האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

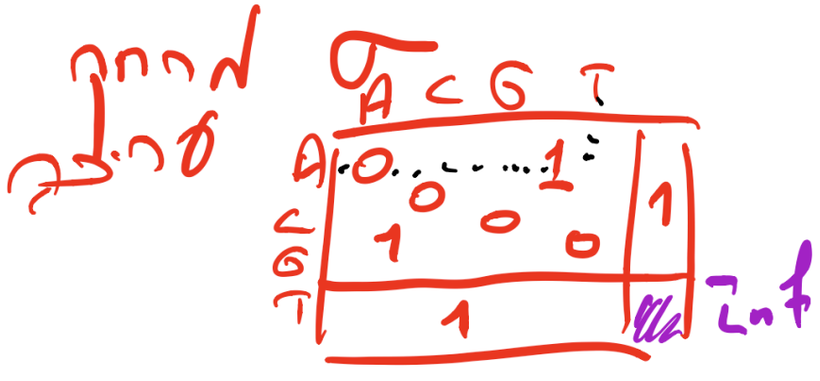
אלגוריתמים בביו' חישובית

76558

תכנון דינאמי ועימוד זוג רצפים

תומי קפלן

2/1/2024



עימוד רצפים ביולוגיים

S: AGT
T: AACT

1. האם יש **אב קדמון משותף** לרצפים S ו־T
2. נלמד על שכיחות מוטציות בדנ"א [שכיחות החלפות]
3. נניח מודל נאיבי, המבוסס על **אי תלות בין העמדות**

$\sigma = \Sigma^2 \to R \qquad \Sigma = \{A, C, G, T\}$

$\sigma = (\Sigma \cup \{\underset{\substack{\uparrow \\ \text{gap}}}{-}\})^2 \to \mathbb{R}$

Indel

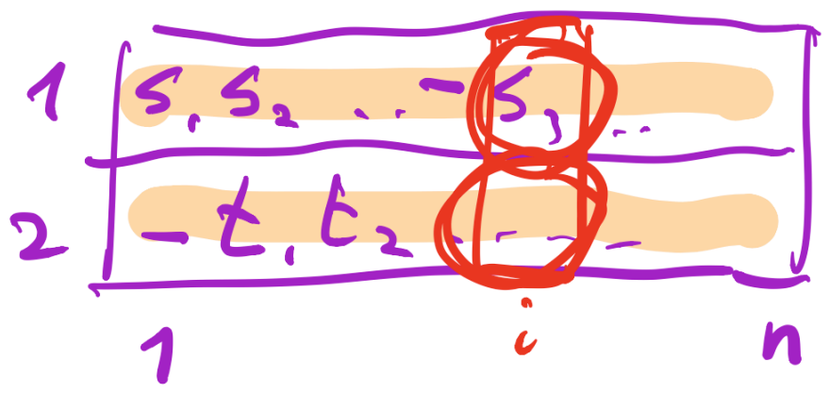
מרחק עריכה

עימוד רצפים ביולוגיים

ציון איכות בין המחרוזות השונות:

$$score(\text{עמודה } i) = \sum_{j=1}^{n} \sigma(a_{1,i}, a_{2,j})$$

עמודה i אורכו n



$$s = \{a_{1,i} : a_{1,i} \neq '-'\}$$

$$t = \{a_{2,i} : a_{2,i} \neq '-'\}$$

S: AGT
T: AACT

דוגמאות

עימוד 3

A	-	G	T
A	A	C	T
1	-2	-1	1

עימוד 2

A	G	-	T
A	A	C	T
1	-1	-2	1

עימוד 1

A	G	T	-	-	-	-
-	-	-	A	A	C	T
-2	-2	-2	-2	-2	-2	-2

σ	A	C	G	T	-
A	1	-1	-1	-1	-2
C	-1	1	-1	-1	-2
G	-1	-1	1	-1	-2
T	-1	-1	-1	1	-2
-	-2	-2	-2	-2	☠

כמה עימודים אפשריים יש?

לשני רצפים באורך n. אולי תנסו לחשב בבית?

