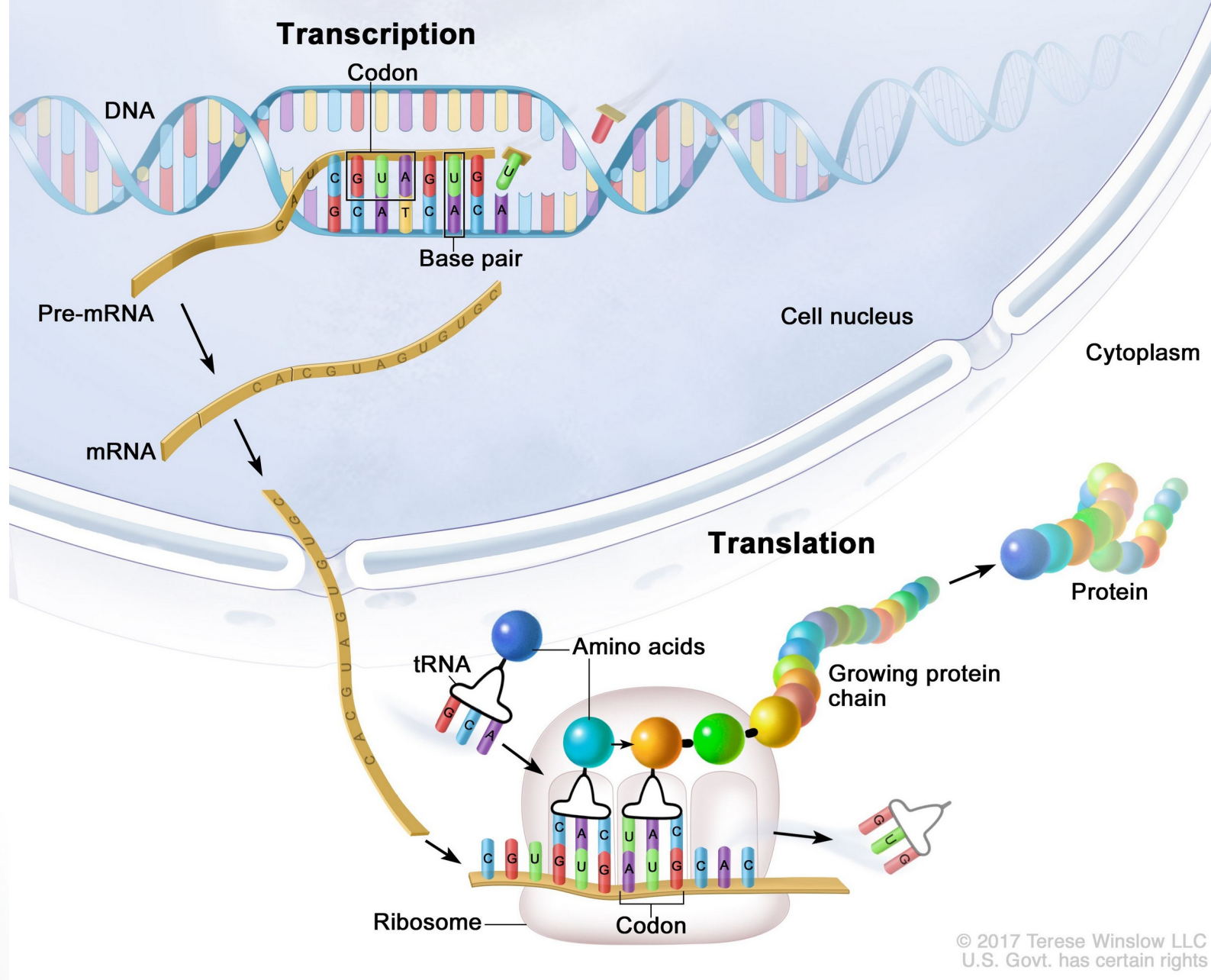
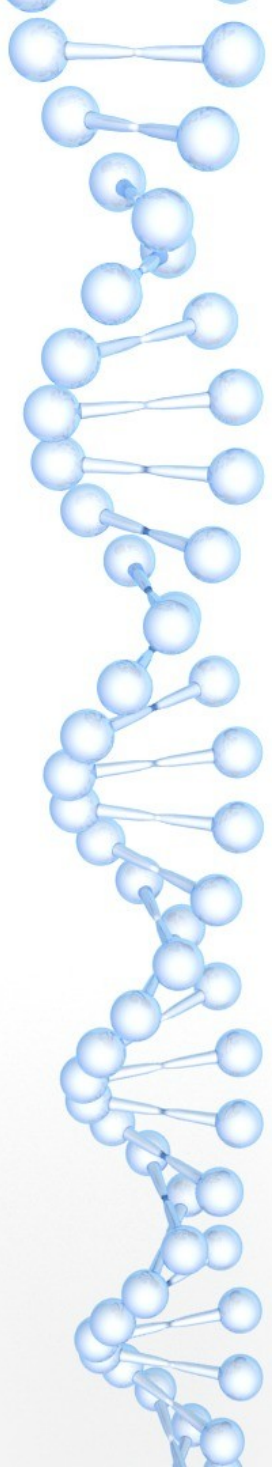


