האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

אלגוריתמים בביו' חישובית

76558

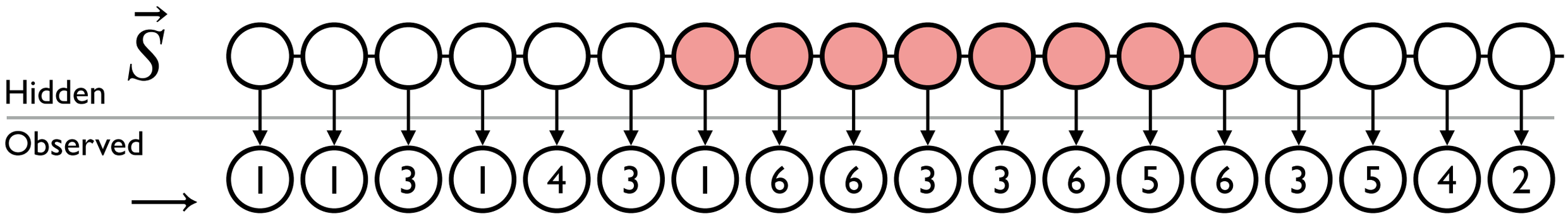
זיהוי מוטיבים בעזרת מודלים מרקוביים חבויים

תומי קפלן

30/1/2024

שרשראות מרקוב חבויות

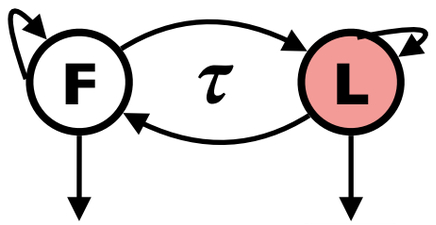
- משתני התהליך המרקובי הם חבויים
- פולטים [בהתאם למצב] משתנים אחרים [ניצפים]



$S_i \in \{\boxed{\text{F}}, \boxed{\text{L}}\}$

$\tau = P(S_i \mid S_{i-1})$

τ	F	L
F	99%	1%
L	5%	95%



1/6	.1
1/6	.1
1/6	.1
1/6	.1
1/6	.1
1/6	.5

$\forall_k,\ \mathsf{e}_F(k) = \frac{1}{6}$

$\mathsf{e}_L(k) = \begin{cases} \frac{1}{10} & k = 1,\ldots,5 \\ \frac{1}{2} & k = 6 \end{cases}$

$X_i \in \{1,2,3,4,5,6\}$

מה עוד היינו רוצים לדעת?

- פוסטריור על מצב חבוי

$$P(S_i \mid X_1, \ldots, X_n)$$

האם סביר שעברנו במצב מסויים בזמן i

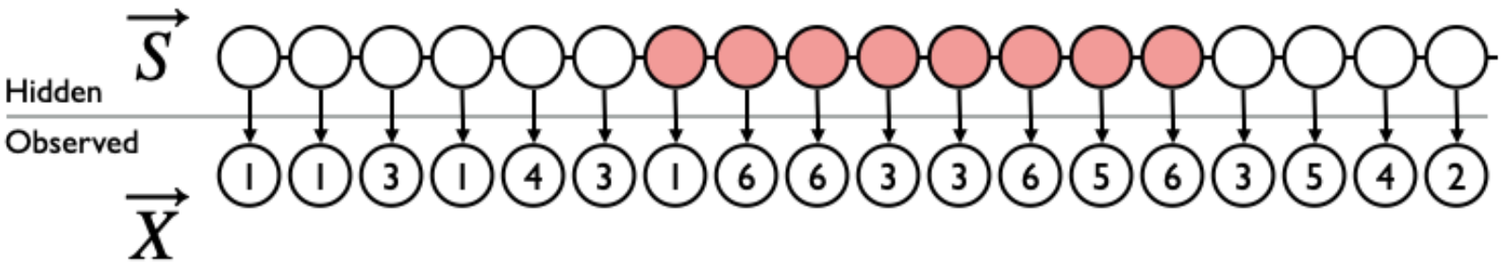
- פוסטריור על מעבר חבוי

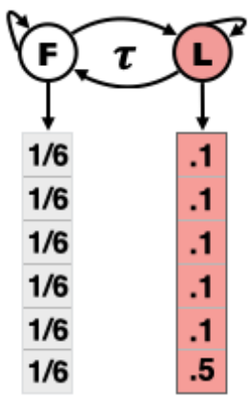
$$P(S_{i-1}, S_i \mid X_1, \ldots, X_n)$$

