Скривени Марковљеви модели у биоинформатици – електронска лекција

Студент: Лазар Васовић
Ментор: Јована Ковачевић

23. септембар 2021.

МАТФ
Универзитет у Београду
Математички факултет

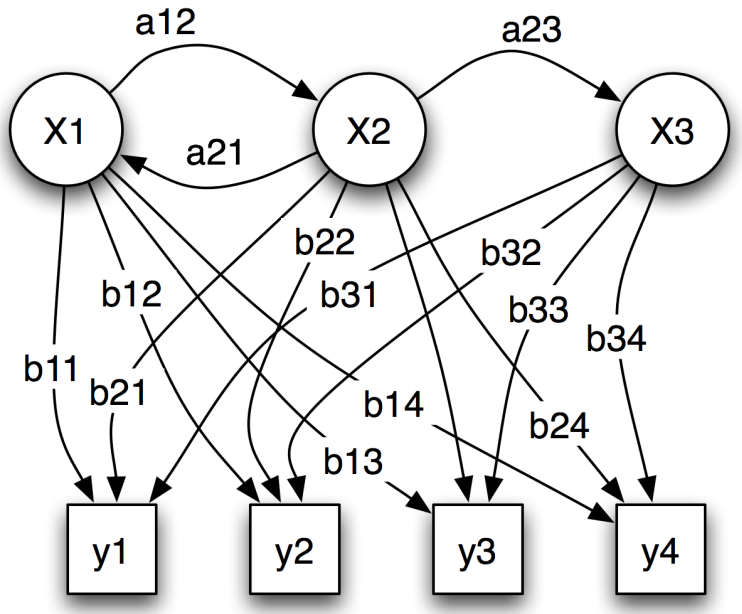
Садржај

Биоинформатика

- Биоинформатика је интердисциплинарна област која се бави применом рачунарских технологија у области биологије и сродних наука, са нагласком на разумевању биолошких података.

Скривени Марковљев модел

- Скривени Марковљев модел (*HMM*, према енгл. *Hidden Markov Model*) представља вероватносни 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}).



Електронска лекција

- Лекција проширује десето поглавље књиге/уџбеника *Bioinformatics Algorithms: An Active Learning Approach*.
- Резултат је *Jupyter* свеска са *Python* кодовима.
- Лекција је јавно доступна на *GitHub* репозиторијуму.

Погађање фенотипа

- Код лечења ХИВ-а, значајно је да ли изолат којим је пацијент заражен ствара синцицијум – нефункционалну вишеједарну цитоплазматичну масу са заједничком ћелијском мембраном.
- Испоставља се да је примарна структура гликопротеина омотача *gp120*, конкретно *V3* петље, важан предиктор фенотипа.
- Електронска лекција – имплементирано правило 11/25.
- Проблем је како прецизно лоцирати (поравнати) геном новог изолата, како би правило могло да се примени.

Потрага за генима

- Већина нуклеотида ДНК не кодира протеине, па је један од важних биолошких проблема управо проналажење места на којима се гени налазе.
- Гене има слисла тражити у стабилним регионима ДНК, који нису подложни метилацији. Такви региони називају се *CG* острвима или *CpG* местима.
- Електронска лекција – имплементиран прозорски приступ.
- Проблем је како одредити ширину прозора и обрадити преклапајуће прозоре.

Коцкање са јакузама

- Једноставан мотивациони пример за увођење *HMM* је непоштена коцкарница – крупије баца новчић, који у сваком тренутку може бити праведан или отежан (два потенцијална новчића).
- Могуће је опазити само резултат бацања, а задатак је на основу тога одредити највероватнији низ коришћених новчића.
- Електронска лекција – имплементиран прозорски приступ.
- Проблеми прозорског приступа остају нерешени.

Дефиниција модела

- Крупије се, уместо као особа, може схватити као аутомат, за који се испоставља да у потпуности одговара појму скривеног Марковљевог модела, као уређене петорке из увода.
- Основну дефиницију погодно је надградити, нпр. увођењем експлицитног почетног стања, заменом матрица мапама или употребом логаритмованих вероватноћа.
- Електронска лекција – имплементирана класа која представља допуњени *HMM* и приказана њена улога у моделовању непоштене коцкарнице.