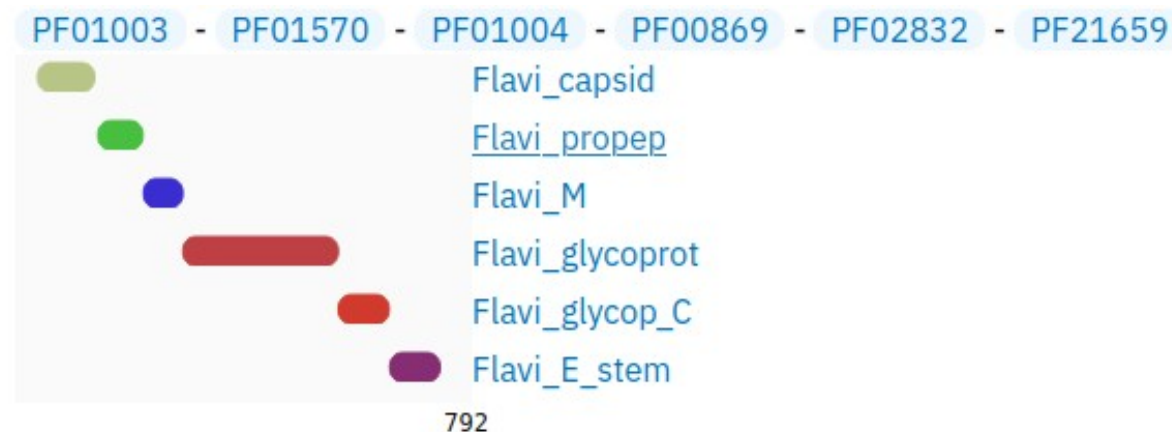
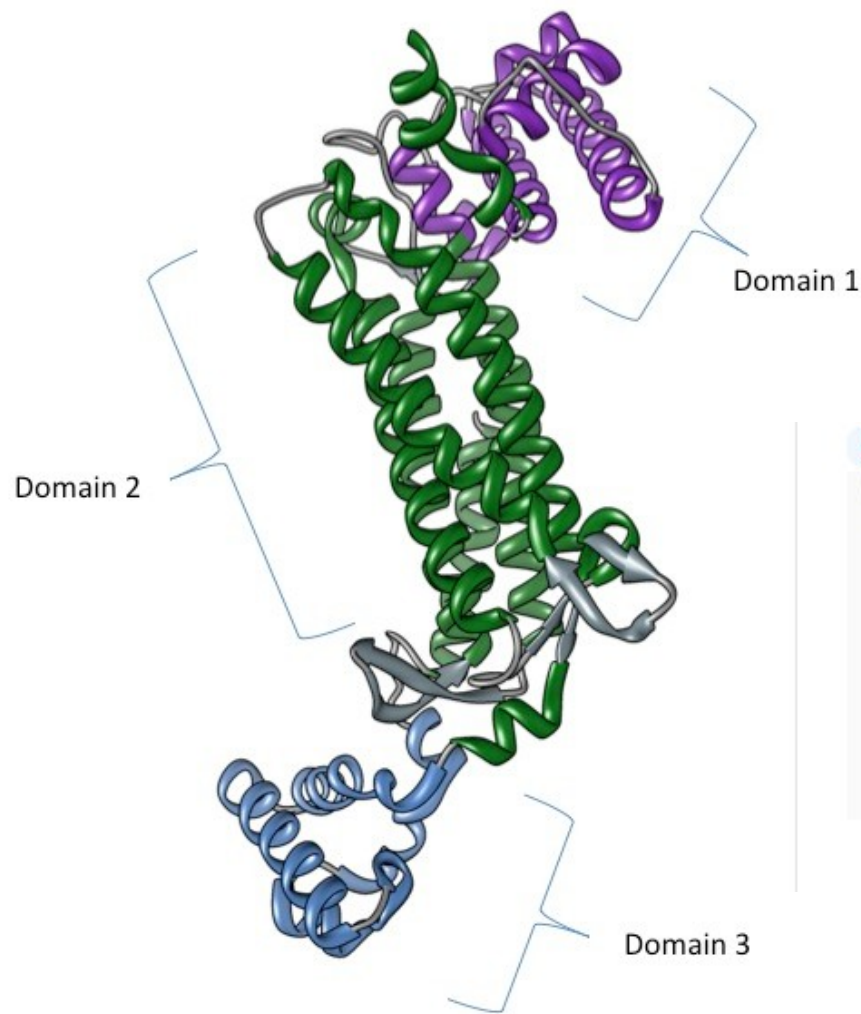
# Implementacija skrivenih Markovljevih modela u domenskoj klasifikaciji proteinskih sekvenci

- **Pristupnik:** Domagoj Sviličić (0036540224)
- **Mentor:** doc. dr. sc. Krešimir Križanović



- A** – Alanin
- C** – Cistein
- D** – Asparaginska  
kiselina
- E** – Glutaminska  
kiselina
- F** – Fenilalanin
- G** – Glicin
- H** – Histidin
- I** – Isoleucin
- K** – Lizin
- L** – Leucin
- M** – Metionin
- N** – Asparagin
- P** – Prolin
- Q** – Glutamin
- R** – Arginin
- S** – Serin
- T** – Treonin
- V** – Valin
- W** – Triptofan
- Y** – Tirozin

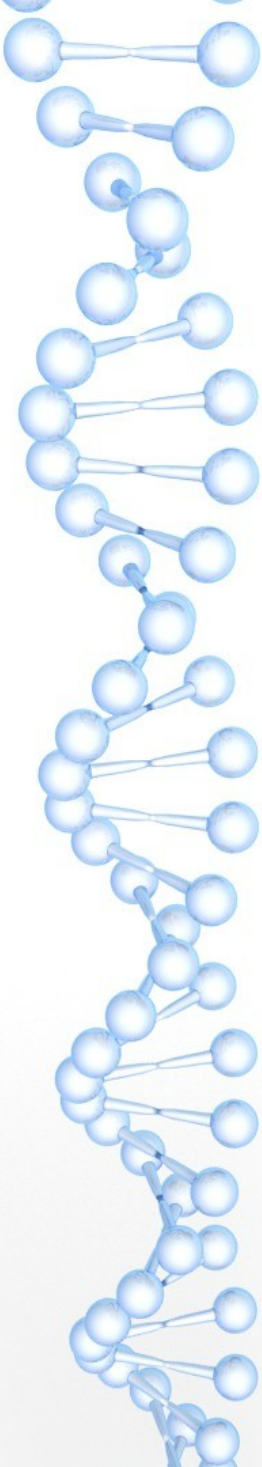
**CILJ:** Odrediti koje se domene pojavljuju na nepoznatoj proteinskoj sekvenci kako bi predvidjeli potencijalne funkcije dotičnog proteina (biotehnologija, medicina, farmacija)