- מסלול ויטרבי

$$\arg\max_{S_1,\ldots,S_n} P(S_1, \ldots, S_n, X_1, \ldots, X_n)$$

MPE / decoding





חישוב נראות: אלגו' Forward

- נגדיר:

$$F_k(i) = P(X_1, \ldots, X_i, S_i = k)$$

- נפתח רקורסיבית את

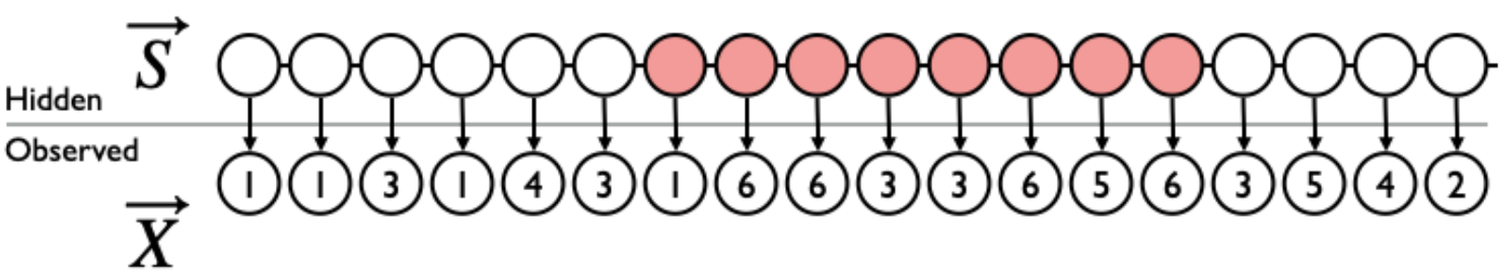
$$F_l(i+1) = P(X_1, \ldots, X_i, X_{i+1}, S_{i+1} = l)$$

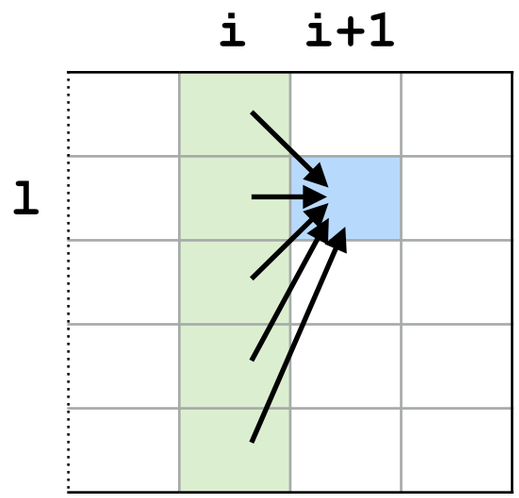
$$= \sum_k P(X_1, \ldots, X_i, X_{i+1}, S_i = k, S_{i+1} = l)$$

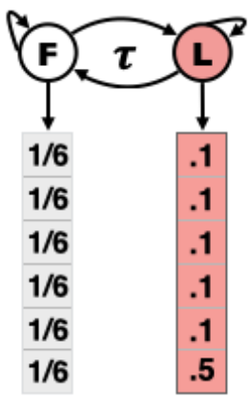
$$= \sum_k P(X_1, \ldots, X_i, S_i = k) \cdot P(S_{i+1} = l \,|\, X_1, \ldots, X_i, S_i = k) \cdot P(X_{i+1} \,|\, X_1, \ldots, X_i, S_i = k, S_{i+1} = l)$$

$$= \sum_k \left[P(X_1, \ldots, X_i, S_i = k) \cdot P(S_{i+1} = l \,|\, S_i = k)\right] \cdot P(X_{i+1} \,|\, S_{i+1} = l)$$

