

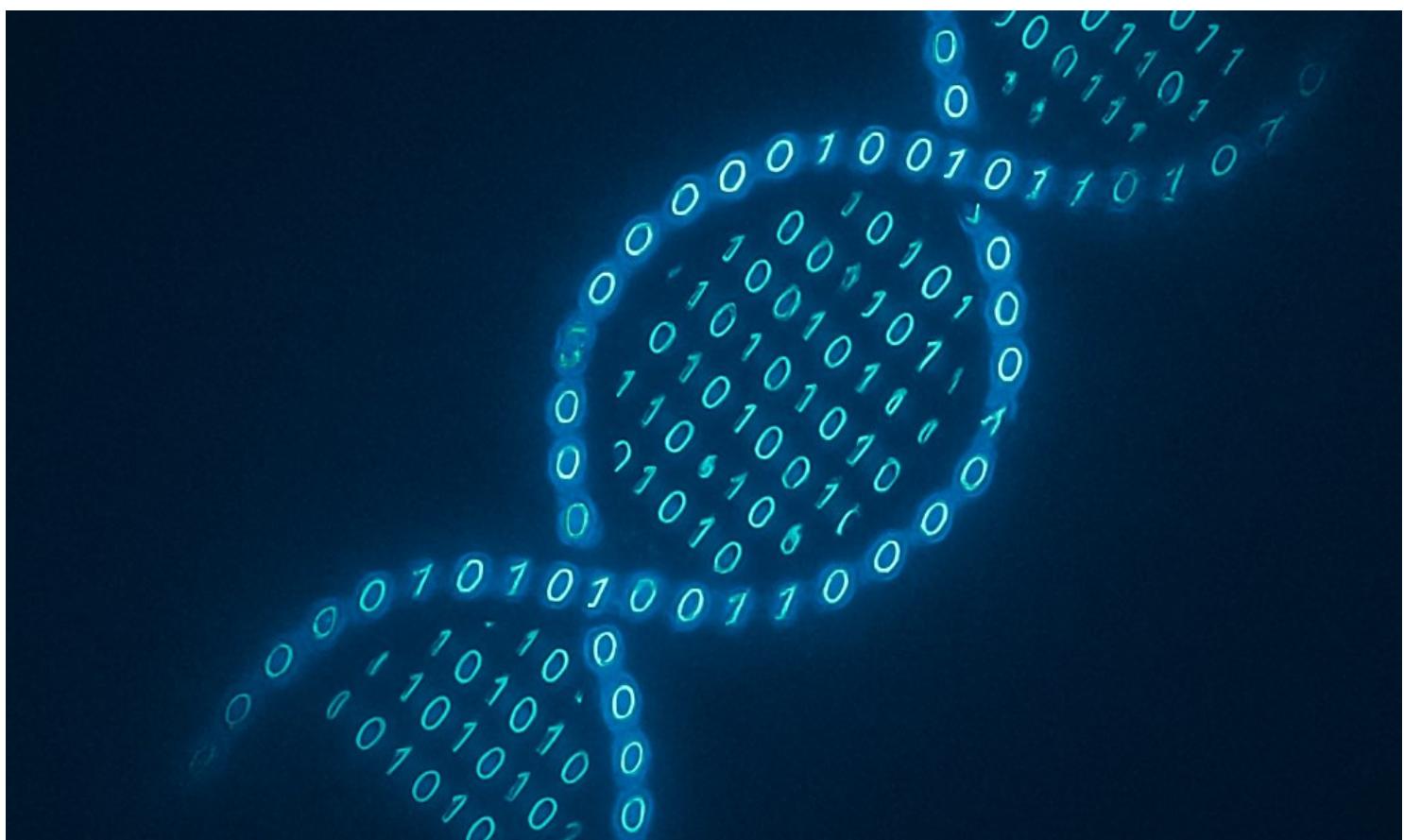
SUMMARIZED BY NOAM KIMHI
noam.kimhi@mail.huji.ac.il

ALGORITHMS IN COMPUTATIONAL BIOLOGY

YEAR 3 – SEMESTER A

COURSE NO. 76558

2025-2026



סיכום הרצאות הקורס

מרצה: ניר פרידמן

תוכן עניינים

שבוע 1 – הרצאה 1	3
הקדמה – ביולוגיה מולקולרית	3
הדוגמה המרכזית של הבiology המולקולרית – Central Dogma of Molecular Biology	4
קידוד לחומצות אמינו	5
פיתוח אלגוריתם לבעה ביולוגית	6
שבוע 1 – הרצאה 2	7
עימוד רצפים – Sequence Alignment	7
תבנון דינמי – Dynamic Programming	8
עימוד במקום לינארי – Linear Space Alignment	10
עימוד מקומי – Local Alignment	12
שבוע 2 – הרצאה 3	13
��型 Probabilistic Models and Decisions	13
Two Hypotheses	13
лемת נימן-פירסון (1933)	16
One Hypothesis	17
שבוע 2 – הרצאה 4	19
למידה על הסתברות מנתונים	19
דרך שיעור בייסיאנית	20
שימוש בהסתברות בעימוד רצפים	21
שבוע 3 – הרצאה 5 (הרצאת אורח)	23
הקדמה	23
מודול אינטגרטיבי של מערכות ביולוגיות דינמיות	24
שלבים ביצירת מודל	25
שימוש במודל	26
שבוע 3 – הרצאה 6	27
היררכיות לעימוד רצפים	27
שרשראות מركוב	28
שבוע 4 – הרצאה 7	31
מודלים מרקוביים חכויים – Hidden Markov Model	31
שבוע 4 – הרצאה 8	36
בעית השחזר – Maximum Probability Reconstruction	36
שבוע 5 – הרצאה 9	41
אלגוריתם Max-Max (MM)	41
אלגוריתם Exp-Max (EM)	42

46	שבוע 5 – הרצאה 10
46	הгинום וברומטידים
53	הרצאה 11
53	סימוניים ביומיים על החלבון
54	ChromHMM

שבוע 1 – הרצאה 1

הקדמה – ביולוגיה מולקולרית

הקורס יתמקד בביולוגיה מולקולרית, ובעשלה כיצד מתמודדים עם הכלים שהביולוגיה מעניקה. עולם הביולוגיה מושפע מאוד מהטכנולוגיה. בקורס נרצה לענות על השאלות – איך עובדים עם המידע הנוכחי לנו? איך מפתחים דברים חדשים? איך עובדים עם המידע הזה מבחינה אלגוריתמית?

מהי ביולוגיה מולקולרית?

ביולוגיה – מדע המתעסק בחקר היצורים בעלי יכולת לגודל ולהתרבות. דוגמאות: בעלי חיים, צמחים, יצורים חד תאים.

בביולוגיה נמצא היררכיה המתחילה באוכלוסיות של יצורים חיים, ונגמרה באטום הבודד:

	Population	All the cats in TLV
$\approx 10^0 - 10^1 m$	Organism	Cat
	Organs/Tissues	Heart
בביולוגיה מולקולרית	$10^{-5} m$	Cells
	$10^{-6} m$	Organelles
	$10^{-7} m$	Macro-molecules
	$10^{-9} m$	Small molecules
	$10^{-10} m$	Atoms
		H, O

מדובר על (!) סדרי גודל בין המולקולה הקטנה לאורגניזם.

