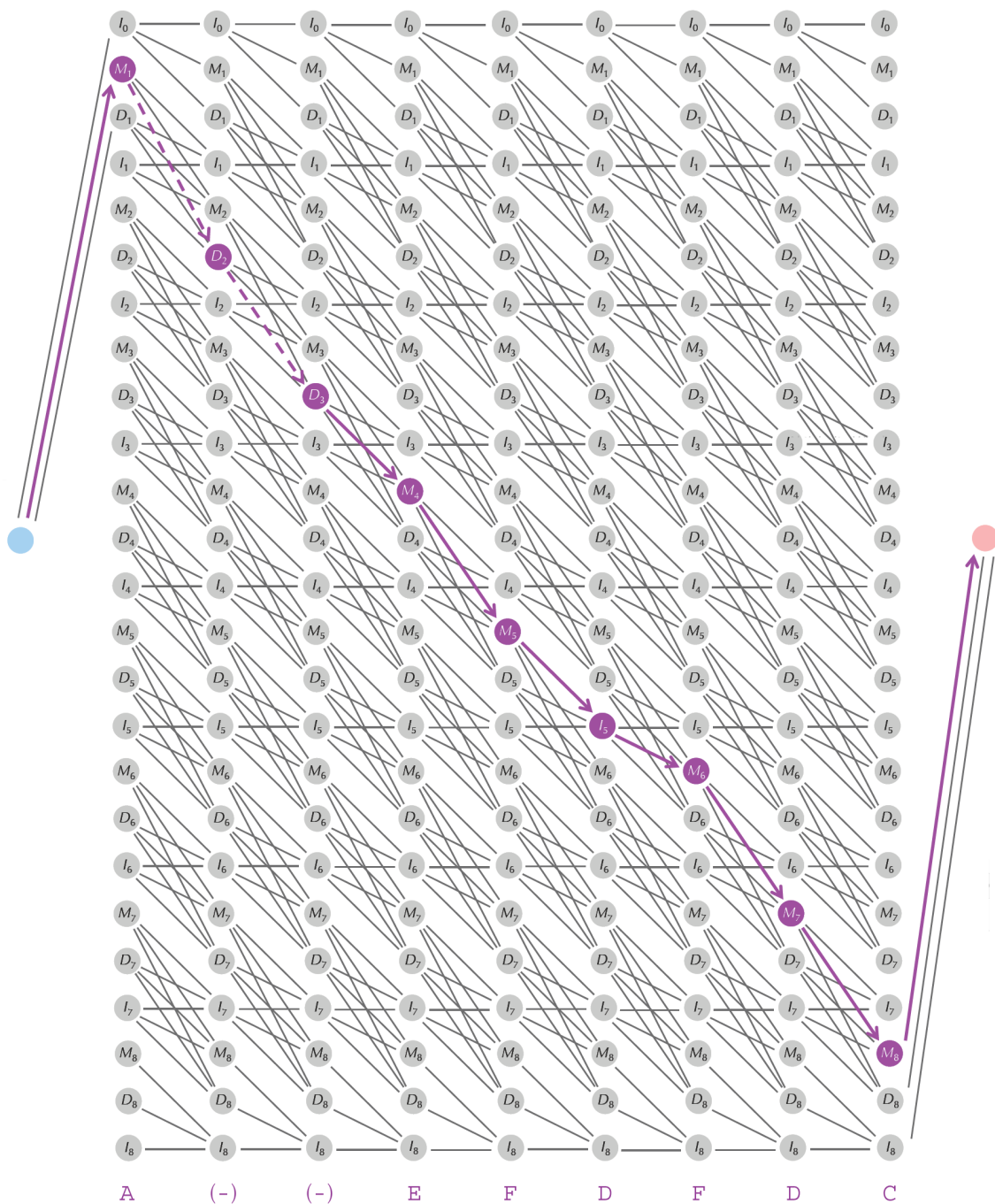
Конкретно, заједнично вишеструко поравнање је следеће. Прва два симбола емитована су из стања поклапања M_1 и M_2 , тако да се налазе у првој, односно другој колони поравнања. Следећа два симбола емитована су из стања инсеције I_2 , тако да се налазе између друге и треће колоне, што је назначено розе сенком. Како је у питању посебан додатак за нову ниску, полазној петорци једноставно се додају по две празнине. Наредна два симбола емитована су из стања поклапања M_3 и M_4 , тако да се налазе у трећој, односно четвртој колони поравнања. Следеће на путу јесте тихо стање делеције D_5 , тако да нема емисије, већ се у пету колону поравнања ставља симбол празнине „-“. Наредни симбол емитован је из стања поклапања M_6 , па се ставља у шесту колону поравнања. Пре њега није било инсерција, тако да се две пречишћене колоне попуњавају празнинама. Следи тихо стање делеције D_7 , што значи да се у седму колону ставља празнина, док је последњи симбол емитован из стања поклапања M_8 , те се ставља у осму колону поравнања.