Могућности модела

- Могуће је дефинисати појам скривеног пута $p = p_1 ... p_k$ као низ k стања кроз која *HMM* пролази, а да притом емитује секвенцу опсервација $o = o_1 ... o_k$. Главна идеја је анализирати у ком су односу p и o, те са којом се вероватноћом реализују.
- Једноставним формулама рачунају се вероватноћа пута $P(p) = \prod_{i=1}^{k} a_{p_{i-1},p_i}$, вероватноћа $P(o|p) = \prod_{i=1}^{k} a_{p_{i-1},p_i} \cdot b_{p_i,o_i}$ исхода на путу, те заједничка вероватноћа пута и исхода $P(p, o) = P(p)P(o|p)$, чијом се максимизацијом за познато o (низ исхода бацања) добија највероватније p (низ новчића).
- Електронска лекција – имплементирано израчунавање вероватноћа према формулама и наивна максимизација грубом силом.

Витербијев алгоритам

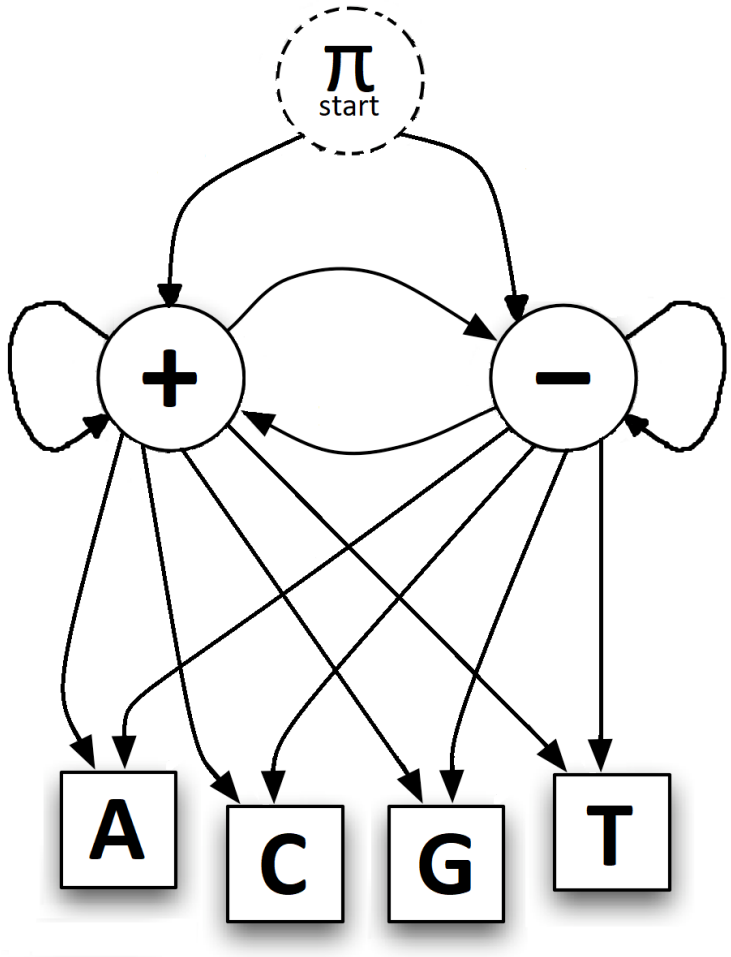
- Наивни приступ је експоненцијалне сложености, па се максимизацији (декодирању) приступа Витербијевим алгоритмом, техником динамичког програмирања заснованом на Витербијевом графу.
- Овај граф моделује све путеве кроз *HMM* истовремено, а осмишљен је на основу основног временског својства *HMM*, према коме текуће стање зависи искључиво од првог претходног.
- Електронска лекција – имплементирана максимизација претходно разматраних вероватноћа $P(p)$, $P(o|p)$, $P(p, o)$, као и $P(p|o)$.

Алгоритам „напред”

- Могуће је моделовати и појединачну расподелу вероватноће опажања $P(o)$, која једина досад није разматрана.
- Ако се примети да Витербијев алгоритам израчунава $\max_p P(p, o)$, а да је вероватноћа опажања $P(o) = \sum_p P(p, o)$, лако је закључити како се израчунава $P(o)$ преко Витербијевог графа.
- Електронска лекција – имплементирано израчунавање вероватноће $P(o)$ и њена максимизација, као и израчунавање $P(p|o)$, чиме је модел комплетиран.