מקורות מולקולות – הרבה מולקולות שמרוכבות יחד מסט מוגבל של יחידות, ואוצר די קטן של אבני בניין שבאמצעותן ניתן ליצור מבנים מסובכים.

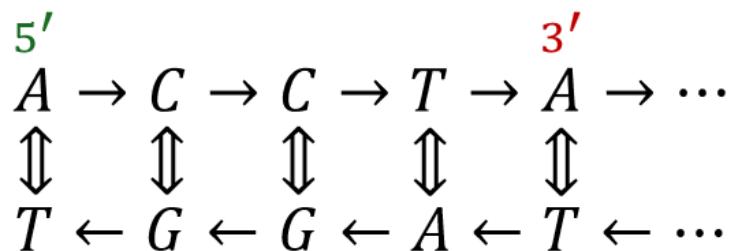
דוגמאות:

באן נתמקד בקורס	נוקלואוטידים (T,G,C,A)	DNA
	נוקלואוטידים (A,U)	RNA
	חומצות אמינו (יש 20 קבניות שונות)	חלבון
	מוносברידים (לא הוזכר בהרצאה)	סוכרים
	גליקול וחומצות שומן (לא הוזכר בהרצאה)	ליפידים
	יצרים מחיצות בין אזוריים נזליים הודיעות להידרופוביות	

נוקלואוטידים של DNA ושל RNA מעט שונים מבחינה כימית.

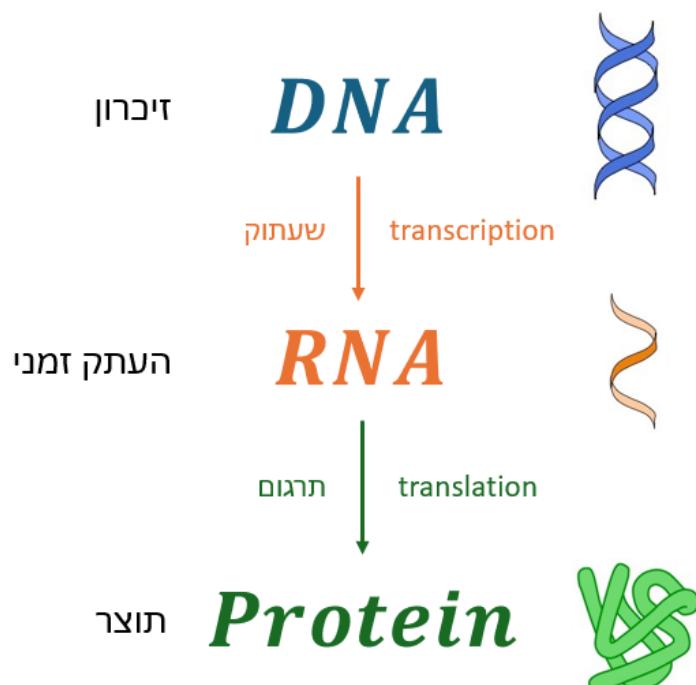
DNA

ה-DNA זו מולקולה גדולה השומרת מידע. ה-DNA הוא דו-גדייל. הנקלאוטידים בגדיילים שונים מצוותים בזוגות כך: $G \leftrightarrow C$, $T \leftrightarrow A$. ככל גדייל יש קשר חזק בין הנקלאוטידים באותו גדייל (לא יפרד בהרתחה), ובין הגדיילים קשר כימי **חלש יחסית** (יפרד בהרתחה, אבל לאור המשיכת החזקה – שני הגדיילים ימצאו אלו את אלו). נהוג לסמן כך:



מרכז אחד ניתן לשחזר את الآخر, וכך אפשר ליצור 2 העתקים של הרצף. תבונה זו עונה על שאלת בסיסית של העברת אינפורמציה – תורשה גנטית. ה-DNA לבדו נטול ערך, מדובר במאגר מידע ללא פונקציונליות משל עצמו. ל-RNA אין נטייה להתחבר למולקولات אחרות (ובכך להיהרס). לתמונה נכנס RNA שונה בהעתקת הבסיסים: $G \leftrightarrow U$, $C \leftrightarrow A$. ה-RNA יותר פעיל ביומית, וממנו מתרגמים את המידע לחלבון בעל פונקציונליות ומטרה.

הדוגמה המרכזית של הבiology המולקולרית – Central Dogma of Molecular Biology



מהתוצר – החלבון, אפשר להרכיב דברים בתא, לייצר מנועים לכיווץ שריר, להעביר סיגנלים, מולקולות וכו'. ייצור החלבון מוכתב לפי הרצף ב-DNA, אשר עברו גם לדoor הבא, והדרך לייצר עוד חלבון היא לבצע קריאה נוספת מה-DNA.

קידוד לחומצות אמינו

מדובר בקידוד משפה של 4 אותיות (A,T,C,G) לשפה של 20 אותיות (חומצות אמינו). מזכיר לנו מעט תרגום ביןארי. יש התאמה ברורה בין מילה באורך 4 לספרה בין 0-15 (בסיס הקסדצימלי). בשעורים למדמים של 4 אותיות אל 20 אותיות, היחסים הללו לא נשמרים – הבiology בדבונית במובן זהה, כי משפה של 4 אותיות התאמות מדויקות יהיו ברכף באורך 16, 64, 4, – אבל 16 מעט מדי, ו-64 יותר מדי.

התרגומים מתבצע על ידי תרגום מילים באורך 3 נוקלאוטידים, ויש לחלק מחומצות האמינו יותר מאשר אחת לתרגום:

64 different options	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px;">AAA</td><td style="padding: 2px;">Lys</td></tr> <tr><td style="padding: 2px;">⋮</td><td style="padding: 2px;">⋮</td></tr> <tr><td style="padding: 2px;">TTT</td><td style="padding: 2px;">⋮</td></tr> </table>	AAA	Lys	⋮	⋮	TTT	⋮	} 20 amino acids
AAA	Lys							
⋮	⋮							
TTT	⋮							

3 מילים מתוך ה-64 מקודדות ל-**STOP**, מילים שמשמעותן להגיד איפה נגמר הרצף שנדרש לתרגום. יש חומצת אmino אחת המקודדת ל-**START** (ATG – מתיוכין).

… CA ATG TAA CTG GTT UGA CC … ATG …
start *stop* *start*

הרץ' ATG יכול להופיע גם באמצע רצף ולפודד למתיוכין, נאמר כי זו חומצת אmino מדירה יחסית, ואולי זו הסיבה שקיבלה את התפקיד הנוסף.

מסגרת קריאה/reading frame – השלשות שמתרגמים ברכף. הזו בודדת תוביל לתוצר שונה לחלוון. רצף DNA יכול להכיל 3 מסגרות קריאה בכל בזען של הרץ', لكن בסך הכל יש 6 אפשרויות לקריאה.
ריצוף/sequencing – התהליך של פיענוח סדר הנוקלאוטידים בקטע DNA לרץ' אותיות. ביום ניתן לרץ' גנים של אורגניזמים שלמים, ומדובר בכ- **$10^9 \times 3$ אותיות**.