Како би поравнање било добро одређено, неопходно је пратити текући карактер нове ниске која се поравнава са фамилијом, односно кренути од првог слова и померати „показивач“ када дође до емисије. Такође, може се посебно пазити да ли се прелази преко пречишћеног дела (сиво сенчење), мада то и није толико важно. Општа правила поравнавања сумирају се следећим списком смерница за тумачење стања оптималног скривеног пута:

- пролазак кроз стање поклапања M_i заправо представља емитовање текућег симбола, те поставља тај симбол управо у колону i поравнања,
- пролазак кроз стање инсерције I_i такође представља емитовање текућег симбола, али поставља тај симбол између колоне i и $i + 1$ поравнања,
- пролазак кроз стање делеције D_i оставља текући симбол на чекању, те поставља алтернативни симбол празнине „-“ у колону i поравнања.

Поравнање је смислено уколико су проласком кроз скривени пут потрошени сви карактери улазног опажања, што се осигурава позадинским алгоритмом.

Остаје још питање како тачно наћи оптимални скривени пут. Јасно је да се може применити Витербијев алгоритам, који је досад коришћен за проблем декодирања. На слици 4.16 приказан је Витербијев граф и одговарајући скривени пут за љубичасто опажање $AEFDFDC$ из мотивационог примера.



Слика 4.16: Наивни Витербијев граф профилног HMM , према [30]