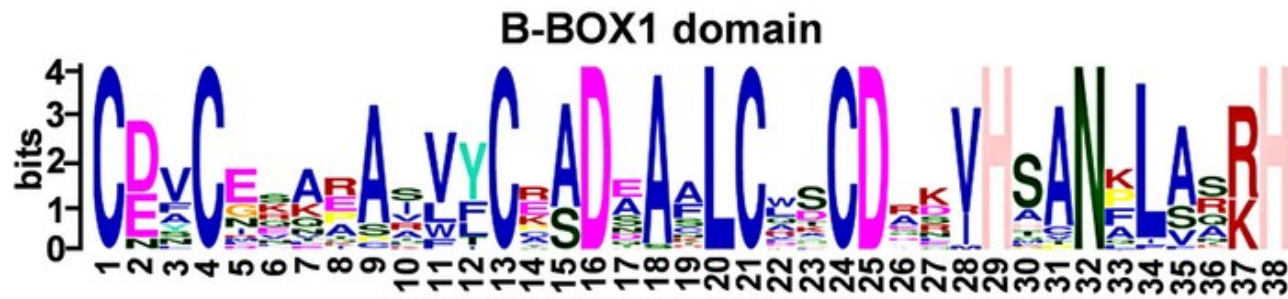
Domena - podniz proteinske sekvence odgovoran za pojedinu funkciju

**Proteinska sekvenca:** ACDRYXGFPMHCDKLVLKNDSPYWTARYW...





A



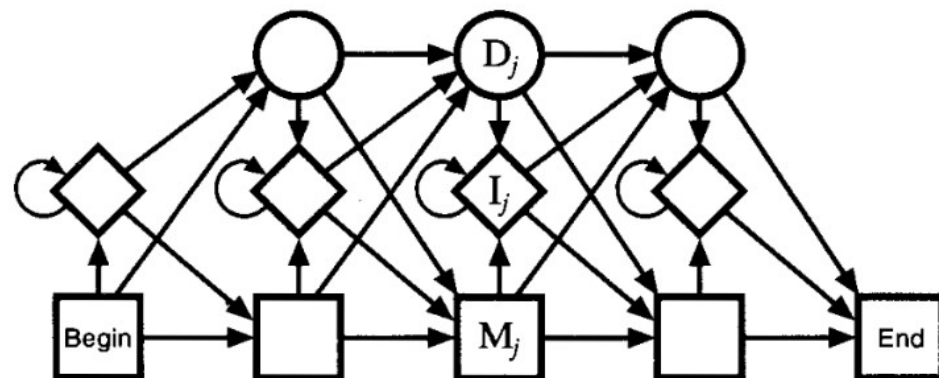
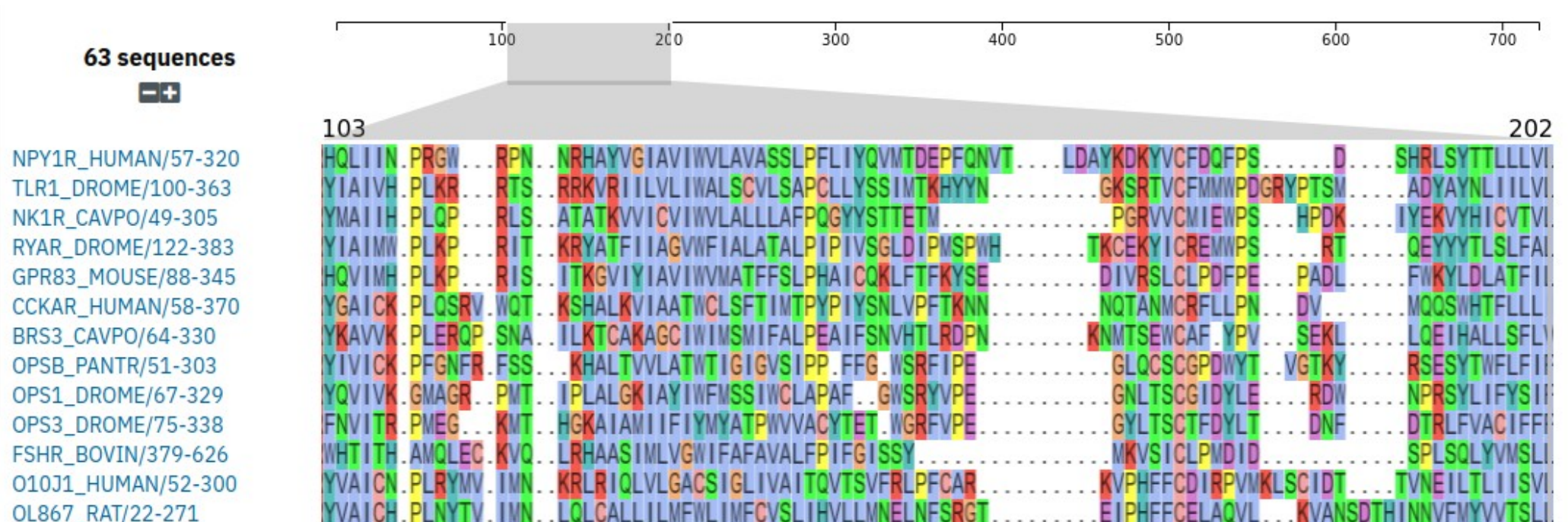
B