מציאת העימוד האופטימלי

4. איך נמצא את העימוד הטוב ביותר בזמן סביר?

אלגוריתם [’58] Needleman-Wunch ע"ם
עקרונות של תכנון דינאמי [’54] Bellman

עקרונות התכנון הדינאמי

נפרק את הבעיה לסדרה של בעיות קטנות יותר,
שחוזרות שוב ושוב. נזכור את תתי הפתרונות, במקום לחשב אותם מחדש.

יהי $L[s,t]$ הסקור של העימוד האופטימלי a^* של s,t

$$Score(s,t) = \max \begin{cases} score(s_{1..n-1}, t_{1..m-1}) + \sigma(s_n, t_m) \\ score(s_{1..n-1}, t_{1..m}) + \sigma(s_n, -) \\ \vdots \text{ score } \dots \end{cases}$$

S_1	S_2	S_3				S_n
T_1	T_2	T_3				T_m

אפשרות א'

						S_n
						T_m

אפשרות ב'

						S_n
						-

אפשרות ג'

						-
						T_m

עקרונות התכנון הדינאמי

בלומר, כדי לחשב את $L(s,t)$ נוכל לקחת את המקסימום מבין 3 פתרונות קצרים יותר + תוספת הציון של העמדה האחרונה.

אפשר לשמור בטבלה V את ציוני עימוד הרישאות:

$$V[i,j] = L(s_{1\ldots i}, t_{1\ldots j})$$

והפתרון האופטימלי יהיה:

$$V[n,m] = L(s,t) = L(s_{1\ldots n}, t_{1\ldots m})$$

	-	A	A	C	T
-	0	-2	-4	-6	-8
A	-2	1	-1	-3	-5
G	-4	-1	0	-2	-4
T	-6	-3	-2	-1	-1

	-	A	A	C	T
-	0	-2	-4	-6	-8
A	-2	1	-1	-3	-5
G	-4	-1	0	-2	-4
T	-6	-3	-2	-1	-1

	-	A	A	C	T
-	0	-2	-4	-6	-8
A	-2	1	-1	-3	-5
G	-4	-1	0	-2	-4
T	-6	-3	-2	-1	-1

A	-	G	T
A	A	C	T

A	G	-	T
A	A	C	T

-	A	G	T
A	A	C	T

Mat	1

Mi	-1

Gap	-2

Needleman-Wunch [70]

$$V(0,0) = 0$$

$$V(i,j) = \max \begin{cases} V(i-1,j-1) & +\sigma(S_i, T_j) \\ V(i-1,j) & +\sigma(S_i, -) \\ V(i,j-1) & +\sigma(-, T_j) \end{cases}$$

$$V(n,m) = Score(a^*) \qquad a^* = \underset{a \in A(S,T)}{\arg\max}(Score(a))$$

סיבוכיות זמן ומקום

- זמן ריצה: $O(n \cdot m \cdot k) = O(n^2)$ (k קבוע)
- מקום: $O(n \cdot m) = O(n^2)$ מטריצה

האם טוב מספיק בדי לעמד גנומים? [10^9]