פיתוח אלגוריתם לבעה ביולוגית

לא נتمكن רק באלגוריתם, אלא גם **בלמה פיתחנו אותו, ועל מה הוא מנסה לענות.** התהילה: ([דוגמה בתכלת](#))

1. **שאלה ביולוגית** – החלק שאותו אנו מעוניינים לחקור.

מה זה רצף החלבן שמולנו? מה מטרתו?

2. **אינפורמציה רלוונטית** – כשבאים לענות על השאלה יש הרבה כיוונים לחקור.

האם אנחנו מעוניינים לבנות מודל **כימי של החלבן? לצורך הדוגמה, בקש לחקר מהכיוון של דמיון בין רצפים נובע ממקור אבולוציוני דומה,** ולכן גם מטרה ופונקציונליות דומות.

sequence similarity → common ancestor → same function

3. **דאטा** – עבור הרעיון צריך בסיס נתונים, בזה שuber בדיקות והוא מהימן.

Protein database sequence function

4. **שאלה אלגוריתמית/מתמטית** –

בhinתן הדאטא, מתן מענה לשאלה.

Sequence comparison/query

בשלב זהה גם מישחו שלא מבין ביולוגיה יכול להיות מסוגל לסייע על ידי כך שניצור הגדרה לרעיון.

Similarity is...

-- עד פה הגדרנו רק את מטרת האלגוריתם, ולא מה הוא עשה --

5. **אלגוריתם** – מציאת האלגוריתם לפתרון הבעיה.

למצוא את המרחק הקטן ביותר.

6. **בחירה פרמטרים** – מדובר בעולם דאטא, האם אפשר להשתמש בה כדי להגיד את מחיר הטעות?

להשתמש במידע של פונקציונליות חלבונים, נכנס **aan almento لمידה.** **aan זה מסתבר** – אפשר להגיד רמות קירוב ומה טווח הטעות שנרצה. הגדרה שיצרנו בשלב 4 צריכה להיות מגובה בפרמטרים: "דמיון בין שני רצפים לפי טבלת 20×20 שבה כל אחת קרובה לאחרת". זו טבלה שאנו נספק לייצר האלגוריתם.

7. **סטטיסטיקה** – נרצה לגבות את התוצאה שלנו תוך השוואה סטטיסטית לתוצאה מקראית, זה ייתן לנו וดาות על נוכחות התוצאה.

מה היה קורה אם לא היה חלבון דומה ב-DB? מה אז הייתה התשובה? נשווה את התוצאה לרץ' רנדומלי של אותיות, האם קיבלנו בערך את אותה תוצאה? "Similar by chance"

8. **יזואיזציה** – איך הגיענו למסקנה הסופית? מה תומך בה?

סיפוק של בסיס שיתמוך במסקנות שלנו.

שבוע 1 – הרצאה 2

עימוד רצפים – Sequence Alignment

הרבה מאקרו-מולקולות שראינו מתוארכות על ידי רצפים של אותיות, והרבה מהbilogיה החישובית מתחילה בעבודה עם רצפים. אחת השאלות המרכזיות היא האם שני רצפים דומים, ואם כן – מה זה אומר?

לאורך השיעור עבדנו כדוגמה עם שני הרצפים הבאים: $s = AACT$ $t = AGT$

עימוד – בהינתן שני רצפים t , s מנפיקים באמצעות הסופת – לקבלת 2 מחרוזות באותו אורך. יש מספר גדול של עימודים, כאשר האיסור היחיד הוא – מול –.

קודם נתבונן בכמה דוגמאות ובדון בדמיון בין רצפים:

CTAACTG
GAGTG
GACTG

שאלת העימוד – איך ניקח 2 רצפים ונדי их שוויון דומים? לשם כך נגדיר את **חוקי המשחק**:

(1) מותר לבצע הדוזות (2) מותר להכניס רווחים –

נגדיר א"ב מוחרב: $\{ -, A, T, C, G \}^{*}$ וنمצא שני רצפים כך שמתקיים $|t^*| = |s|$ וכן שמתקאים $s = (t^*) = t$, $remove - (s^*) = remove - t$, $remove - (s^*)$ מסירה את הרווחים שהוספנו. לצורך הדוגמה נתבונן ברצפים:

C | – G A G T G
C T A A C T G

בירוק – רואים התאמה/match כי זיהינו את אותן אותות.

באפור – רואים indel¹ כי מופיעה אותה מול רווח.

בכתרום – מופיעה אי-התאמה/mismatch כי אין שווין באותיות.

הערה: לא נרצה עימוד של – מול – לאחר שעימוד זה הוא חסר משמעות.

עבור הדוגמה של הרצפים t , s לעיל, ניצור את הרצפים t^* , s^* כך:

$\frac{s^*}{t^*}$ A A C T A A – C T A A C T – – –
A – G T A – G – T – – – A G T

¹ הלוחם של המילים של insert ושל deletion, לאחר שבעבור הרץ' העליון עליינו לבצע (T) ובעבור הרץ' השני insert(T). עליינו לבצע (T).

באבולוציה של רצפים יש מוטציות או טעויות העתקה שגורמות להכנסה/אובדן רצף. Indel מתאר אירוע של אובדן/תוספת חתיכת מרצף. ברצפים של חלבונים אלו אירועים יותר נדירים – אבל קורים.

נרצה להגיד פונקציית ציון σ עבור רמת קיבלה לרצף נתון. פונקציית ציון כזו יכולה להיות למשל:

$$\sigma(x, y) \rightarrow \mathbb{R} \quad \sigma(x, y) = \begin{cases} +1 & x = y \quad (match) \\ -1 & x \neq y \quad (mismatch) \\ -2 & x = - \vee y = - \quad (indel) \end{cases}$$

עבור הדוגמה שראינו קודם, נראה עבשו ציון:

s^*	A	A	C	T		A	A	-	C	T		A	A	C	T	-	-	-
t^*	A	-	G	T		A	-	G	-	T		-	-	-	-	A	G	T
σ	+1	-2	-1	+1		+1	-2	-2	-2	-1		-2	-2	-2	-2	-2	-2	-2

כעת יש בידינו את הכלים להגדיר בצורה טובה את הבעיה:

בහינתן שני רצפים t , s ופונקציית ציון σ , מצא את העימודים עם הציון המקסימלי.

הערה: תמיד קיים ממשו טוב יותר מהעימוד הנוכחי ביותר (יש דוגמה בטבלה לעיל, כל התאים Indel). כמו כן, ניתן מצב בו יותר מעימוד אחד מקבל את אותו הציון, שכן יכול להיות יותר מעימוד אחד עם ציון מקסימלי.

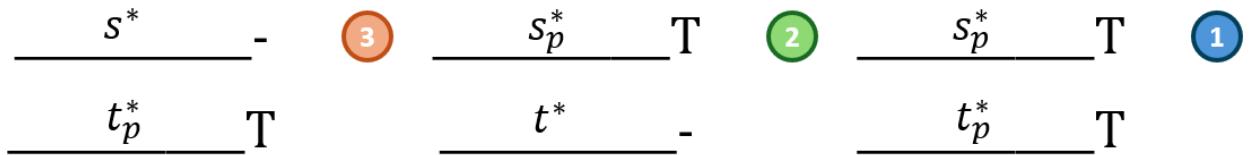
תכנות דינמי – Dynamic Programming

בקורס אלגוריתמים נתקלנו בעיית התכנון הדינמי, ובמסגרתו דנו בעיית **מრחיק עריבה**. נבחן כי יש דמיון רב בין בעיה זו לבין רצפים. נרצה לפתור את בעית עימוד הרצפים באותה דרך שבה פתרנו בתכנון דינמי – באמצעות הגדרת תתי בעיות ושימוש חוזר בפתרונותות ביניים על מנת לצמצם את מספר הפעולות ובכך לצמצם משמעותית את זמן הריצה.