	B-BOX1 domain	
SIBBX1	CDSRSVTCTIYCQADSAYLOADCDARIEAASIVTSRE	58
SIBBX3	CDSCHSATCTIVYCRADSAYLOAGCDARIEASIMASRE	51
SIBBX2	CDSRSTACAVYCRADSSFLOAGCDTRMEANLIASRE	58
SIBBX4	CDACKATPSTVECKADNAFLQLCDSKIEAANKIASRE	50
SIBBX5	CDSKTSFATVECRADSAFLQLGCDCKIEAANKIASRE	50
SIBBX6	CEYCHLAAALVECRDNTFTVOLSCTRLEAR.....E	58
SIBBX18	CDVCSAAAILECAADEAALORACDEKVEMCNKIASRE	42
SIBBX19	CDVCSAAAILECAADEAALORSCEKVELCNKIASRE	42
SIBBX20	CDVKNKEAIVECTADEAALODDCHRVHVNKIASRE	42
SIBBX21	CDVNNNEASVECVADEAALODSCHRVHANKIASRE	42
SIBBX25	CDVDKEEASVYCSADEATLCQSCDYQVHANKIASRE	42
SIBBX22	CNVCEVAEANVLCCADEAALQWSCDEKVEAANKIASRE	42
SIBBX23	CNACEVAEAKVLCCADEAALQWYCDKVEAANKIANKE	42
SIBBX24	CDVCEKAQATVICCDADEAALQAKCDIEVEAANKIASRE	42
SIBBX11	CDECGNNTALLYCRADSAKLOFTCDREVESTNQFTKE	53
SIBBX12	CDEFNQQIAVLYCRADIAKLOLECDQIVESANALSKE	46
SIBBX8	CEFCGEQRSIVYCRSDAACLQLSCDRNVESANALSQRE	39
SIBBX9	CEYCGEQRSIVYCRSDAACLQLSCDRNVESANALSQRE	42
SIBBX7	CEYCGEQRSIVYCRSDAACLQLSCDRNVESANALSQRE	42
SIBBX10	CDLCEVRAVVYCKSDSARLQLQCDYVESFNLSRRE	42
SIBBX27	CEFCMLLKFFVYCEADAHLQLSCDAKVESANALSNRE	44
SIBBX13	CDNCIRKFRARWYCAADDAFLCQSCDSSVESANPIARRE	56
SIBBX15	CDNCIKKFRARWYCEADDAFLCQNCIASVESANPIARRE	57
SIBBX14	CDSCLSKFRARWYCEADDAFLCQSCDVSVIESANQIASRE	57
SIBBX16	CELCK.SEAYVYCEADNAFLQKKCDKIVETANFFAQRE	58
SIBBX17	CALCS.SEASVYCEADNAFLQKKCDRSVEGANFLAQRE	71
SIBBX28	CELON.GIARIYCESDHANLQWCDLKVESANFIVAKE	40
SIBBX30	CELCT.GIARMYCESDNASLQWCDLKVESANFLAARE	40
SIBBX26	CELCARVYKAVYKESDASLQWCDLKVESANFLAARE	44

- **Multiple Sequence Alignment (MSA)** algoritam
- „Seed alignment” - skup poravnatih reprezentativnih podnizova koji predstavljaju jednu proteinsku domenu (Pfam baza)
- Praćenje redosijeda i frekvencije aminokiselina po pozicijama



# Konstrukcija profilnog HMM-a

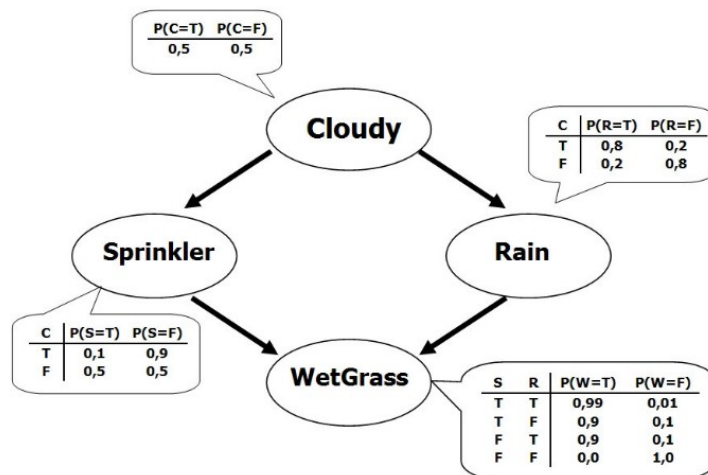


**M** – Match  
(podudaranje)

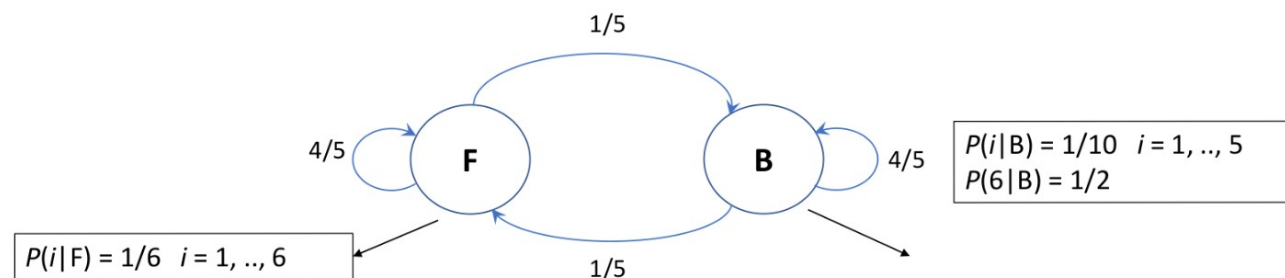
**I** – Insert (umetanje)

**D** – Delete (brisanje)

# Bayesovska mreža vs. HMM



- modelira kauzalne odnose



$$S = \{F, B\}$$

$$E = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/10 & 1/10 & 1/10 & 1/10 & 1/10 & 1/2 \end{pmatrix}$$

$$\Pi = \{1/2, 1/2\}$$

$$O = \{1, 2, 3, 4, 5, 6\}$$

$$A = \begin{pmatrix} 4/5 & 1/5 \\ 1/5 & 4/5 \end{pmatrix}$$

- modelira slijed skrivenih stanja i vidljivih promatranja
- sekvencijalni podaci