(ביולוגי הישראלי (True) שחצה הקשים) ∨ c

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

אלגוריתמים בביו' חישובית

76558

תכנון דינאמי ועימוד זוג רצפים

תומי קפלן

7/1/2024

עימוד אופטימלי בשטח לינארי

- אבחנה 1: קל למצוא את הסקור האופטימלי L בשטח לינארי

	-	A	A	C	T	T	C	G	G
-									
A									
G									
T									
A									
C									
T									
G									
G									

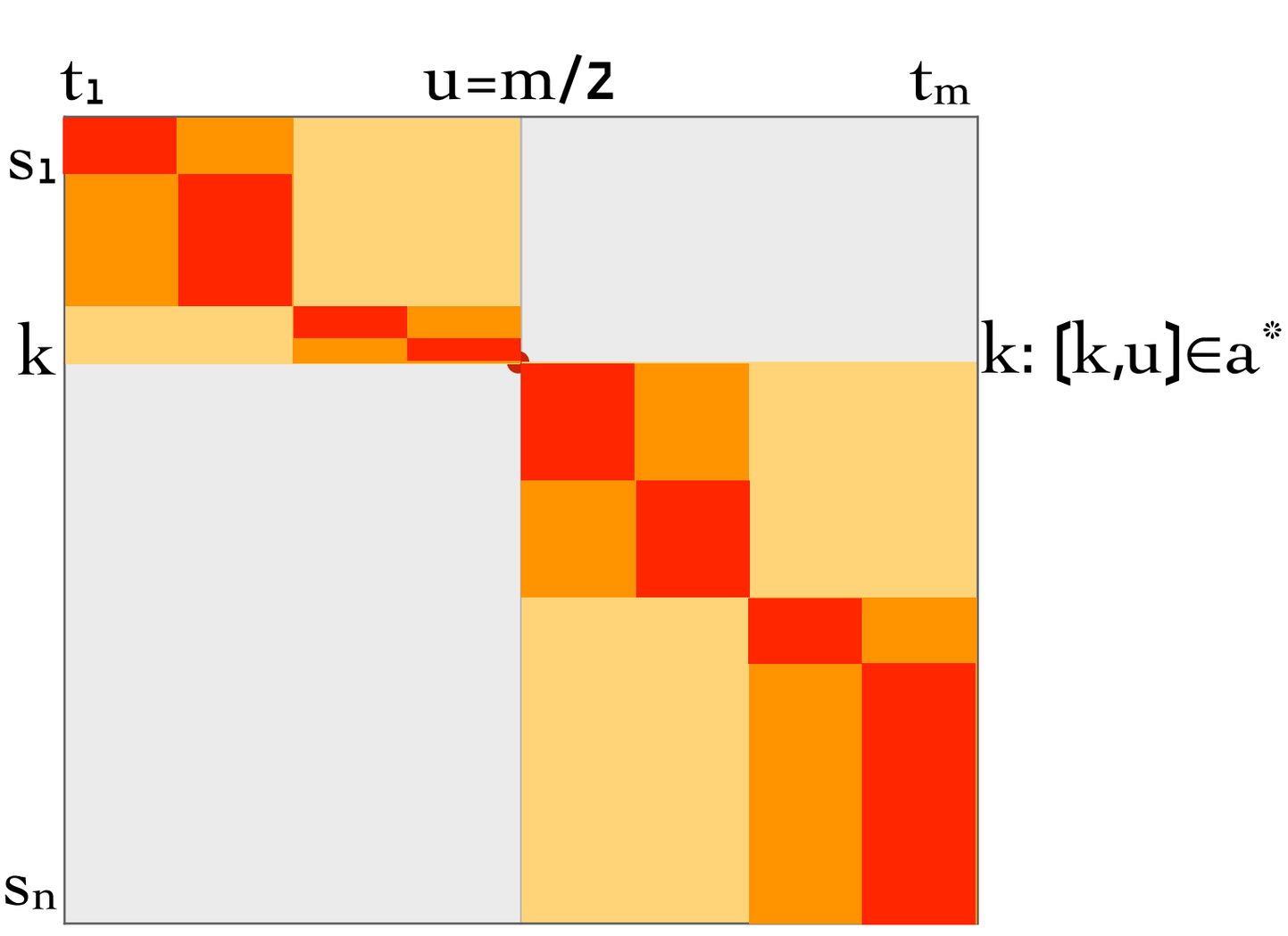
$v[n,m] = L(s,t)$

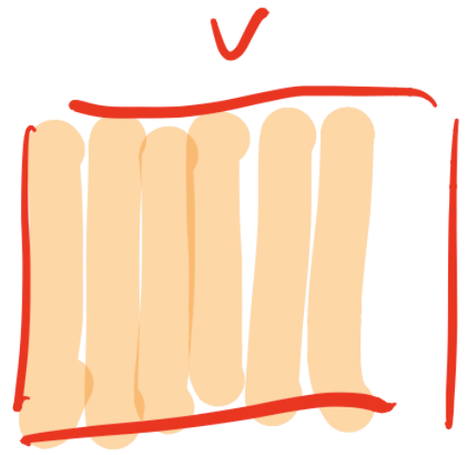
ב"ו: חשב את הערכים מן.