$$= \sum_k \left[F_k(i) \cdot \tau_{k,l}\right] \cdot \mathsf{e}_l(X_{i+1})$$







חישוב נראות: אלגו' Backward

- נגדיר:

$$B_k(i) = P(X_{i+1}, \ldots, X_n \,|\, S_i = k)$$

- נפתח רקורסיבית את

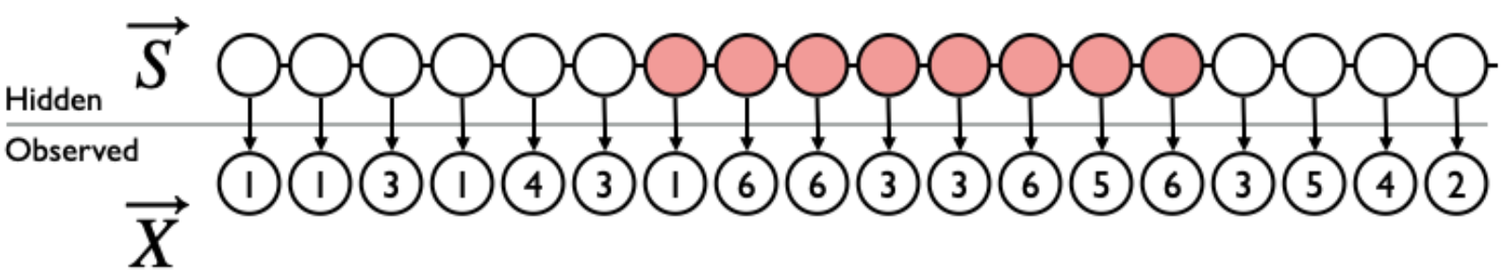
$$B_l(i-1) = P(X_i, X_{i+1}, \ldots, X_n \,|\, S_{i-1} = l)$$

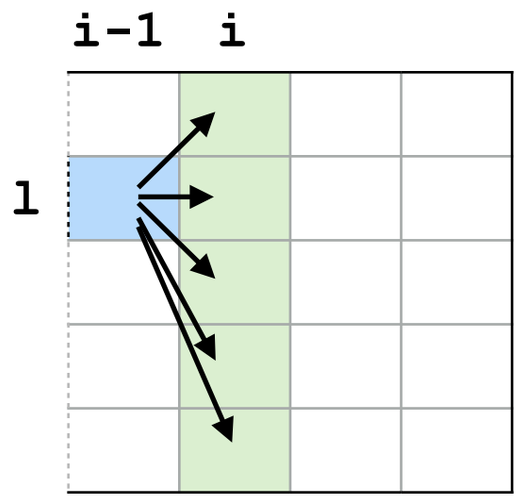
$$= \sum_k P(S_i = k, X_i, X_{i+1}, \ldots, X_n \,|\, S_{i-1} = l)$$

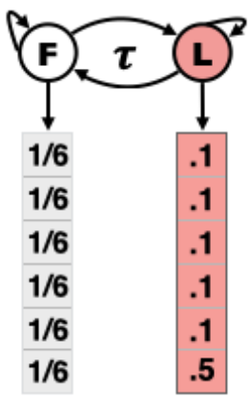
$$= \sum_k P(S_i = k \,|\, S_{i-1} = l) \cdot P(X_i \,|\, S_i = k, S_{i-1} = l) \cdot P(X_{i+1}, \ldots, X_n \,|\, X_i, S_i = k, S_{i-1} = l)$$

$$= \sum_k P(S_i = k \,|\, S_{i-1} = l) \cdot P(X_i \,|\, S_i = k) \cdot P(X_{i+1}, \ldots, X_n \,|\, S_i = k)$$

$$= \sum_k \tau_{l,k} \cdot \mathrm{e}_k(X_i) \cdot B_k(i)$$







חישוב הפוסטריור

- ומכאן ש:

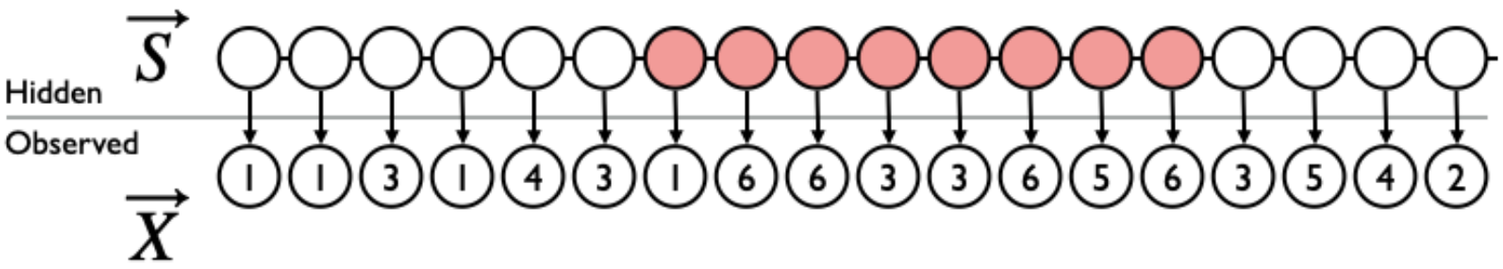
$$P(\overrightarrow{X}, S_i = k) = P(X_1, \ldots, X_n, S_i = k)$$
$$= P(X_1, \ldots, X_i, S_i = k) \cdot P(X_{i+1}, \ldots, X_n \mid X_1, \ldots, X_i, S_i = k)$$
$$= P(X_1, \ldots, X_i, S_i = k) \cdot P(X_{i+1}, \ldots, X_n \mid S_i = k)$$
$$= F_k(i) \cdot B_k(i)$$

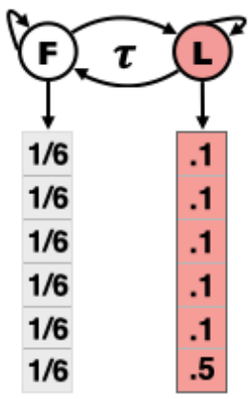
- ונוכל לחשב את הפוסטריור:

$$P(S_i = k \mid \overrightarrow{X}) = \frac{F_k(i) \cdot B_k(i)}{P(\overrightarrow{X})}$$

$$P(S_{i-1}=l, S_i=k \mid \vec{X}) = \frac{F_{i-1}(l) \cdot \tilde{T}_{lk} \cdot e_k(x_i) \cdot B_i(k)}{P(\vec{X})}$$

כדאי לי:
פוסטריור
מתברים.
תופן שלמה
המלוכ:





מציאת מסלול ויטרבי

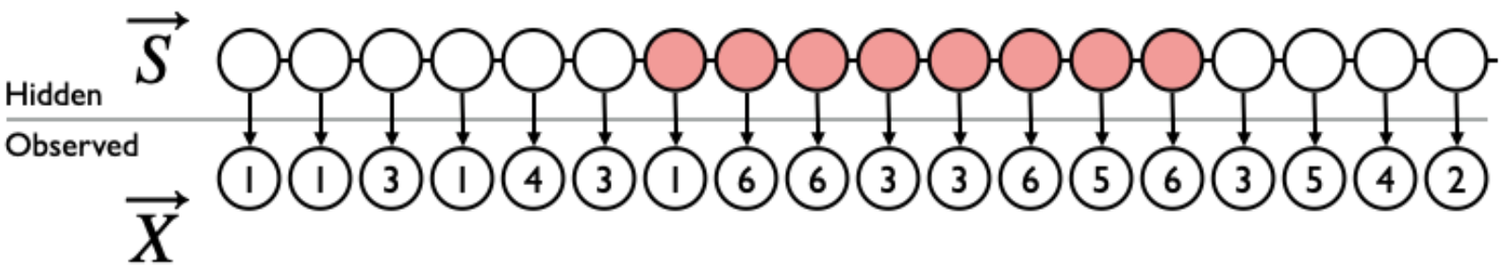
- איך נמצא את המסלול הסביר ביותר, בהנתן התצפיות?