# HMM-algoritmi

- **Forward** (omogućuje računanje vjerojatnosti promatranog niza s obzirom na model)
- Baum Welch, Viterbi, backward algoritam

**Forward algorithm:**

*for*  $i \in S$  // initialization

$$\alpha_{\theta}(i) = \pi_i \cdot e_{io_0}$$

*for*  $t = 1$  *to*  $T$

*for*  $i \in S$

$$\alpha_t(i) = e_{io_t} (\sum_{j \in S} \alpha_{t-1}(j) a_{ji}) \text{ // prob. at time } t \text{ and state } i \text{ emitting } e_{io_t}$$

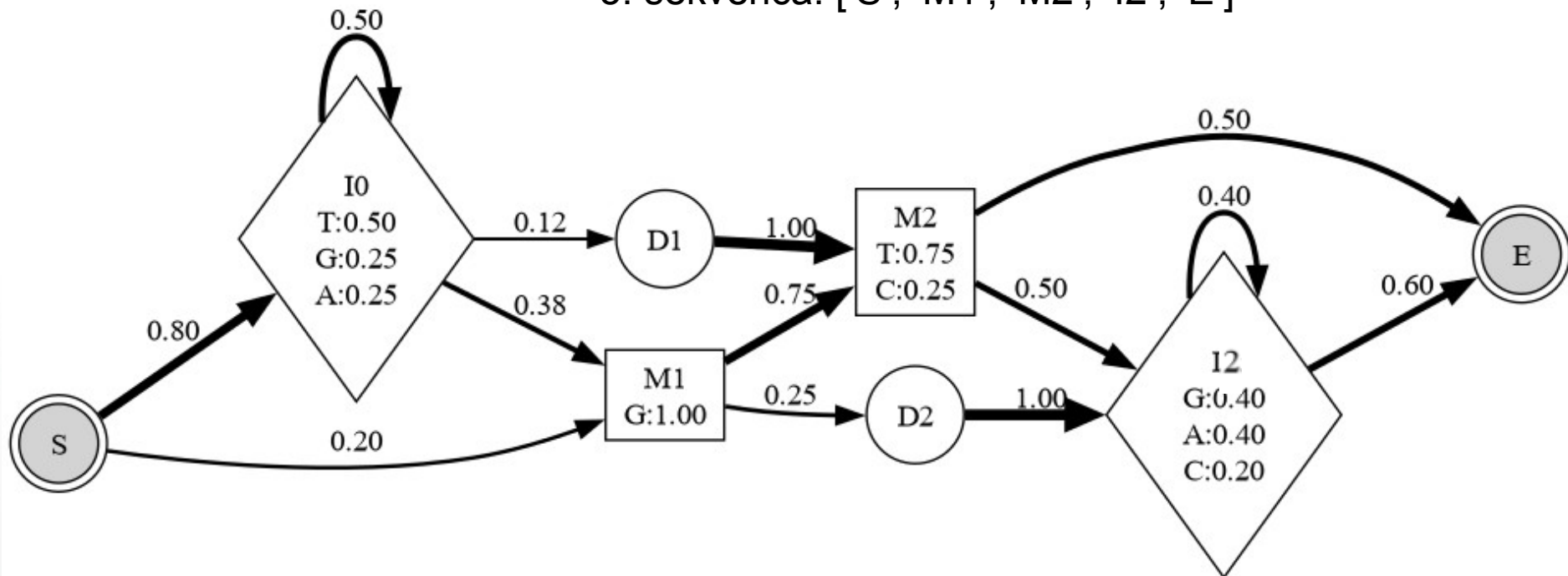
**return**  $\sum_{i \in S} \alpha_T(i)$



# Postupak konstrukcije HMM-a

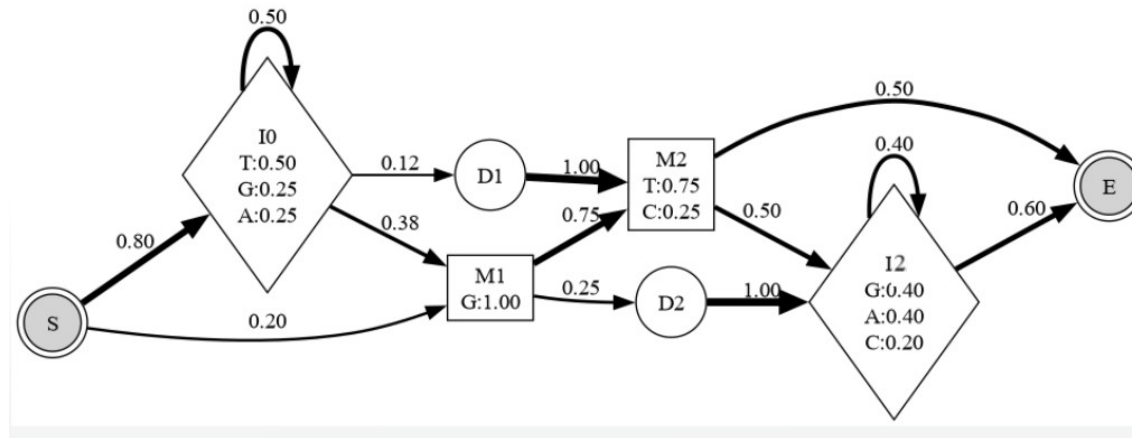
-GAT-T---  
T---GTG--  
TG-TG-GAA  
--A-GC---  
----GTC--  
\*\*

- 1. sekvenca: ['S', 'I0', 'I0', 'I0', 'D1', 'M2', 'E']
- 2. sekvenca: ['S', 'I0', 'M1', 'M2', 'I2', 'E']
- 3. sekvenca: ['S', 'I0', 'I0', 'I0', 'M1', 'D2', 'I2', 'I2', 'I2', 'E']
- 4. sekvenca: ['S', 'I0', 'M1', 'M2', 'E']
- 5. sekvenca: ['S', 'M1', 'M2', 'I2', 'E']