- הבעיה שבינתיים שכחנו את המסלול האופטימלי

עימוד אופטימלי בשטח לינארי

- אבחנה 1: קל למצוא את הסקור האופטימלי בשטח לינארי
- אבחנה 2: בהינתן נקודה אחת על המסלול האופטימלי, קל לחלק את הבעיה לשתי בעיות קטנות (הפרד ומשול)





עימוד אופטימלי בשטח לינארי

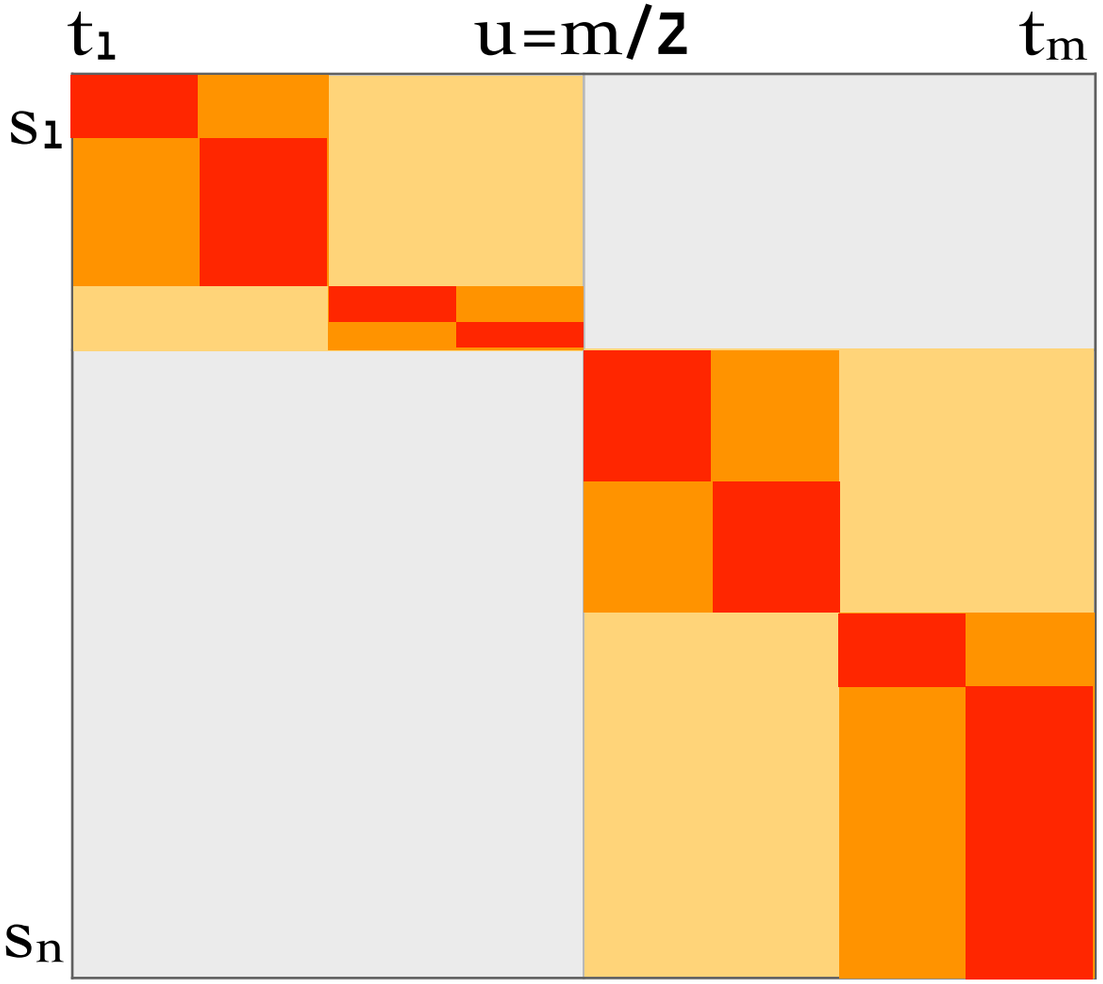
$$T = T_1 + T_2 + T_3 + T_4 + \dots$$

$O(n \cdot m)$ $O(n \cdot m)/2$ $O(n \cdot m)/4$ $O(n \cdot m)/8$

M $M/2$ $M/4$ $M/8$

- זמן ריצה כולל:
- בלומר

$$T = M\left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} \dots\right) \leq 2M$$