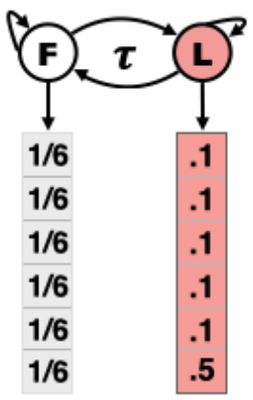
$$\pi^* = \arg\max_{\vec{S}} P(S_1, \ldots, S_n \,|\, X_1, \ldots, X_n) = \arg\max_{\vec{S}} P(\vec{S}, \vec{X})$$

למה?

- אלגו' תכנון דינאמי. נשמור בטבלה V את ההסתברות למסלול הסביר ביותר [עד i] המסתיים במצב $S_i=k$

$$V_k(i) = \max_{S_1,\ldots,S_{i-1}} P(S_1, \ldots, S_{i-1}, S_i = k, X_1, \ldots, X_i)$$



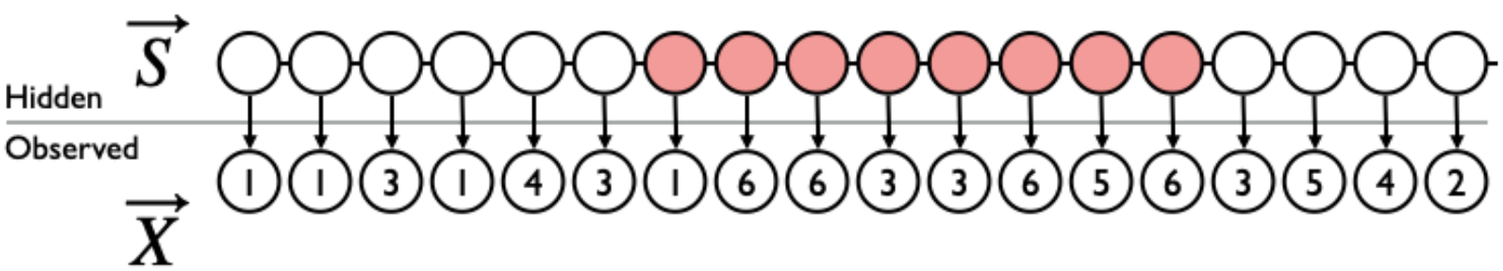


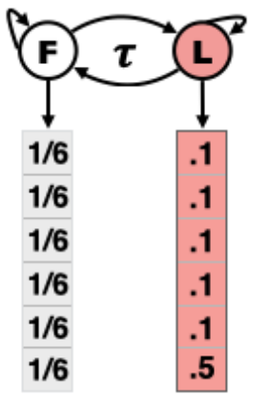
מציאת מסלול ויטרבי

- לפיכך $$V_k(n) = \max_{S_1,\ldots,S_{n-1}} P(S_1, \ldots, S_{n-1}, S_n = k, X_1, \ldots, X_n)$$
- ולכן $$\max_k V_k(n) = \max_k \max P(S_1, \ldots, S_n = k, X_1, \ldots, X_n) = \max_{\vec{S}} P(\vec{S}, \vec{X})$$
- כלומר נמצא את המצב האחרון במסלול האופטימלי

$$\pi^*[n] = \arg\max_k V_k(n)$$

- באופן דומה נוכל לשמור טבלה נוספת, לשיחזור המסלול





מציאת מסלול ויטרבי

Init: $\forall_k V_k(0) = 0$

For $i = 1 \ldots n$

For $k = 1 \ldots K$

$$V_k(i) = \mathsf{e}_k(X_i) \cdot \max_l \left[v_l(i-1) \cdot \tau_{l,k} \right]$$