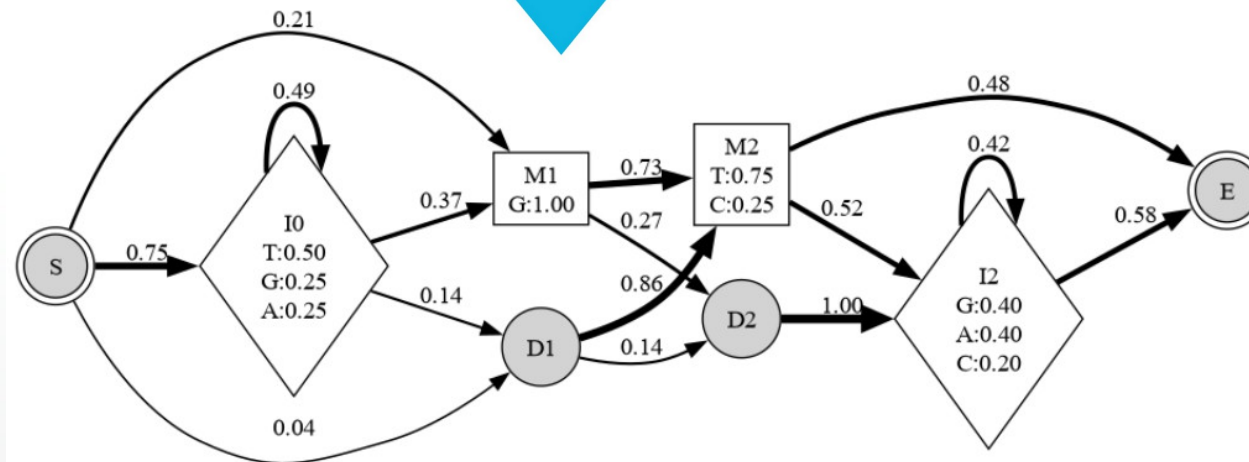




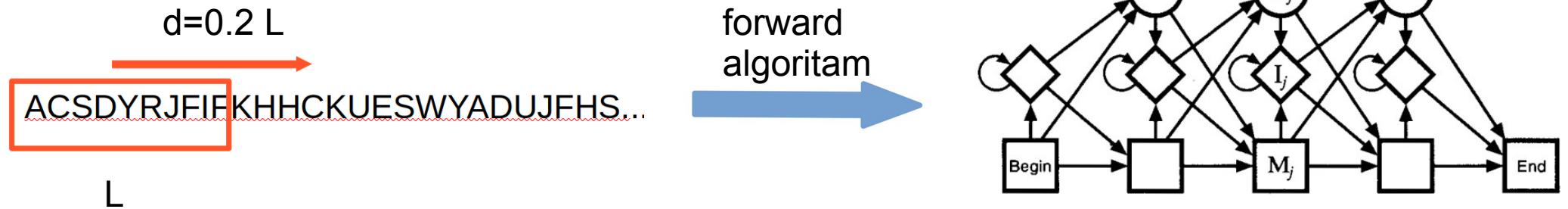
# Laplaceovo zagladivanje



$$P(x|y) = \frac{N(x, y) + \alpha}{N(y) + \alpha \cdot |X|}$$



# Korištenje posmačnog prozora



- duljina prozora pojedinog HMM-a određena preko prosječne duljine slučajno generirane sekvence (Viterbi)
- posmak prozora jednak je 20% duljine prozora



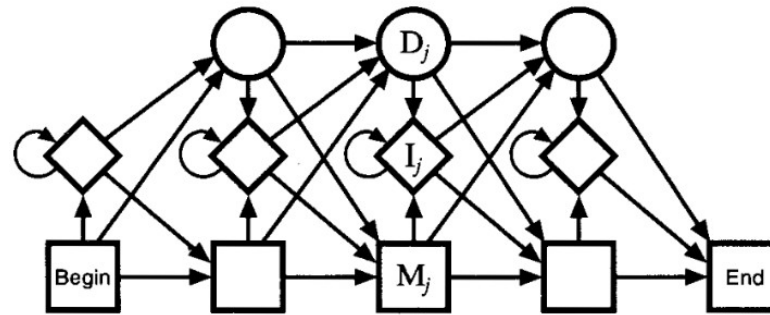
# Postupak klasifikacije

- 1) Konstrukcija 10 profilnih HMM-a iz 10 pripadnih „seed alignment” prikaza različitih domena poravnatih MSA algoritmom
- 2) Nepoznatu proteinsku sekvencu ispitujemo prolaskom posamačnog prozora za svaki od 10 profilnih HMM-ova
- 3) Sadržaj posamačnog prozora pri svakom pomaku prosljeđujemo forward algoritmu i računamo log-odds
- 4) Najveći log-odds daje HMM koji je konstruiran nad „seed” poravnanjem odgovarajuće domene koja je pronađena u sekvenci