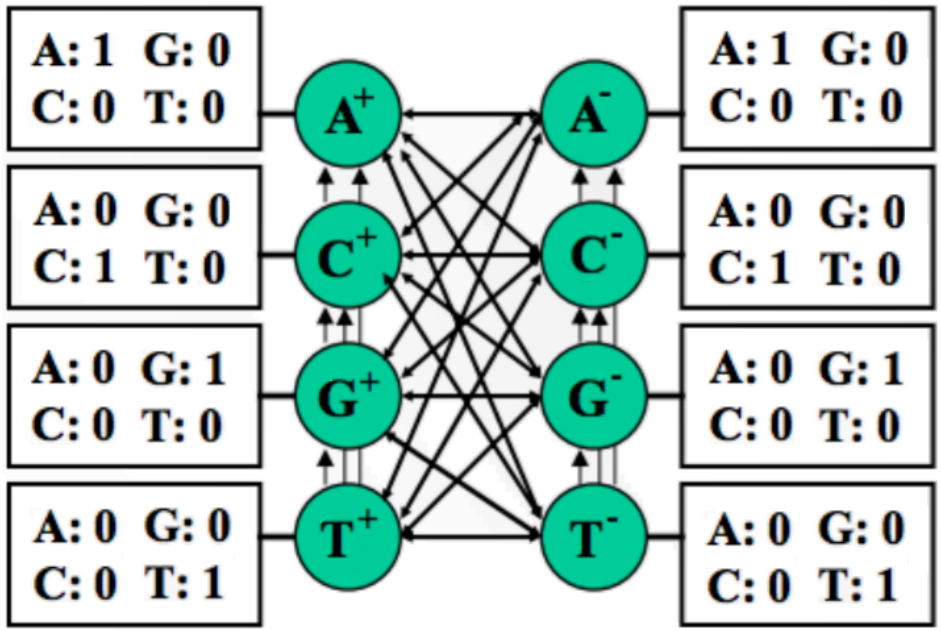
Гени – два стања

- Улазни низ нуклеотида посматра се као секвенца опажања коју треба декодирати.
- Параметри модела одређују се на основу знања из генетике или емпиријски.
- Најуспешнији је модел који посматра динуклеотиде уместо појединачне симболе.
- Електронска лекција – упоређени различити модели.



Гени – више стања

- *CG* острва и региони ван њих могу се моделовати као два одвојена Марковљева ланца, која се могу спојити у *HMM* са осам скривених стања.
- Свако стање представља емисију одговарајућег нуклеотида у неком региону.
- Електронска лекција – упоређени различити модели.

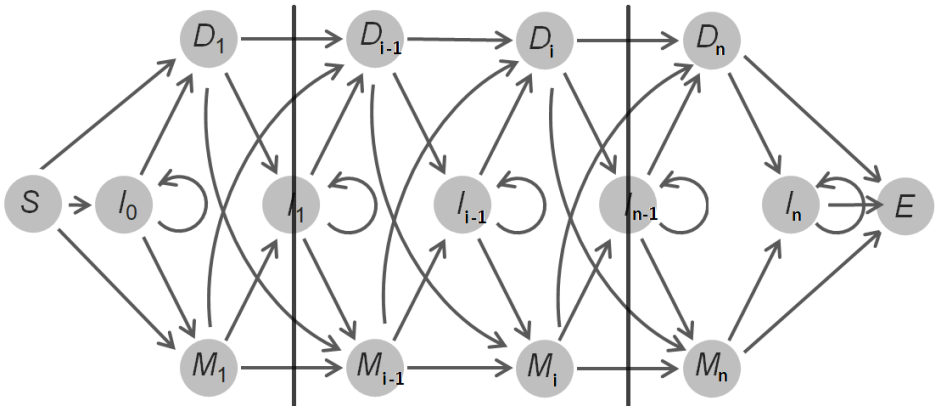


Профилни модели

- Протеини су организовани у разнолике протеинске фамилије, а чест биолошки задатак јесте додељивање новооткривеног полипептида некој од познатих фамилија.
- Користан алат за класификацију протеина јесу профилни *HMM* или *HMM* профили, који статистички описују фамилије протеина, а граде се на основу вишетруког поравнања.
- Идеја класификације је да се нови полипептиди декодирају профилним моделима неких фамилија, а затим одабере профил у односу на који је вероватноћа припадности изолата највећа или макар прелази предефинисану границу.