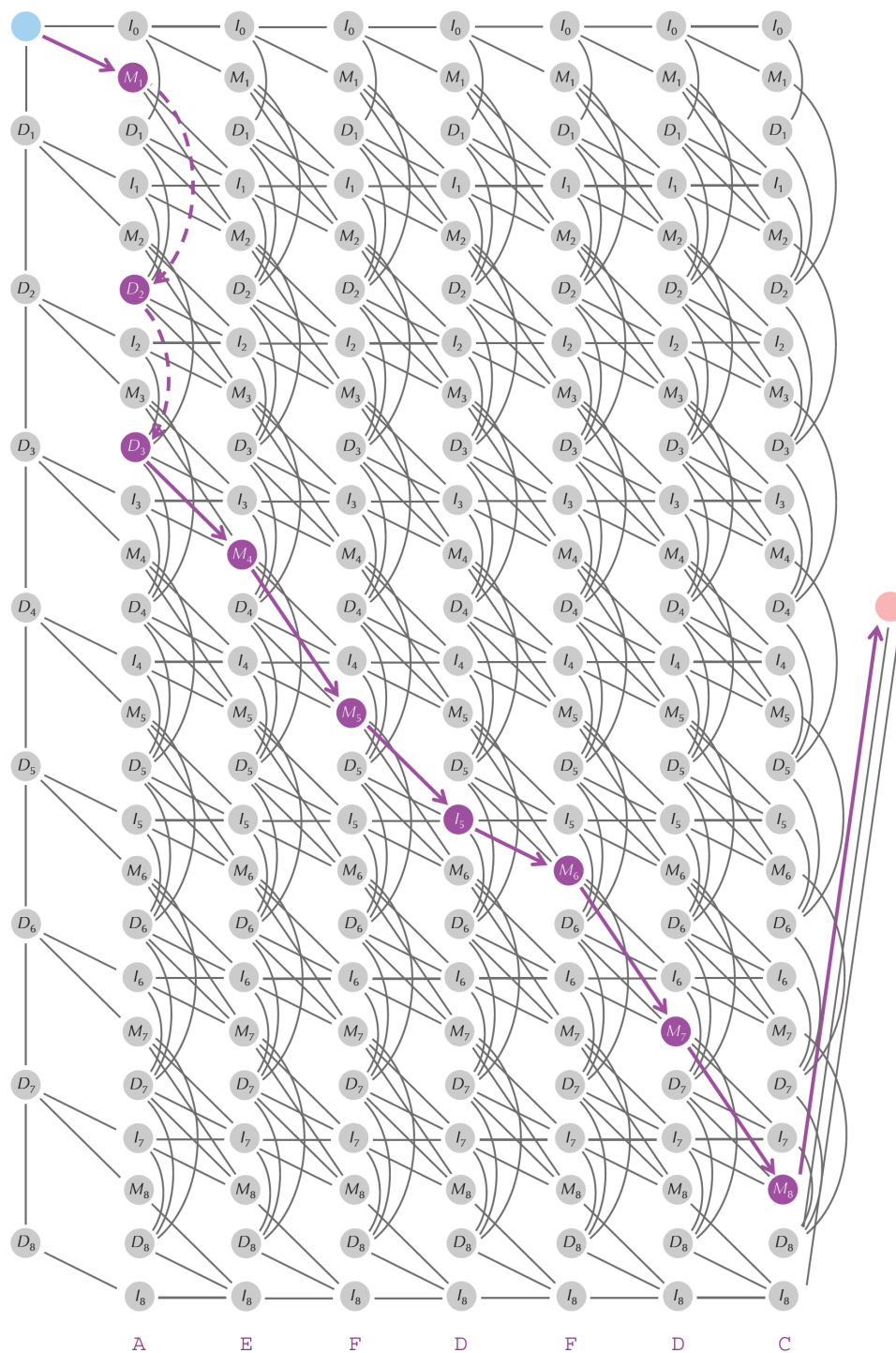
Подсећања ради, основа Витербијевог графа је мрежа чворова чије редове чине сва могућа скривена стања (у конкретном случају профилног модела, за $n = 8$, има их 25, без експлицитног почетног и завршног, а у општем $3n + 1$), док колоне означавају ток времена, тренутак t . Из сваког чвора у колони $t - 1$ усмерена је по једна грана у сваки чвор из колоне t , али искључиво ако је дозвољен прелаз између та два стања. Тако је на основу чињенице да се из сваког стања у тренутку $t - 1$ може прећи у било које стање у тренутку t , уколико је вероватноћа преласка ненулта. Поред ове основе, мрежа има и два посебна чвора – извор (експлицитно почетно стање – плави кружић) и понор (експлицитно завршно стање – црвени кружић). Из извора иду три гране – ка I_0 , M_1 и D_1 – док у понор увиру такође само три гране – од M_8 , D_8 и I_8 (односно од индекса n уместо 8 у општем случају). Број колона (тренутака) заправо је дужина пута k , а замисао овакве мреже управо и јесте да истовремено моделује све скривене путеве дужине k кроз упитни *HMM*.

Проблем код графа са слике 4.16 управо је превише прецизно одабран број колона. Приказани граф као да зна да је оптимални пут дужине $k = 9$, а љубичасто опажање заправо $A - -EFDFDC$ уместо $AEFDFDC$. Овакво знање, међутим, не може бити доступно пре покретања алгоритма, што значи да граф са слике 4.16 технички и није Витербијев. Штавише, могао би се направити већи број графова сличних претходном, али нпр. без друге и/или треће колоне, и потпуно једнако употребити за одређивање оптималног пута. Резултат, међутим, не би био тачан, јер би се разматрали само путеви дужине $k < 9$, па се свакако не би могао добити оптимални, чија је дужина $k = 9$.

Витербијев граф стога мора имати фиксан број колона, и то овде тачно $k = 7$. У питању је број емитованих симбола, односно дужина познатог опажања, а не непознатог скривеног пута. Ово пре увођења стања делеције није било проблематично, пошто су дужина опажања и дужина оптималног скривеног пута увек биле једнаке. Тиха стања, међутим, нарушавају ову једнакост. У конкретном случају профилних *HMM*, опажање дужине k може настати на скривеном путу који је најмање дужине баш k , уколико пут садржи само стања поклапања, а највише $n + k$, уколико пут не садржи ниједно стање поклапања. Сада је циљ Витербијевим графом са k колона некако моделовати путеве различитих дужина, а на којима се емитује тачно k симбола.