# Računanje log-odds vrijednosti

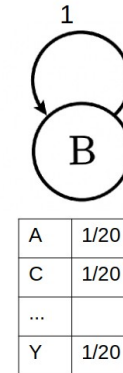
Profilni HMM za neku domenu



forward  
algoritam



Pozadinski trivijalni model

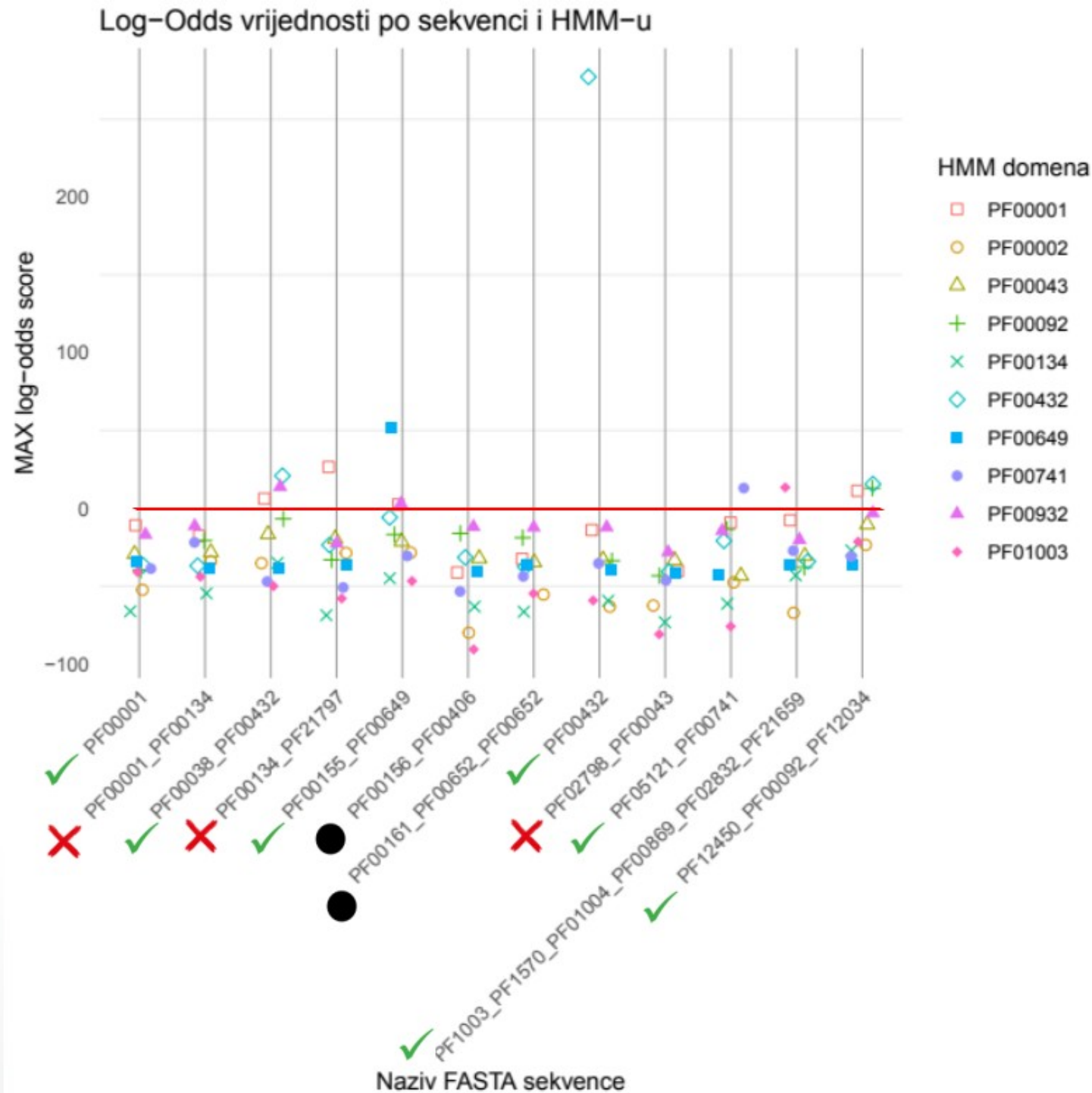


forward  
algoritam

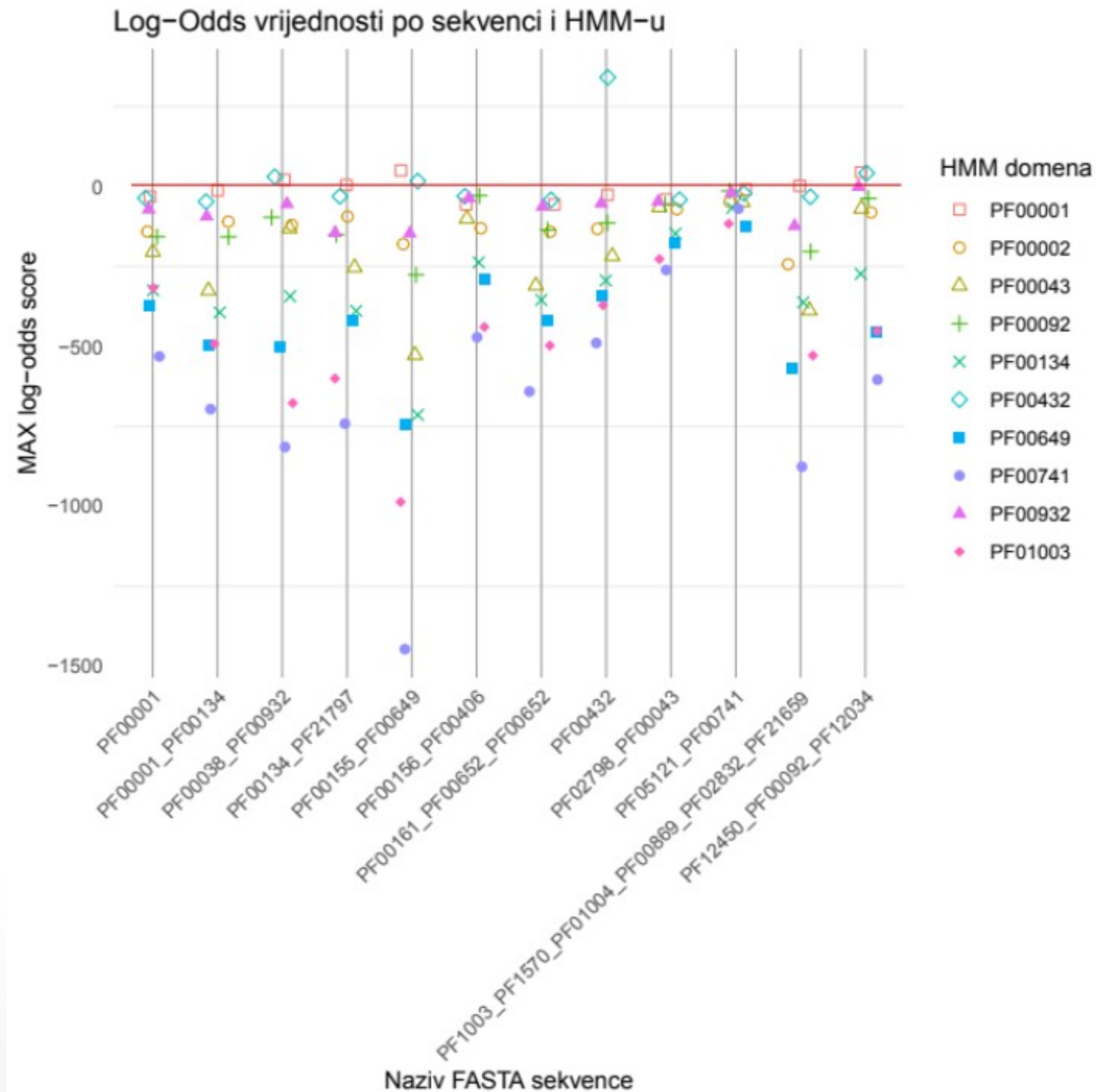
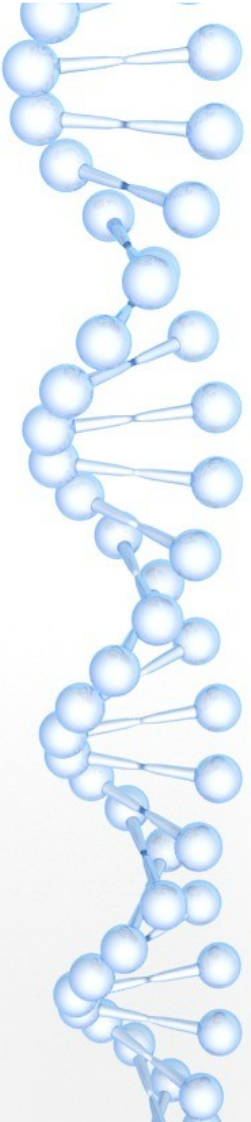