- זמן ריצה כולל: $O(n^2)$
- וסיבוכיות מקום: $O(n)$

עימוד לוקאלי לעומת גלובאלי

Smith-Waterman [81]

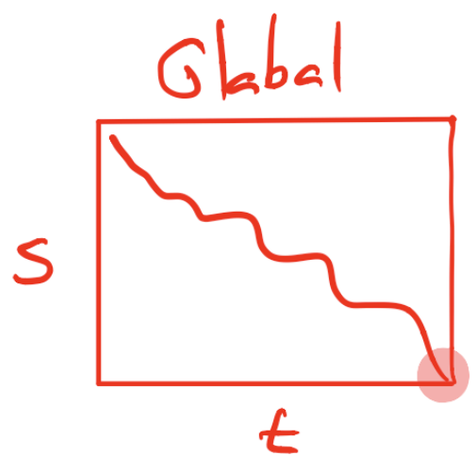
עימוד לוקאלי – עימוד אופטימאלי בין תתי-רצפים של s,t

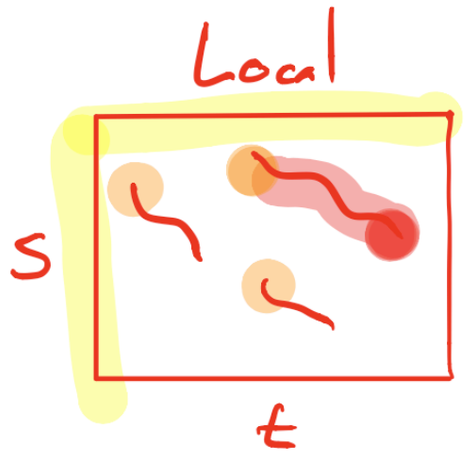
משמעות האפס – נקודות התחלה/סוף לכל תת עימוד

$$V(0,0) = V(0,\cdot) = V(\cdot,0) = 0$$

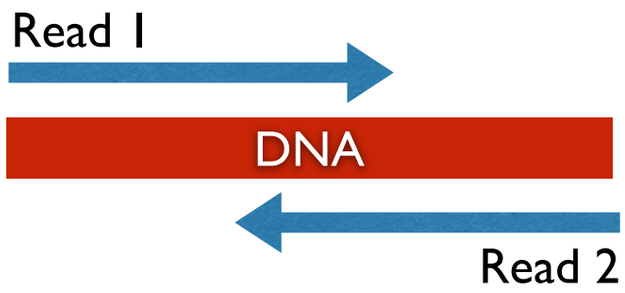
$$V(i,j) = \max \begin{cases} 0 & \\ V(i-1,j-1) & +\sigma(S_i, T_j) \\ V(i-1,j) & +\sigma(S_i, -) \\ V(i,j-1) & +\sigma(-, T_j) \end{cases}$$