Проблем, дакле, праве тиха стања делеције, пошто мењају смисао чвора (x_i, t) графа. Првобитно, пролазак кроз тај чвор значио је да се у тренутку t

модел налази у стању x_i , односно да је симбол o_t улазног опажања o емитован из стања x_i . Ово, међутим, нема смисла код тихих стања, која ништа не приказују. Поправљени Витербијев граф приказан је на слици 4.17.



Слика 4.17: Коначни Витербијев граф профилног *HMM*, према [30]

Смисао се сада допуњује у следећи: пролазак кроз чвор (x_i, t) графа значи да се *НММ* налазио у стању x_i када је емитовао симбол o_t уколико стање x_i није тихо, односно да се налазио у тихом стању x_i након емитовања симбола o_t , а пре емитовања o_{t+1} . У конкретном случају профилних модела, пролазак кроз (M_i, t) или (I_i, t) значи да је у стању поклапања M_i или инсерције I_i емитовано o_t , док пролазак кроз (D_i, t) означава да је аутомат био у тихом стању делеције D_i између емитовања суседних симбола опажања o_t и o_{t+1} .

У коначној верзији Витербијевог графа са слике 4.17, претходно објашњена допуна смисла читава се двама променама. Прво, сваки прелаз у стање делеције, дакле прелаз облика $I_i \mapsto D_{i+1}$, $M_i \mapsto D_{i+1}$ или $D_i \mapsto D_{i+1}$, сада се дешава у оквиру исте колоне. Ово непосредно осликава чињеницу да делеција не емитује нови симбол, што значи да индекс (тренутак) t остаје непромењен. Наравно, преласци ка стањима поклапања или инсерције остају скок у следећу колону, пошто за собом повлаче нову емисију. Ово свеукупно значи да једна колона строго одговара опажању једног симбола, мада се то може десити проласком кроз различите скривене путеве, што је и био циљ.

Друга промена одговара проширењу нулте колоне, у којој се досад налазио само извор. Она се проширује низом (ланцем) повезаних стања делеције, свим могућим од D_1 до D_8 , односно D_n у општем случају. Наиме, уколико не би било тог проширења, пут би завршио у ћорсокаку уколико би започео првим стањем делеције D_1 у првој колони. Прелази су једноставно такви да након уласка у стање делеције није више могуће емитовати симбол у текућој колони. Стога се тај прелаз пребацује у нулту колону, сачињену искључиво од тихих стања. Подразумевано, све гране у истој колони оријентисане су надоле, док су прелази између колоне оријентисани надесно. И слика 4.17, попут претходне, декодира љубичасту ниску *AEFDFDC* из уводног поравнања.

Поравнање помоћу профилног *НММ*-а формално се дефинише кроз проблем 14. У питању је, дакле, нешто напредније декодирање, дорађена верзија проблема 1. Решење оба је Витербијев алгоритам, само се разликује Витербијев граф над којим се ради, пошто сад постоје и тиха стања делеције.

Проблем 14: Поравнање са профилним моделом[19]