בහינתן שני רצפים t , s שמסתויימים באות T נסתכל עליהם כך:

$$\frac{*}{t_{prefix}} \text{T} \qquad \frac{*}{s_{prefix}} \text{T}$$

יש למעשה שלוש אפשרויות למקם את הרצפים זה מול זה:



ולכן:

$$V(s, t) = \max \begin{cases} V(s_p, t_p) + \sigma(T, T) \\ V(s_p, t) + \sigma(T, -) \\ V(s, t_p) + \sigma(-, T) \end{cases}$$

נגיד $n = |s|$, $m = |t|$ ו i, j

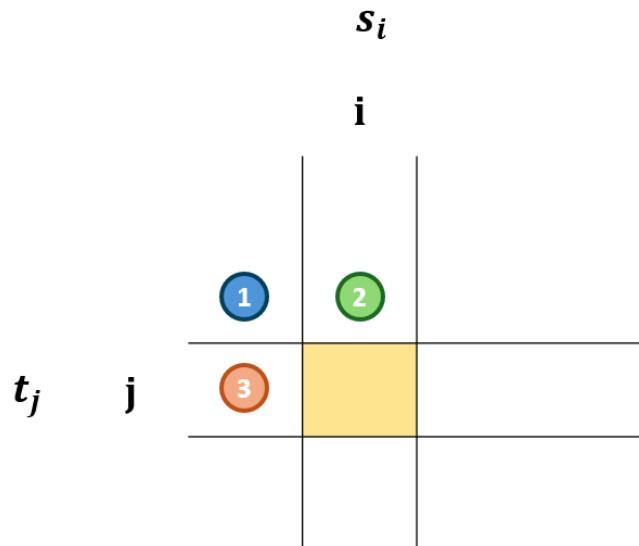
$$V[i:j] = \max(s[1:i], t[1:j]) \quad \text{מרקורי בסיס: } V[-1,*] = \infty \quad V[*,-1] = \infty \quad V[0,0] = 0$$

הגדרת הרקורסיה:

$$V[i:j] = \max \begin{cases} V[i-1:j-1] + \sigma(s_i, t_i) \\ V[i-1:j] + \sigma(s_i, -) \\ V[i:j-1] + \sigma(-, t_i) \end{cases}$$

מילוי הטבלה:

באופן כללי אם נתבונן בתא בטבלה, אלו הכללים למילוי:



1. התקדמות לפני s_i מול t_j
2. התקדמות לפני t_j מול $indel$
3. התקדמות לפני s_i מול $indel$

שיטה זו מוצעה לנו גם דרך לשחזר את העימוד. נסמן בכל שלב אילו מ-3 האופציות בחרנו כשמילאנו את התא.

גיעה למילוי באופן הבא:

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

סיבוכיות מקום: $(m \cdot n)^O$ כמספר התאים בטבלה.

סיבוכיות זמן: בכל תא בחרנו 3 אופציות, זמן חישוב קבוע לתא – לכן סה"כ כמספר התאים $(m \cdot n)^O$.

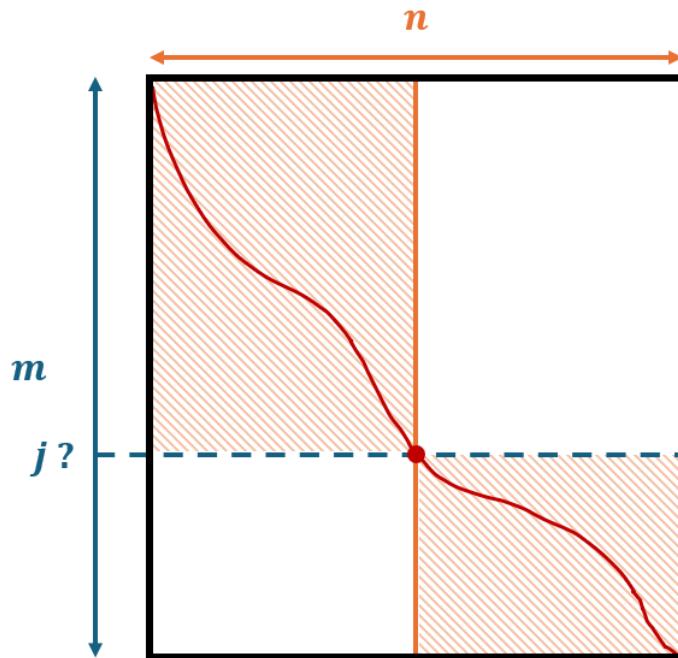
זמן שקיבלנו הוא פולינומי, אבל אם מבצעים עימוד גנים של אדם מול גנים של עבר, עדין מדובר בסדרי גודל עצומים ומתיילה גם בעיית מקום ומוגבלות hardware. כיצד נתמודד? כדי לחשב את הציון בלבד, אין צורך במטריצה כולה, אבל עבור העימוד הסופי علينا לשמור יותר מידע.

עימוד במקום לינארי – Linear Space Alignment

פתרון נאייבי: יכלו נסות בכל שלב להתחשב ב-3 תתי מטריצות עבור 3 האפשרויות. במקרה זמן הפור להיות בשלישית, לא מאד עוזר.

נחפש פתרון טוב יותר.

עבור העימוד הטוב ביותר, נשאל איפה עבר קו האמצע. את העמודות נחצה באמצע, אבל איפה עבר הקו בשורות?



כך נוכל לפתור כל צד באופן בלתי תלוי. חתכנו את s באמצע, ומשם נפתחו שתי בעיות עימוד. העימוד הכי טוב של הרישות בתוספת העימוד הכי טוב של הסיפות ייתן את העימוד הכלול הטוב ביותר.

$$\text{רישות: } V[i:j] = \max(s[1:i], t[1:j])$$

$$\text{סיפות: } U[i:j] = \max(s[i+1:n], t[j+1:m])$$

מבחינת חישוב, U ו- V דומים. מכיון מעבר קו המחזית ב- j הוא $\left[\frac{n}{2}, j\right]$ ו- $V\left[\frac{n}{2}, j\right] + U\left[\frac{n}{2}, j\right]$ ולכן השאלה היא מיצאת ה- j המיטבי:

$$j^* = \arg \max V\left[\frac{n}{2}, j\right] + U\left[\frac{n}{2}, j\right]$$

מדוע אפשר לעשות זאת במקום לינארי? המקום הוא שמיירה של 2 עמודות, והזמן הוא הגודל של הטריטוריה (גודל המטריצה). לכן צעד ה-"Divide" הוא בזמן $(m \cdot n) O$ והמקום הוא $O(n \log n)$. עומק הרקורסיה הוא $\log 2$ כי בכל פעם אנחנו חוצים את מספר העמודות ב-2.