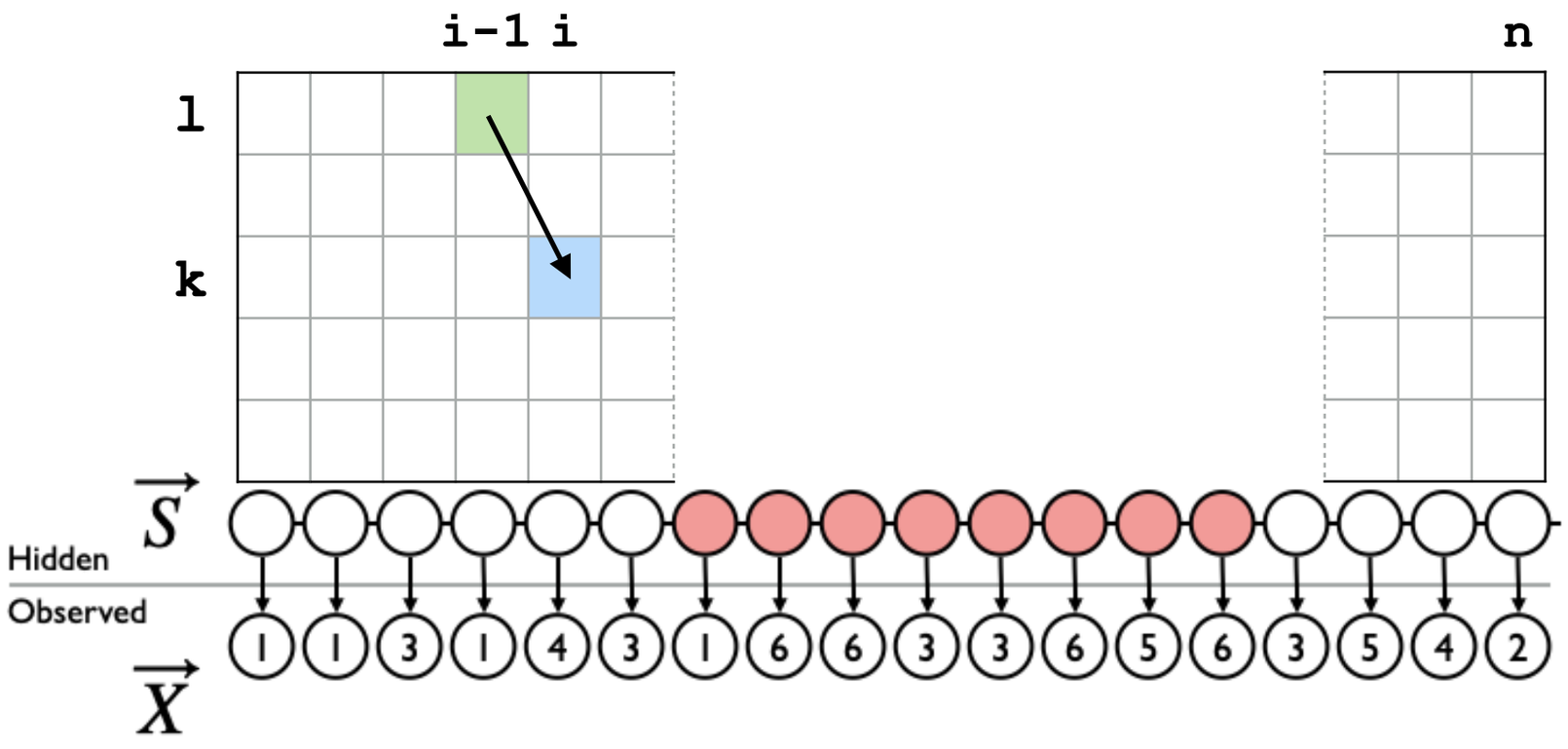
$$Ptr_k(i) = \arg\max_l \left[v_l(i-1) \cdot \tau_{l,k} \right]$$

Traceback:

$$\pi^*[n] = \arg\max_k V_k(n)$$

For $i = n \ldots 2$

$$\pi^*[i-1] = Ptr_{\pi^*[i]}(i)$$



דגימה ממרחב הפוסטריור

- האם נדע לדגום מסלולים [שלמים] מהסתברות הפוסטריור?
  [שימו לב להבדל, לעומת הפוסטריור המקומי $P[S_i|X]$]
- נפתח:

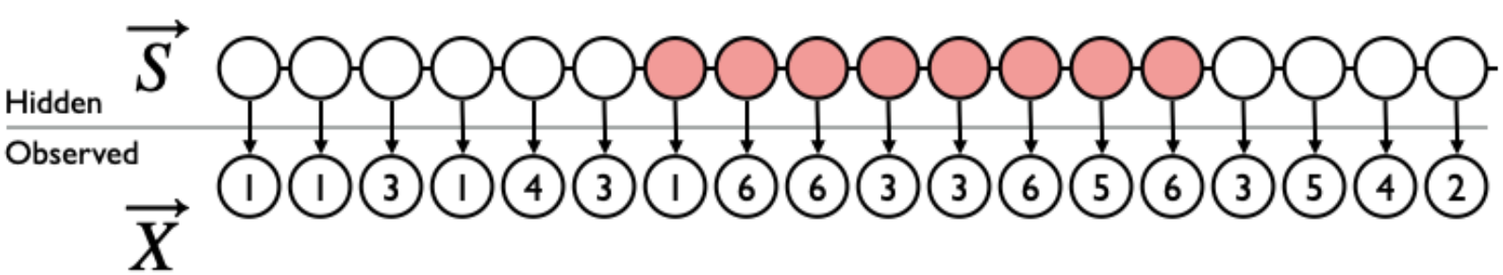
$$P(\vec{S}\,|\,\vec{X}) = \prod_i P(S_i\,|\,S_{i-1}, \vec{X})$$

$$= \prod_i P(S_i\,|\,S_{i-1}, X_1, \ldots, X_i, \ldots, X_n) = \prod_i \frac{P(X_i, \ldots, X_n, S_{i-1}, S_i)}{P(X_i, \ldots, X_n, S_{i-1})}$$

$$= \prod_i \frac{P(X_i\,|\,S_i) \cdot P(X_{i+1}, \ldots, X_n\,|\,S_{i-1}, S_i) \cdot P(S_i\,|\,S_{i-1})}{P(X_i, \ldots, X_n, S_{i-1})}$$

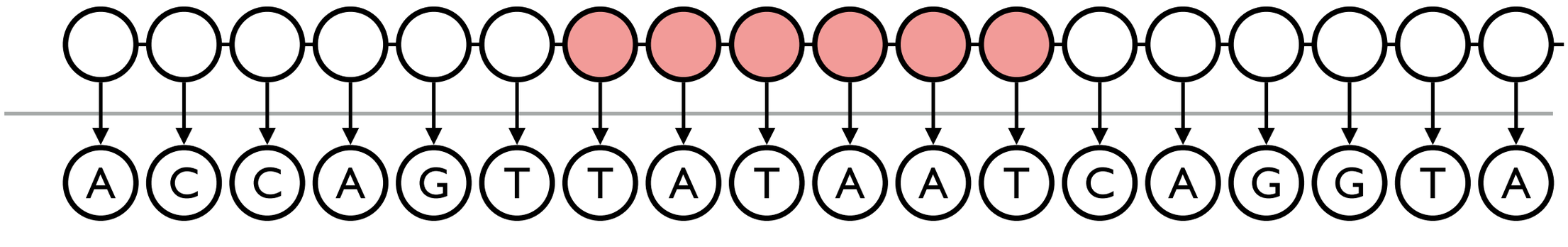
- בלומר נוכל לדגום עמדה עמדה:

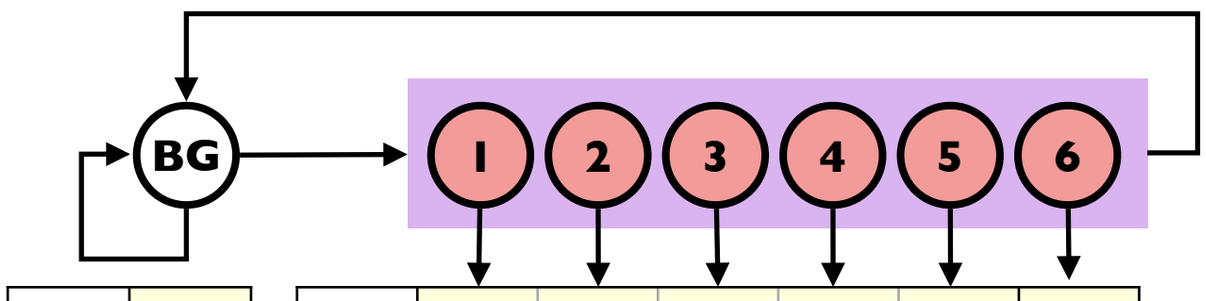
$$P(S_i\,|\,S_{i-1}, \vec{X}) \propto \; \mathsf{e}_{S_i}(X_i) \cdot B_{S_i}(i) \cdot \tau_{S_{i-1},S_i}$$