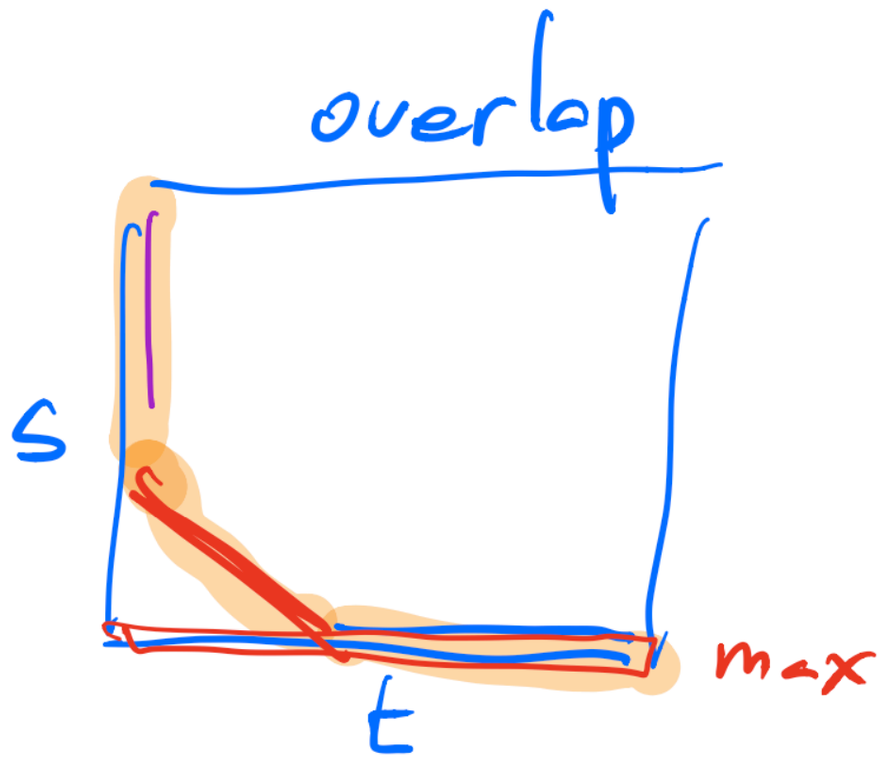
$$Score(a) = \max_{i,j} V(i,j)$$

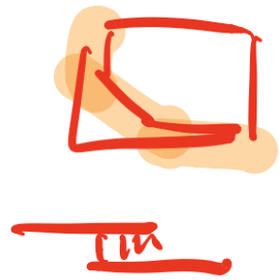


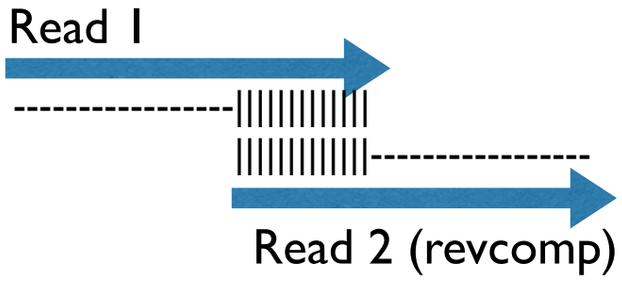


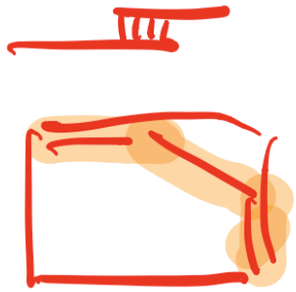
[overlap] עימוד חופף











ידוע מראש שנרצה לעמד את הסיפא של s מול הרישא של t.

איך נשנה את הסקור?

איך נשנה את שחזור המסלול האופטימלי?