$$\text{log-odds} = \log(P(X|HMM)) - \log(P(X|\text{pozadinski model}))$$

# Analiza rezultata



# Analiza rezultata – bez korištenja posmačnog prozora

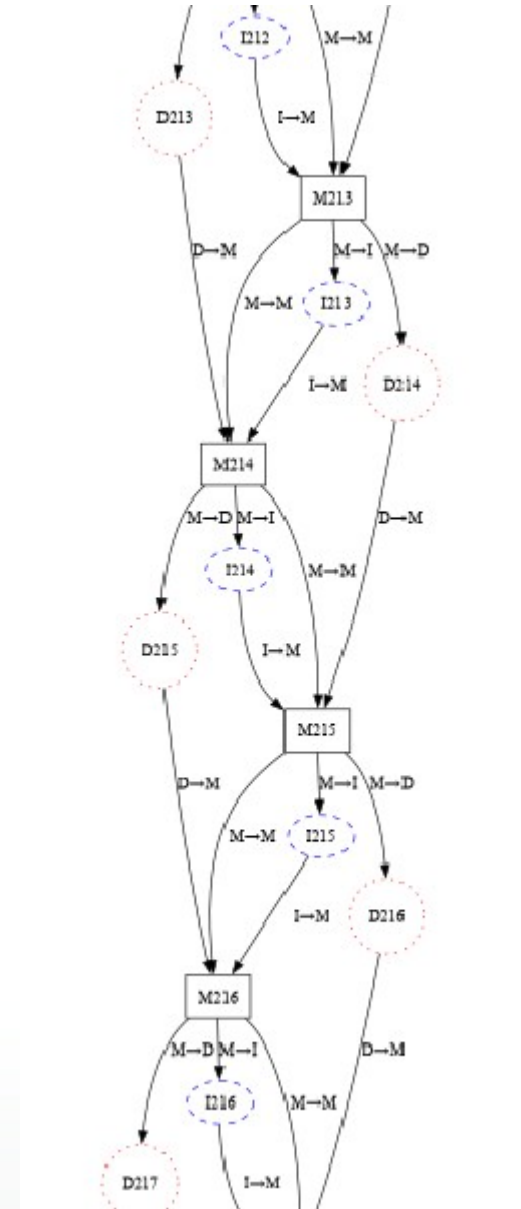




# HMMER

```
1 #
2 # target name      accession  tlen query name      accession  qlen  --- full sequence --- ----- this domain -----
3 # -----
4 7tm_1              PF00001.26  260 A0A821XNW9          -          782   1.4e-62  197.6  12.8   1   2   1.4e-62  1.4e-62  197.6  12.8
   receptor (rhodopsin family)
5 7tm_1              PF00001.26  260 A0A821XNW9          -          782   1.4e-62  197.6  12.8   2   2       0.18    0.18   -2.7   2.2
   receptor (rhodopsin family)
6 #
7 # Program:         hmmscan
8 # Version:         3.4 (Aug 2023)
9 # Pipeline mode:   SCAN
10 # Query file:      A0A821XNW9.fa
11 # Target file:     binaries/PF00001.hmm
12 # Option settings: hmmscan --domtblout rezultati.tbl binaries/PF00001.hmm A0A821XNW9.fa
13 # Current dir:     /home/domagoj/Desktop/Seminar2_Implementation_of_Hidden_Markov_Models_in_the_Phylogenetic_Classification_of_Protein_Sequer
14 # Date:            Fri May 16 12:37:25 2025
15 # [ok]
16
17
18
```

# HMMER- vizualizacija HMM-a





# Zaključak

- HMM-ovi – pogodni za ugradnju vanjskog domenskog znanja
- mogućnost treniranja na inicijalno malim količinama podataka
- ovisnost evaluacije o eksperimentalnom postavu