באופן נאיבי יש לנו ברגע אלגוריתם שהזמן שלו הוא $O(n \log n)$ אבל זה לא בדיק המצב, מדוע? כי מדובר במציאות חיוני:

$$T(n, m) = O(n \cdot m) + T\left(\frac{n}{2}, \textcolor{green}{j}\right) + T\left(\frac{n}{2}, \textcolor{green}{m} - j\right)$$

(שתי המטריצות המקווקוות בכתום באירור מעלה). נבחן שהחקלים שמסומנים במשווה בירוק מסתכנים יחד $\frac{n}{4} \cdot m$. אם נמשיך את הרקורסיה הלאה, בסיבוב הבא נבעוד על 4 מטריצות שסכומן יחד $\frac{n}{8} \cdot m$, ועוד על $\frac{n}{8} \cdot m$ וכן הלאה. הטור הזה מתכנס ל-2, ולכן סיבוכיות הזמן לינארית.

עימוד מקומי – Local Alignment

האם יש תת מחרוזת s - t ותת מחרוזת t - s שיש ביניהן עימוד טוב? כלומר תת מחרוזת s - t ותת מחרוזת t - s שאם נוסיף אליהן עוד אותיות מימין ומשמאלו רק יגרעו מהצין.

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

מה השוני מהעימוד הגלובלי? לא נדבר יותר על התא האחרון במטריצה, כי אולי העימוד הטוב ביותר ממוקם בתא אחר, לכן צריך תמיד להחזיק את הערך הגבוה ביותר של V (שעד בה היה מוגדר העימוד הבן טוב מהירישה עד לנקודה זו). נגיד V חדש:

$$V_\ell[i:j] = \max \begin{cases} V_\ell[i-1, j-1] + \sigma(s_i, t_j) \\ V_\ell[i-1, j] + \sigma(s_i, -) \\ V_\ell[i, j-1] + \sigma(-, t_j) \\ 0 \end{cases}$$

מילוי טבלה:

	0	A	G	T
0	0	0	0	0
A	0	1	0	0
A	0	1	0	0
C	0	0	0	0
T	0	0	0	1

יש 3 תאים עם ציון מקסימלי. נבחן כי מטריצה כזו אינה שלילית, מחפשים את התא הטוב ביותר וממנו הולכים אחורה עד להגעה ל-0. מבחינת סיבוכיות זמן, זהה לאלגוריתם שראינו קודם. העלינו לדין גם את נושא עימוד של סיפה מול סיפה, רישה מול רישה או סיפה מול רישה (באמצעות היפוך).

Probabilistic Models and Decisions

הסתברות – דרך שלנו לתאר א-ודאות על העולם. איך נמדד אותה? לפי דגימת תכיפות של אירוע (frequency), לפי תיאור סובייקטיבי של האירוע..

המטרה שלנו בנושא זה היא שטבל להמשיך לדבר על הדוגמה הרצה מתחילה הקורס: בהינתן שני רצפים נרצה לדעת האם הם ממקור אבולוציוני. אומנם, בשיעור התחלנו להסביר את הנושא באמצעות דוגמה פשוטה יותר: נניח שיש לנו מדידות אובייקט שאפשר לאסוף על הרבה אנשים כגון נתוני בדיקת דם. אפשר לקבל בך שיעור של איך נראה אדם בריא: $(x) P_H$ באשר x זו ספירת דם של אובייקט (אלמנטי ו-ממדוי). בהינתן x נוכל לקבל את ההסתברות $\neg x$ מאובטשיה בזו או אחרת.

דבר על שתי שיטות הכרעה בבחינת היפותזות.

Two Hypotheses

באן נתעסק במקרה של איך נראה אנשים חולים מול אנשים בריאים, עם 2 הипOTESIS:

$$P_{Healthy}(x) \quad P_{Flu}(x)$$

על מנת לקצר, נסמן ב- $(x) P_H$ וב- $(x) P_F$. בעת נוכל עבור פרט לדבר על הערכות שמתקבלים (התפלגות לפי חוליה/בריא). איך נקבע את איכות ההחלטה שקיבלנו על סמך ההתפלגות? באמצעות **כל החלטה**:

$$\tau(x) \mapsto \{H, F\}$$

נוהג לסדר את סוג ההחלטות שהתקבלו גם בצורה טבלה באופן הבא:

		Predicted Condition	
		(predicted) Positive	(predicted) Negative
Actual Condition	(actual) Positive (P)	True Positive TP	False Negative FN (type II error)
	(actual) Negative (N)	False Positive FP (type I error)	True Negative TN

סיווג להתבוננות בטבלה: הצביע של המילה בתוך התאים של הטבלה נגזר מהעמודות. למשל נתבונן בתא השמאלי עליון **True Positive**, הנגזר מຕור זה שחזינו **Positive** (לכן יחקק), וה-**True** נובע מזיהה שהמצב האמתי הוא אכן **Positive** (לכן סגול).

נגידır **Positive** כאדם חולה (זה האובייקט אותו אנחנו בודקים). כל חוק הכרעה מתחשב בטעויות שנעשות, אם נקבע שכולם חולמים יהיה הרבה **FP** (טעית בכר שקבעתי שהם **Positive** למחלה, כי הם למעשה **Negative**). לעומת זאת, אם נגיד שכולם בריאות יהיו הרבה **FN** (טעית בכר שקבעתי שהם **Positive** למחלה, כי הם למעשה **Positive**).

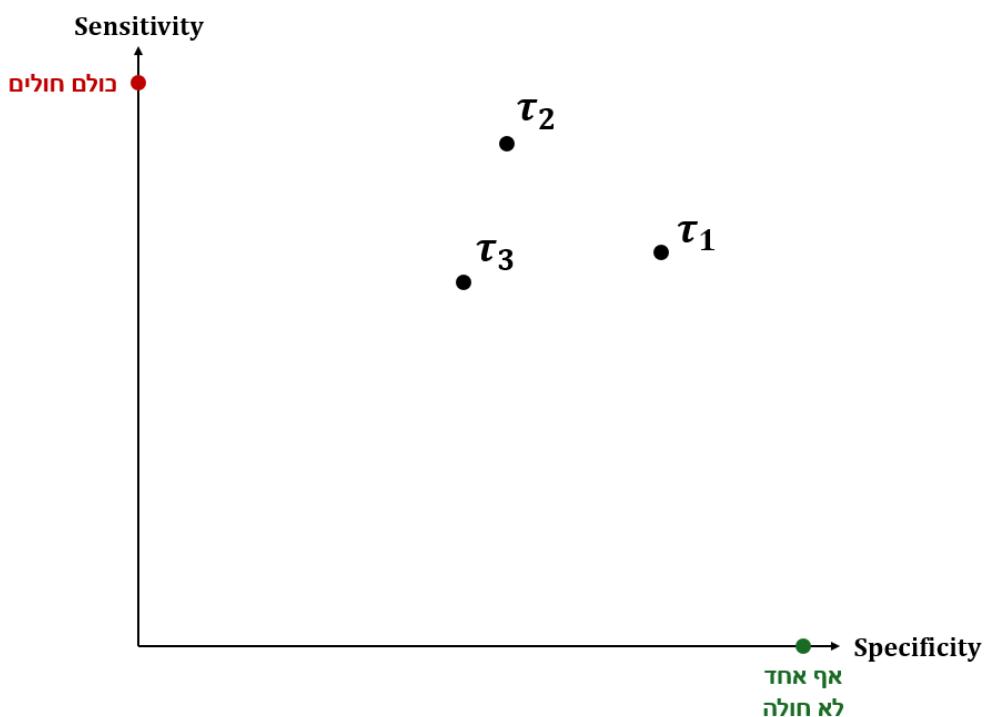
העלינו במה ייחסים מעוניינים:

$$\frac{TP}{P} - \text{שיעור החיזוי הנכון, סנסיטיביות.}$$

$$\frac{TP}{FP+TP} - \text{מתוך מה שהגדכנו חיובי, בכמה צדקנו? (ספציפיות)}$$

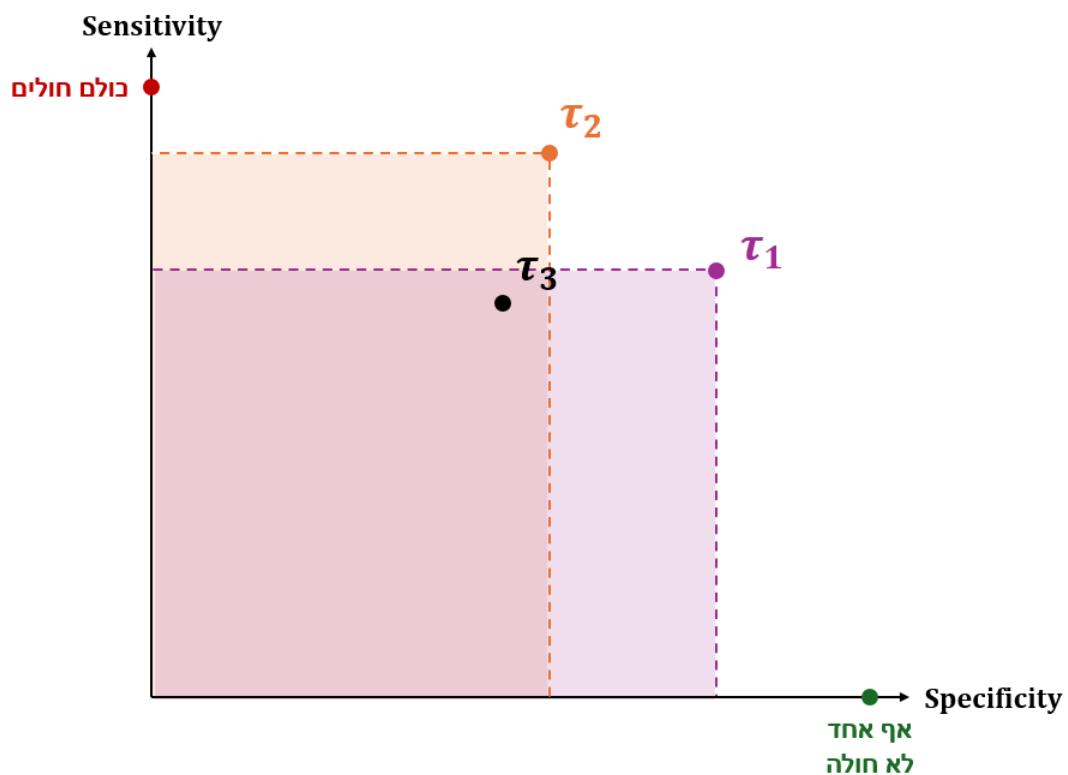
לפירוט המלא: [Sensitivity and specificity - Wikipedia](#)

אפשר להמחיש גם בגרף:

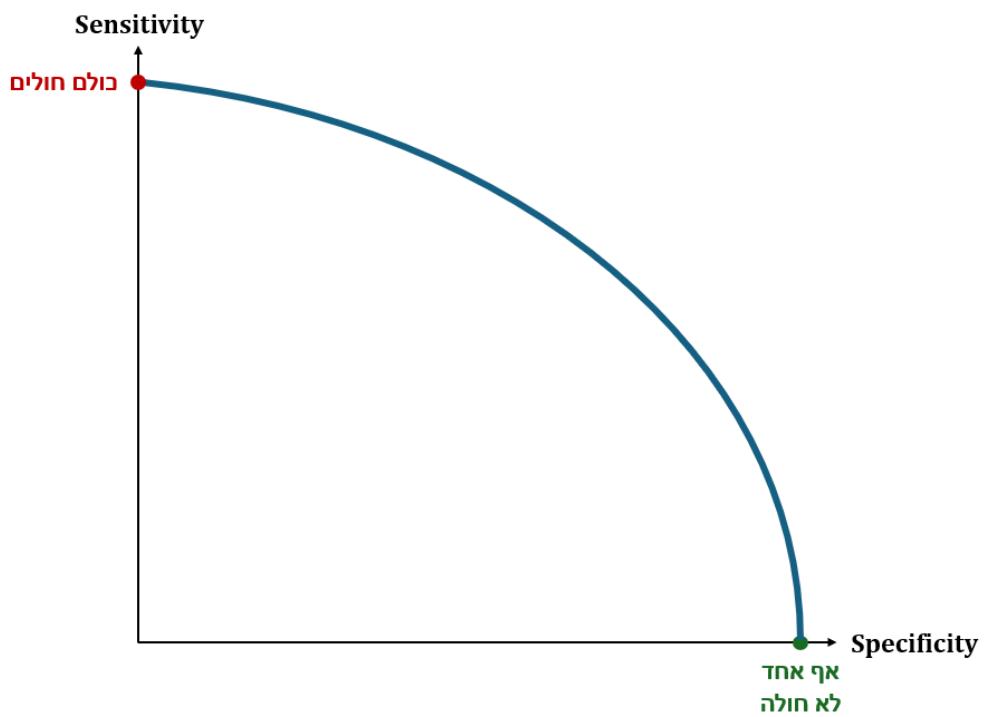


כל חוק החלטה הוא נקודה במרחב זהה (עם ביצועים בשני הממדים). האם נעדייף את τ_1 או את τ_2 ? תלוי במה שנדרש מאייתנו. לעומת זאת, לעיתים לא נעדייף את τ_3 על פני אחד מהם, כי הוא גורע מהם בשני הפרמטרים. בשלב זהה, זה שיקול של המשתמש מה חשוב יותר.