Рад са профилима

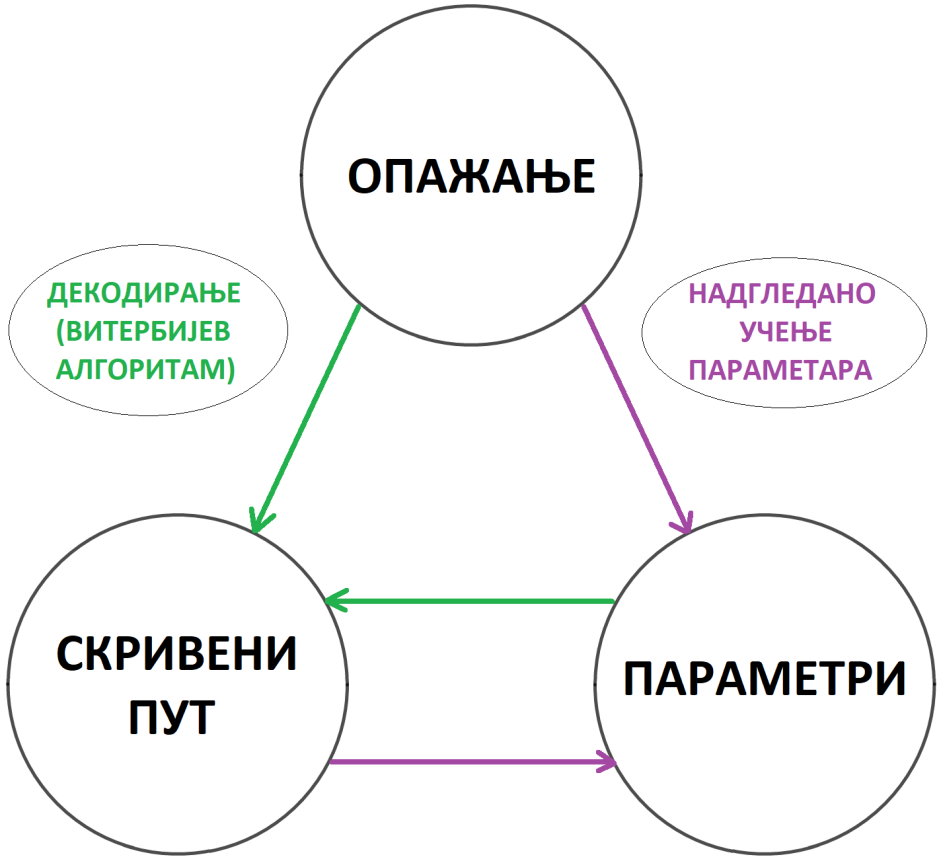
- Профилни *HMM* имају три типа стања, која одговарају мутацијама које настају у геному, уз поклапања.
- Параметри модела одређују се емпиријски, по улазном вишеструком поравнању.



- Електронска лекција – имплементирана класа која представља профилни *HMM* и приказана њена улога у раду са секвенцама.

Витербијево учење

- Још једна важна способност *HMM* јесте то да је могуће научити све параметре модела само на основу опажања.
- Пре таквог ненадгледаног учења, параметри се могу научити и надгледано (на основу опажања и пута).
- Надгледано учење одређује вероватноће на основу фреквенција, док ненадгледано представља алгоритам максимизације очекивања.



Баум-Велчово учење

- Основна верзија ненадгледаног учења параметара је Витербијево учење, али се чешће користи оптималније Баум-Велчово учење.
- Декодирање у кораку очекивања мења се новим алгоритмом заснованим на Витербијевом графу: „напред-назад". Алгоритам одређује вероватноћу да је *HMM* у неком тренутку био у неком скривеном стању.
- Надгледано учење у кораку максимизације мења се сумирањем одговарајућих индикатора.
- Електронска лекција – имплементирани сви типови учења, као и повезани концепти, попут „меког" и апостериорног декодирања.

Закључак

- У раду је изложен појам скривених Марковљевих модела, као и њихов биоинформатички значај. Дата је детаљна мотивација за увођење статистички поткованог аутомата, након чега је појам *HMM* разрађен и примењен на решавање биолошких проблема.
- Суштински најзначајнији допринос рада је електронска лекција, која, уз детаљну теоријску позадину, садржи и многобројне имплементације. Замисао јој је да допринесе усвајању знања о скривеним Марковљевим моделима и њиховој примени у биоинформатици, а притом буде јавно и свима доступна.

ХВАЛА НА ПАЖЊИ!

Питања?

Библиографија


Лазар Васовић, *HMM u bioinformatici*, 09 2021, GitHub, репозиторијум са електронском лекцијом доступан на: `https://github.com/matfija/HMM-u-bioinformatici`.


Phillip Compeau and Pavel Pevzner, *Bioinformatics Algorithms: An Active Learning Approach, 2nd Edition, Vol. II*, Active Learning Publishers, LLC, 2015, званични сајт књиге/уџбеника: `https://www.bioinformaticsalgorithms.org/`.