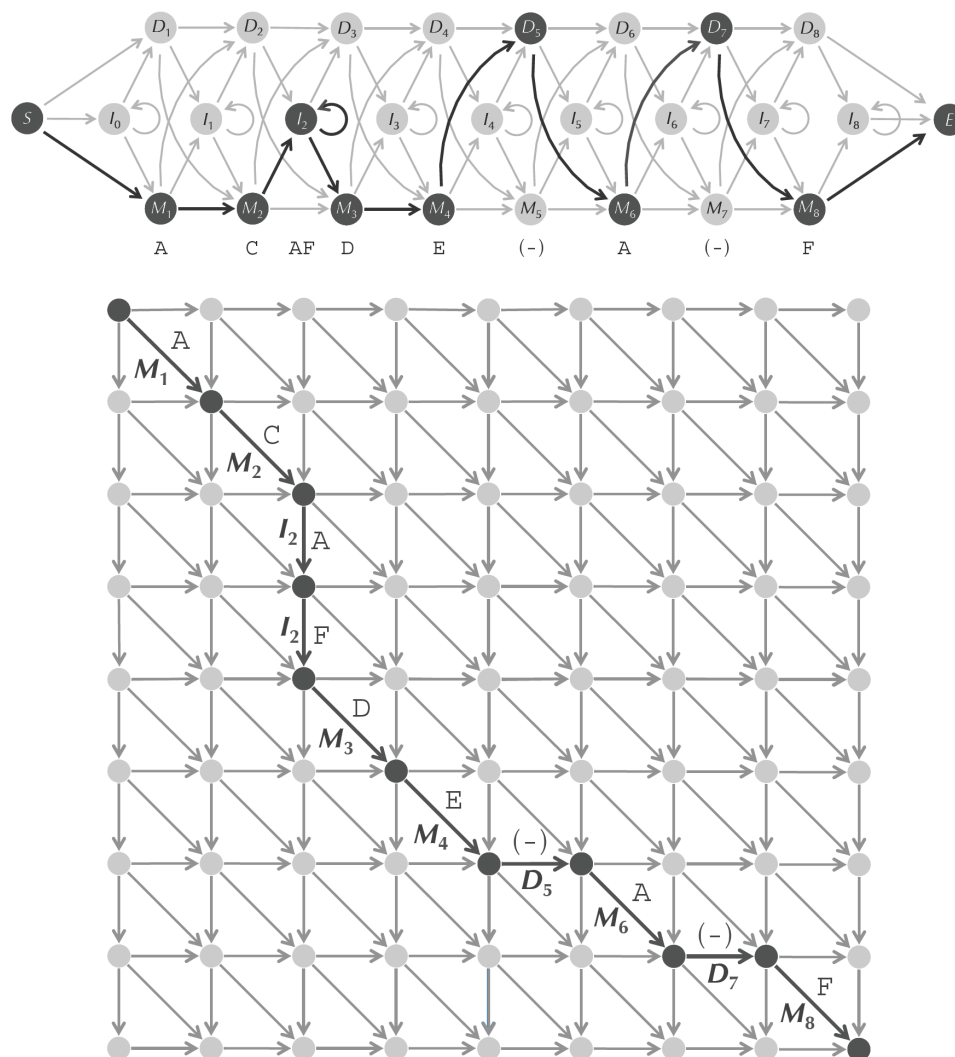
Поравнавши нову ниску са породицом – профилни НММ.

Улаз: вишеструко поравнање P , граница одсецања θ ,
псеудовредност σ , нова ниска (опажање) o .

Излаз: оптимални пут p кроз $НММ(P, \theta, \sigma)$ за ниску o .

На сличан начин је, иначе, могуће конструисати Витербијев граф за произвољни *HMM* са тихим стањима. Једини услов који мора да буде испуњен јесте да не постоје петље или циклуси који се састоје искључиво од тихих стања. У супротном, није могуће решити проблем декодирања, али ни друге сличне проблеме засноване на Витербијевом графу (алгоритам „напред” итд.).

Ово је аналогно чињеници да није могуће решити проблем оптималног обиласка Менхетн графа из петог поглавља (*Chapter 5: How Do We Compare DNA Sequences? – Dynamic Programming*) са циклусима. Разлог томе је што се, како код Менхетна, тако и код Витербија, вредности морају рачунати према тополошком редоследу чворова. Другим речима, неопходно је да буду познате све родитељске вредности како би се евалуирао нови, текући чвор.



Слика 4.18: Поређење поравнања мотивационог примера, према [30]

Граф, дакле, мора бити без циклуса – усмерени ациклички граф. Тополошки редослед у супротном не постоји. У конкретном случају профилних *НММ*, уобичајен редослед рачунања је слева надесно, од врха надолу. Успут се, наравно, чувају путокази, како би се оптимални пут на крају могао реконструисати. Без њих, добила би се само вероватноћа најбољег пута.

Претходна сличност између Витербијевог и Менхетн графа није случајна. Испоставља се да су то у неку руку аналогне структуре, а не само проблеми. На слици 4.18 приказана је њихова наизглед једнакост на већ познатом примеру ниске *ACAFDEAF* и њеног поравнања са слике 4.15. Слика илуструје како пут у Менхетн графу одговара скривеном путу кроз профилни *НММ*. Дијагоналне ивице Менхетна одговарају поклапањима, усправне инсерцијама, а водоравне делецијама. Важно је истаћи да ова аналогија ипак није једнакост (еквиваленција), пошто код *НММ*-а постоје променљиве вероватноће (тежине) прелаза и емисија, што није могуће моделовати помоћу Менхетна.