כך למעשה אפשר להציג את זה שכל נקודה במרחב מגדרה איזור (מיקווקו) שהוא טוב יותר מכל הנקודות השוכנות בתוכו:



בנחתה שהנתונים שבידינו מאפשרים ריצוף נקי של קווים. האיזור הכלוא מתחתיתו ניתן למימוש, אבל כל מה שמעליו בלחטוי ניתן להשגה:



למת נימן-פירסון (1933)

אם נשתכל על חוק החלטה בלשא:

$$\tau_t = \begin{cases} + & \frac{P_F(x)}{P_H(x)} \geq t \\ - & \frac{P_F(x)}{P_H(x)} < t \end{cases}$$

הזהת t (הסף/threshold) היא משחק بعد כמה מחמירים בדרישה להגיד שימושו בריא/חולה. הלמה טוענת שמשפחת החוקים זו מקיימת:

$$\tau \text{ is maximal} \Leftrightarrow \exists t \text{ s.t. } \forall x \quad \tau(x) = \tau_t(x)$$

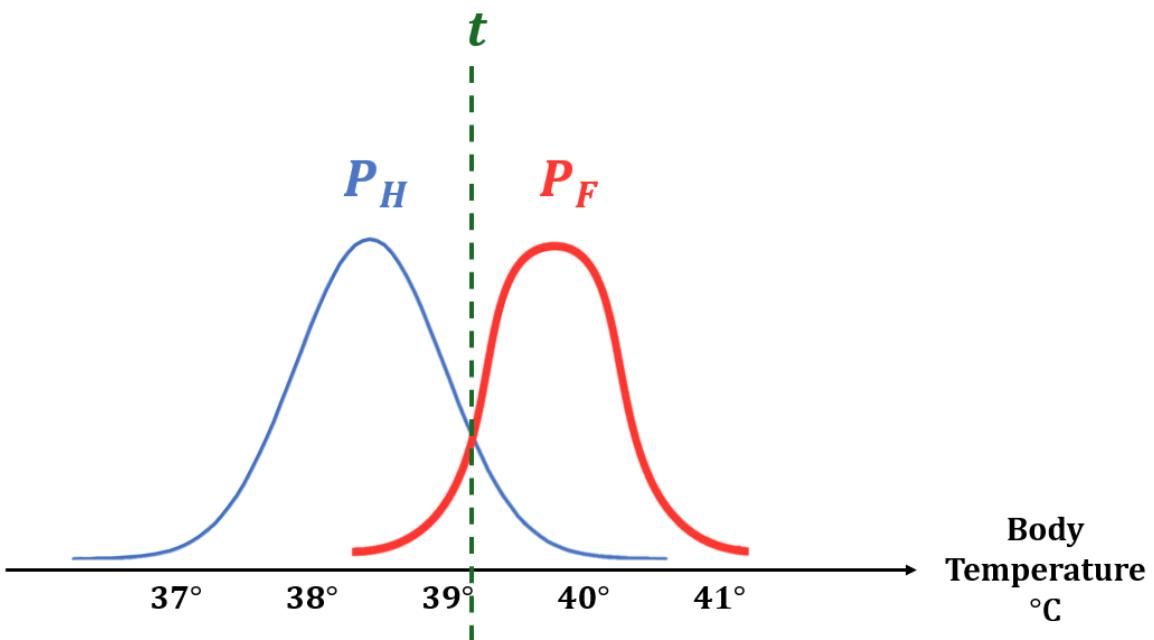
בולם, החוק מקסימלי אם קיים סף t שעבורו זה מתנהג כמו יחס הנראות. במקרה אחר, אם מציבים את חוקי המשחק (2 הסתברויות) כל החוקים שצריך לשקל בעולם יהיו מהצורה זו (אם לא, נהייה בתוך **האזור שמחחת לקו הבהיר** שראינו לעיל).

גישה ביידיאנית אומרת שבהינתן תצפית אפשר להבריע איך פרט מתנהג:

$$\mathbb{P}(+ | x) \stackrel{\text{Bayes}}{=} \frac{\mathbb{P}(x|+) \cdot \mathbb{P}(+)}{\mathbb{P}(x)} = \frac{\mathbb{P}(x|+) \cdot \mathbb{P}(+)}{\mathbb{P}(+) \cdot \mathbb{P}(x|+) + \mathbb{P}(-) \cdot \mathbb{P}(x|-)}$$

ואת זה אפשר להמיר ליחס נראות. ביטויים שימושיים במשוואה ולא הגדרנו:

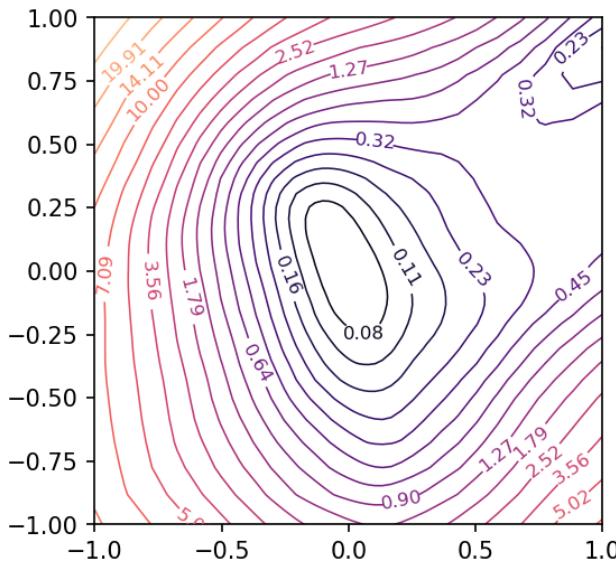
$$\mathbb{P}(x|+) = P_F(x) \quad \mathbb{P}(x|-) = P_H(x) \quad \mathbb{P}(+), \mathbb{P}(-) = \text{Pr}(\text{+}), \text{Pr}(\text{-}) \quad (\text{ידע מוקדים})$$



One Hypothesis

נתון לנו (x) ו אנחנו רוצים לבדוק השערות: H_0 – בריא, H_1 – לא. המבחןים האלה לא נתונים לנו אלטרנטיביה, אלא רק הכרעה האם מי שמולנו עונה על השערת האפס או לא. במקרה זה, יהיה לנו רק את P_H בדיאגרמה לעיל.

עבשו נוכל להחליט לדוחות את השערת האפס בהתאם לנחותים, למשל $P_H(\text{temp} > 38.5^\circ) = \overbrace{0.0001}^{\text{P value}}$. העניין מסתבר באשר הבדיקה שלנו היא ביוטר מממד אחד (לא רק חום גוף, למשל רמת כולסטרול, לחץ דם...). איך אז ייראה הגרף?



במקרים אלו נרצה לתרגם את הכל לחד-ממד, ויש יותר מדרך אחת לעשות זאת (נורמה, ממוצע..).
השאלות שצירר לשאול: מה ההתפלגות של P_{H_0} ? איך מתרגמים את הבעה לבעה הסתברותית בציר אחד?

חזרה לדוגמת הרცפים

בהינתן שתי סדרות t , s ו-2 היפותזות נרצה לשפטו ביניהן:

$$\begin{cases} H_0 & s, t \text{ are independent} \\ H_1 & s, t \text{ share common ancestor} \end{cases}$$

סימוניים:

$$P_{H_1} \equiv P_1 \quad \text{וגם} \quad P_{H_0} \equiv P_0 \quad -$$

p_0, p_1 תהיה דגימה של אותות

s_i ו- t_i האות ה- i ברצף.

הנחות:

- אורך בלשונו, לא נרשה עימוד מסווג indel.
- הרცפים הם d . i.e., ככלומר כל אות ברצף נדבגה באקראי ובאופן ב"ת. ההסתברות של רצף היא דגימתאותיות מאותו "סל".