זיהוי מוטיבים עם מודלים חבויים

- נניח שנתונים רצפי דנ"א, וברצוננו למצוא בהם מוטיב ידוע [למשל TATA-box, אתר הקישור למכונת השעתוק]

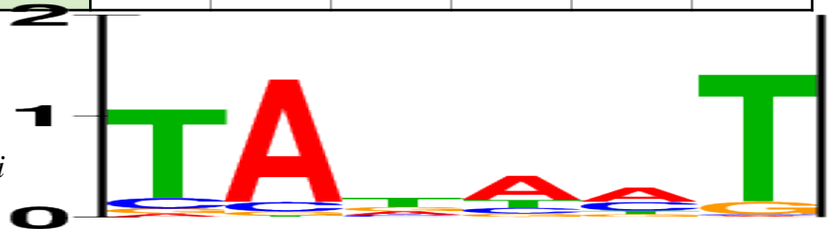




BG	1
A	0.25
C	0.25
G	0.25
T	0.25

TBP	1	2	3	4	5	6
A	0.03	**0.89**	0.2	**0.59**	**0.49**	0.01
C	0.10	0.07	0.13	0.14	0.30	0.01
G	0.05	0.03	0.25	0.08	0.10	0.09
T	**0.82**	0.01	**0.52**	0.19	0.11	**0.89**

$$H(p) = -\sum_i p_i \log_2 p_i$$



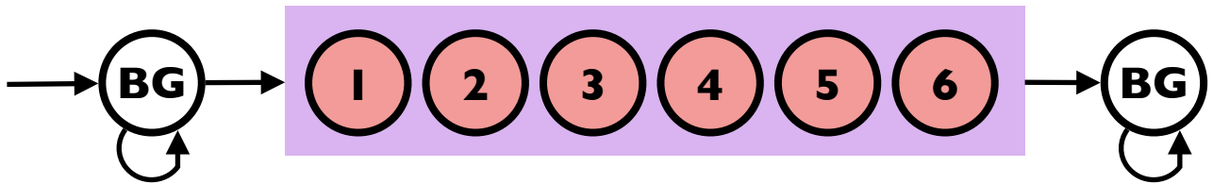
GCGGCACCCATCTCTTATAATTCGCTTGATAGTAACGTTCGAAAGCACAA

AAAGCGGTAGACTTCAGGCATAAAAGGATTAATTTTTGACAATCCCCGAT

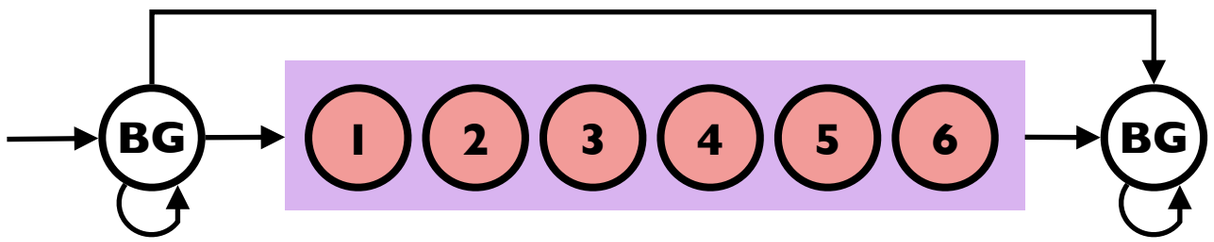
CATCCACAGTTAAGTTATCCGAAGGGGGTCGGCATCGGAAGCTTGAAATT

זיהוי מוטיבים עם מודלים חבויים

- מודל OOPS



- מודל ZOOPS



- מודל TCM