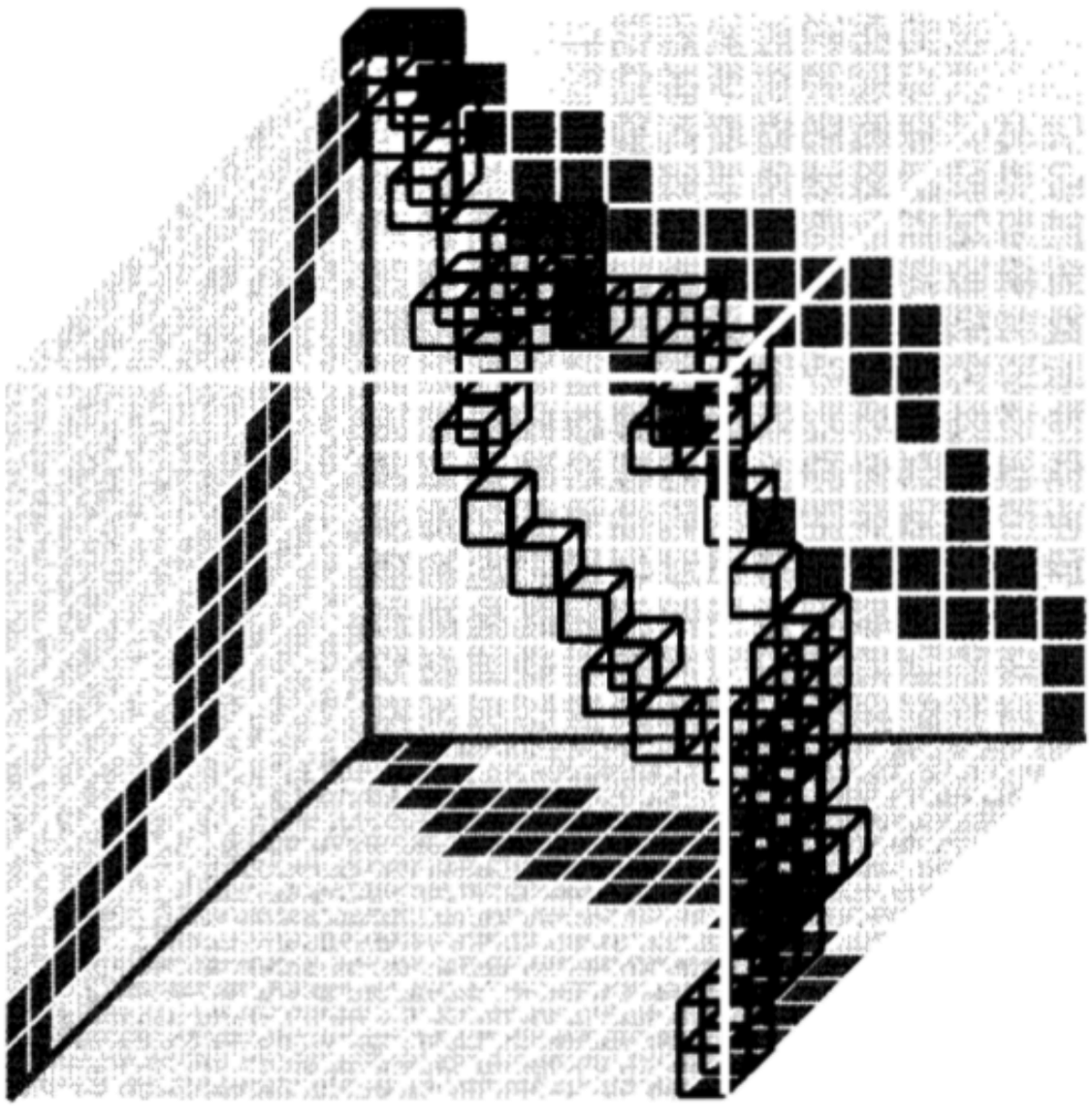
נושאים לדיון

- שחזור פתרון בשטח ריבועי? לינארי?
- היוריסטיקות עימוד [אלכסון, סקור מינימלי]

	-	A	A	C	T	T	C	G	G
-	0								
A									
G									
T									
A									
C									
T									
G									
G									

נושאים לדיון

- שחזור פתרון בשטח ריבועי? לינארי?
- היוריסטיקות עימוד [אלכסון, סקור מינימלי]
- עימוד מרובה [טנזורים?]



זמן ריצה? MSA

יישומים היוריסטיים?

[Carrillo & Lipman, 88'] [https://www.cs.tau.ac.il/~rshamir/algmb/archive/msa.pdf]

נושאים לדיון

- **המודל שהצענו מניח אי-תלות בין העמדות.**
- לגבי איזה רצפים ההנחה אולי סבירה?
- לגבי איזה לא?

[מוטיבים – אתרי קישור של פקטורי שעתוק]
רצף גאפים [אינדלים: הוספות/מחיקות]

- איפה חוסר הזכרון נכנס למודל שלנו?
  הציעו מודל חלופי בעל זכרון?

[אפשר להכין על זה סקרייב]