Способност профилних скривених Марковљевих модела да различито оцењују различите колоне матрице поравнања издваја их као прецизније у односу на једноставне методе поравнања засноване на једној матрици са истим скоровима. Профилни модели тако могу ухватити суптилне сличности, које једноставна поравнања пропуштају. Можда најбољи део свега јесте да, упркос тој предности, сложеност остаје подједнако добра, и износи $O(nk)$ за n колона (дужина полазног поравнања) и k редова (дужина опажања).

Код Менхетна је сложеност очигледна, јер се оперише над матрицом димензија $n \times k$, где се вредност сваког чвора израчунава кроз највише три гране. Ни код профилних *НММ* није тешко одредити је. Већ је више пута напоменуто да је сложеност Витербијевог алгоритма сразмерна броју грана у Витербијевом графу. У општем случају, када се из сваког стања може прећи у било које друго, износи $O(n^2k)$, што је последица постојања по n прелаза између n стања у $k - 1$ промени тренутка. Код профилних *НММ*, међутим, постоји већи број стања $3n + 1$, као и већи број промена тренутка k , али је прелаза највише по три (константа), тако да је асимптотски производ $O(nk)$.

За крај, следе тачне формуле максимизације вероватноће пута у чворовима Витербијевог графа код профилних *НММ* са n колона поравнања за познато опажање o дужине k . Још једном, нека мапа скорова s буде таква да $s_{x_i,t}$ складишти вероватноћу оптималног пута дужине t који се завршава скривеним стањем x_i . У конкретном случају, уместо општих стања x_i , одвојено се

разматрају специјализована стања поклапања M_i , делеције D_i и инсерције I_i .

База рекурзије овако постављеног проблема јесте (први чланови низа):

$$\begin{matrix} (i = 1) \\ (t = 1) \end{matrix} s_{M_1,1} = a_{S,M_1} \cdot b_{M_1,o_1},$$

$$\begin{matrix} (i = 1) \\ (t = 0) \end{matrix} s_{D_1,0} = a_{S,D_1},$$

$$\begin{matrix} (i = 0) \\ (t = 1) \end{matrix} s_{I_0,1} = a_{S,I_0} \cdot b_{I_0,o_1}.$$

Рекурзивне формуле максимизације су (непостојећи индекси се занемарују):

$$\begin{matrix} (\forall i \in \{2, \dots, n\}) \\ (\forall t \in \{2, \dots, k\}) \end{matrix} s_{M_i,t} = \max \begin{cases} s_{M_{i-1},t-1} \cdot a_{M_{i-1},M_i} \cdot b_{M_i,o_t} \\ s_{D_{i-1},t-1} \cdot a_{D_{i-1},M_i} \cdot b_{M_i,o_t} \\ s_{I_{i-1},t-1} \cdot a_{I_{i-1},M_i} \cdot b_{M_i,o_t} \end{cases},$$

$$\begin{matrix} (\forall i \in \{2, \dots, n\}) \\ (\forall t \in \{1, \dots, k\}) \end{matrix} s_{D_i,t} = \max \begin{cases} s_{M_{i-1},t} \cdot a_{M_{i-1},D_i} \\ s_{D_{i-1},t} \cdot a_{D_{i-1},D_i} \\ s_{I_{i-1},t} \cdot a_{I_{i-1},D_i} \end{cases},$$

$$\begin{matrix} (\forall i \in \{1, \dots, n\}) \\ (\forall t \in \{2, \dots, k\}) \end{matrix} s_{I_i,t} = \max \begin{cases} s_{M_i,t-1} \cdot a_{M_i,I_i} \cdot b_{I_i,o_t} \\ s_{D_i,t-1} \cdot a_{D_i,I_i} \cdot b_{I_i,o_t} \\ s_{I_i,t-1} \cdot a_{I_i,I_i} \cdot b_{I_i,o_t} \end{cases}.$$

Коначан оптимални (највероватнији) пут добија се додатном максимизацијом:

$$P\{p_{opt}, o\} = \max_p P\{p, o\} = \begin{matrix} (i = n) \\ (t = k) \end{matrix} \max\{s_{M_n,k}, s_{D_n,k}, s_{I_n,k}\}.$$

У уџбеничкој верзији претходних формула налази се трећи најављени пропуст из књиге. Конкретно, код рекурзивне формуле за стања поклапања M_i , која је једина и приказана, стављене су вероватноће прелаза a_{I_{i-1},I_i} и a_{D_{i-1},D_i} уместо a_{I_{i-1},M_i} и a_{D_{i-1},M_i} . Ова грешка, међутим, није присутна на Певзнеровој презентацији, где су формуле тачно написане, мада користе друге ознаке.

Као и досад, аналогно се формирају логаритамске верзије формула, које множење мењају сабирањем. Идентична је и општа верзија формула са произвољним тежинама τ . Поред декодирања као проблема 14, једнако се приступа и осталим задацима заснованим на Витербијевом графу – табела 3.1. Решење је такође у одабиру одговарајућих оператора уместо максимума.

Глава 5

Учење модела

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 коктелом 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[24].

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[22]. Она се

данас састоји од скоро 20.000 вишеструких поравнања разних протеинских домена и рутински се користи у анализи нових протеинских секвенци[6].

Све у свему, скривени Марковљеви модели прешли су дуг пут од својих првих употреба у рачунарској биологији (нпр. Черчил 1989[27], Круг и сарадници 1994[44], Балди и сарадници 1994[21]) 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[12]. 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] HMMER: biosequence analysis using profile hidden Markov models. алат бесплатно доступан на интернет адреси: <http://hmmer.org/>.
- [6] Pfam 34.0 (March 2021, 19179 entries). информације о алату и сам алат бесплатно доступни на интернет адреси: <http://pfam.xfam.org/>.
- [7] The GENSCAN Web Server at MIT. информације и алат бесплатно доступни на интернет адреси: <http://argonaute.mit.edu/GENSCAN.html>.
- [8] 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/>.
- [9] Андрей Андреевич Марков. Распространение закона больших чисел на величины, зависящие друг от друга. *Известия физико-математического общества при Казанском университете*, 2(15):135–6, 1906.

- [10] Лазар Васовић. Биоинформатика, 07 2021. Classtools.net, ауторски Венов дијаграм: <https://www.classtools.net/Venn/202107-QTgda5>.
- [11] Лазар Васовић. COVID u Srbiji, 05 2021. GitHub, репозиторијум доступан на интернет адреси: <https://github.com/matfija/COVID-u-Srbiji>.
- [12] Лазар Васовић. HMM u bioinformatici, 09 2021. GitHub, репозиторијум доступан на: <https://github.com/matfija/HMM-u-bioinformatici>.
- [13] Compute the Probability of a Hidden Path, 09 2015. ROSALIND, задатак из уџбеника доступан на: <http://rosalind.info/problems/ba10a/>.
- [14] Compute the Probability of a String Emitted by an HMM, 09 2015. ROSALIND, доступно на: <http://rosalind.info/problems/ba10d/>.
- [15] Compute the Probability of an Outcome Given a Hidden Path, 09 2015. ROSALIND, доступно на: <http://rosalind.info/problems/ba10b/>.
- [16] Construct a Profile HMM, 09 2015. ROSALIND, задатак из уџбеника доступан на: <http://rosalind.info/problems/ba10e/>.
- [17] Construct a Profile HMM with Pseudocounts, 09 2015. ROSALIND, задатак доступан на: <http://rosalind.info/problems/ba10f/>.
- [18] Implement the Viterbi Algorithm, 09 2015. ROSALIND, доступно на интернет страници: <http://rosalind.info/problems/ba10c/>.
- [19] Perform a Multiple Sequence Alignment with a Profile HMM, 09 2015. ROSALIND, доступно на: <http://rosalind.info/problems/ba10g/>.
- [20] 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.
- [21] 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>.
- [22] 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/>.
- [23] 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>.
- [24] 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>.
- [25] 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/>.
- [26] Chris Burge and Samuel Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1):78–94, 04 1997. линк: <http://www.bx.psu.edu/old/courses/bx-fall07/genscan.pdf>.
- [27] 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/>.
- [28] 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/>.
- [29] 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>.

- [30] 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.
- [31] Aron Culotta, David Kulp, and Andrew McCallum. Prediction of complete gene structures in human genomic DNA. 2005. Department of Computer Science, University of Massachusetts, Amherst, чланак бесплатно доступан на адреси: <http://ciir.cs.umass.edu/pubfiles/ir-419.pdf>.
- [32] David DeCaprio, Jade P. Vinson, Matthew D. Pearson, Philip Montgomery, Matthew Doherty, and James E. Galagan. Conrad: Gene prediction using conditional random fields. *Genome Research*, 17(9):1389–98, 09 2007. доступно на: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950907/>.
- [33] 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/>.
- [34] 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>.
- [35] 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.
- [36] J Henderson, S Salzberg, and Fasman K H. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Journal of Computational Biology*, 4(2):127–41, 04 1997. чланак доступан на: <https://drum.lib.umd.edu/bitstream/handle/1903/8004/FindingGenes.pdf>.
- [37] Susan Holmes. Hidden Markov Model for CpG islands, 2012. Stats366 / Stats 166 Course Notes, белешке са предавања доступне на: <https://web.stanford.edu/class/stats366/exs/HMM1.html>.
- [38] Daniel Huson. Grundlagen der Bioinformatik, 07 2020. SoSe’20, Universität Tübingen, професорске белешке са предавања на курсу посвећеном осно-

- вама биоинформатике доступне на: <https://software-ab.informatik.uni-tuebingen.de/download/public/GBi-2020-Script.pdf>.
- [39] Jean-Jacques de Jong, Anthony de 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: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC240176/pdf/jvirol00042-0547.pdf>.
- [40] 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>.
- [41] Manolis Kellis. 8.3: Encoding Memory in a HMM- Detection of CpG islands, 01 2021. Biology LibreTexts, ауторски текст професора доступан на: <https://bio.libretexts.org/@go/page/40962>.
- [42] 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>.
- [43] Keith Knapp and Yi-Ping Phoebe Chen. An evaluation of contemporary hidden Markov model genefinders with a predicted exon taxonomy. *Nucleic Acids Research*, 35(1):317–24, 01 2007. чланак доступан на интернет адреси: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1802560/>.
- [44] 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/>.
- [45] Anthony Lee and Sergei Lebedev. hmmlearn, 02 2021. документација доступна на: <https://hmmlearn.readthedocs.io/en/latest/>, GitHub репозиторијум: <https://github.com/hmmlearn/hmmlearn>.

- [46] P. A. Levene. On the biochemistry of nucleic acids. *Journal of the American Chemical Society*, 32(2):231–240, 1910. чланак делимично доступан на: <https://pubs.acs.org/doi/10.1021/ja01920a010>.
- [47] 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/>.
- [48] 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. доступно на: <https://bmcbgenomics.biomedcentral.com/track/pdf/10.1186/s12864-016-3097-0.pdf>.
- [49] 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.
- [50] 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.
- [51] Jacob Schreiber. pomegranate, 05 2021. документација доступна на: <https://pomegranate.readthedocs.io/en/latest/>, GitHub репозиторијум: <https://github.com/jmschrei/pomegranate>.
- [52] Ron Shamir, Ron Wein, and Nir Avrahami. Lecture 6: January 10, 1999 – Hidden Markov Models, 01 1999. Algorithms for Molecular Biology, Fall Semester, 1998, Blavatnik School of Computer Science, Tel Aviv University, белешке са предавања доступне на: <http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/pdf/lec06.pdf>.
- [53] Mark Stamp. A Revealing Introduction to Hidden Markov Models. 2021. доступно на: <http://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf>.
- [54] Tdunning. File:HiddenMarkovModel.png. Wikimedia Commons, илустрација скривеног Марковљевог модела са Викимедије: <https://commons.wikimedia.org/wiki/File:HiddenMarkovModel.png>.

- [55] 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>.
- [56] 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>.
- [57] Byung-Jun Yoon. File:HMM Eukaryotic.jpg. Wikimedia Commons, илустрација скривеног Марковљевог модела еукариотске *DNA* са Викимедије: https://commons.wikimedia.org/wiki/File:HMM_Eukaryotic.jpg.
- [58] Byung-Jun Yoon. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, 10(6):402–15, 09 2009. чланак доступан на: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2766791/>.