לכן תחת הנחות אלו:

$$P_0(s, t) = P_0(s) \cdot P_0(t)$$

$$P_0(s) = \prod_{i=1}^n p_0(s_i) \quad P_0(t) = \prod_{i=1}^n p_0(t_i)$$

הערה: המודל זהה מבונן לא מציאותי, אבל הוא יכול ל佗צאות שימושיות ואז אפשר להחזיר להנחות המקלות של קחנו.

עבור P_1 נניח שמתקיים:

$$P_1(s, t) = \prod_{i=1}^n p_1(s_i, t_i)$$

הנחה אי-תלוות ונדבגה באקראי מתיקיות, אבל t_i, s_i בן תלוויות ונdagmo יחיד. נוכל למשל לקבוע:

p_1	A	C	G	T		p_0
A					0.3	A 0.3
C					0.2	C 0.2
G					0.2	G 0.2
T					0.3	T 0.3
	0.3	0.2	0.2	0.3	1	

מה שחייב להציג הוא שבכל שורה בטבלה של p_1 מתקיים $\sum_c p_1(T, c) = p_0(T)$, כלומר שהערכים נסכמים לאותו ערך כמו בטבלה של p_0 . בעת נוכל להגיע לפיתוח המרכז:

$$\frac{P_1(s, t)}{P_0(s, t)} = \prod_{i=1}^n \frac{p_1(s_i, t_i)}{p_0(s_i)p_0(t_i)} \stackrel{\log}{\Rightarrow} \log \frac{P_1(s, t)}{P_0(s, t)} = \sum_{i=1}^n \log \frac{p_1(s_i, t_i)}{p_0(s_i)p_0(t_i)}$$

מצביר פונקציית ציון עבור עימוד שנתקלנו בה בהרצאות הקודמות: $score(s, t) = \sum_{i=1}^n \sigma(s_i, t_i)$:

ולכן נגדיר:

$$\sigma(s, t) = \log \frac{P_1(s, t)}{P_0(s)P_0(t)}$$

למידה על הסתברות מנתונים

בاهינתן דאטה נרצה למדוד על התפלגות. הדאטה מורכבת מוקטורים של תצפיות. מספר נתונים:

$\vec{x}[1], \dots, \vec{x}[N]$: הדגימה ה- N .

המימד ה- i -י בדגם ה- i : \vec{x}_i .

מה זה אומר שהצלחנו למדוד טוב הסתברות? האם קיבל עוד דאטה מאותו מקור נצליח עליו בזורה טובה.

הטלה נעזים

דמייה להטלה מטבע, בנטרול הדעה הקודמת שיש לנו על מטבעות.

T



H



הדאטה נראה למשל בזורה הבאה: $x[1], \dots, x[N] \quad x[i] \in \{H, T\}$

נרצה לשערק את הסתברות המאורע H (זהה לשערק הסתברות T , מאורעות משלימים).

פרמטרים: נגדיר θ באופן הבא:

$$\mathbb{P}_\theta(H) = \theta \quad \mathbb{P}_\theta(T) = 1 - \theta$$

אפשר להגיד פרמטר גם בזורה אחרת, למשל:

$$\eta = \frac{\mathbb{P}_\theta(H)}{\mathbb{P}_\theta(T)} = \frac{\theta}{1 - \theta} \Leftrightarrow \theta = \frac{\eta}{1 + \eta}$$

$$\mathbb{P}_\eta(H) = \frac{\eta}{1 + \eta} \quad \mathbb{P}_\eta(T) = \frac{1}{1 + \eta}$$

לצורך הדוגמה ניצמד לפרמטר θ . נתחל בהגדרת פונקציה שמתארת איך הפרמטר מושפע על ידי הדאטה שלנו. הפונקציה בה השתמש היא לוגריתם הנראות:

$$L(\theta) = \prod_n \mathbb{P}_\theta(x[n]) \quad ; \quad \ell(\theta) = \sum_n \log \mathbb{P}_\theta(x[n])$$

נסתכל במקרה הטלת הנעץ עם הדאטה $\{H, T, T\}$:

$$L(\theta) = \mathbb{P}_\theta(H) \cdot \mathbb{P}_\theta(T) \cdot \mathbb{P}_\theta(T) = \theta \cdot (1 - \theta)^2$$

אנו מעריכים **בנראות המיטבית**, עקרון לפיו כאשר אנחנו מעריכים פרמטרים, נבחר את אלו שambilאים

$$\hat{\theta} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \ell(\theta)$$

נסמן $MLE := Maximum Likelihood Estimation$. נבחן כי פונקציה כדוגמת זו שיש לנו במקרה הנעץ תמיד תהיה מהצורה $\theta^{N_H} \cdot (1 - \theta)^{N_T}$ כאשר:

$$N_T = \sum_n \mathbb{1}\{x[n] = T\} = N - N_H \quad ; \quad N_H = \sum_n \mathbb{1}\{x[n] = H\}$$

זה נקרא **סטטיסטיק** – פונקציה שמתארת את הדאטה.

במקרה זה, שתיהן ייחד מרכזות את כל האינפורמציה על הדאטה. הסדר כאן לא משנה.

$$L(\theta) = \theta^{N_H} \cdot (1 - \theta)^{N_T} \quad ; \quad \ell(\theta) = N_H \log \theta + N_T \log(1 - \theta)$$

נרצה למצוא את המקסימום של הפונקציה ℓ . נגזר ונשווה לאפס:

$$\ell'(\theta) = \frac{N_H}{\theta} - \frac{N_T}{1 - \theta} = 0 \Leftrightarrow N_H(1 - \theta) = N_T\theta \Leftrightarrow N_H = (N_H + N_T)\theta \Leftrightarrow \boxed{\hat{\theta} = \frac{N_H}{N_H + N_T}}$$

מדוע רצינו את השלבים האלה? לאחר שבמידול של בעיות מורכבות יותר, עבורם דרך שלבים דומים.

דרך שיעורן בבייסיאנית

ל-maximum likelihood יש יתרונות שונים: תמיד אפשר להפעיל אותם, ה- MLE אינו תלוי בפרמטר (היה אפשר לעבוד גם עם $\hat{\theta}$), הוא אינו מוטה (אם ניקח הרובה **datasets** – הממוצע עליהם יהיה הנכון) והוא גם מדויק אסימפטוטית. **איפה הבעה?**

נחזיר במקרה הטלת הנעץ ונניח שבידינו דאטה מהצורה $\{T, T, T\}$ השערוך בדרך שראינו בעבר שהסיכוי לקבל H הוא 0, אין פה מקום לביטוי שלמו על ידע קודם. שיטת השיעורן הזאת תיתן את אותו הערך גם לדאטה מהצורה $\{T, \dots, T\}$ כר 100 פעמים – ברור לנו שתי קבוצות הדאטה שונות זו מזו.

שיטת הביסיאנית

בשיטת זו אנחנו לא מעריכים θ האמיתית של הנעץ, אלא מה אנחנו **חושב** שהוא θ הנכון. מתחילה עם מידע קודם על מהו θ ואיך מחשבים אותו בהינתן ההנחה הזו.

שימוש בהסתברות בעימוד רצפים

הדאטה שלנו יהיה מהצורה $x \in \{A, T, C, G\}^N$, כלומר $x[1], \dots, x[N]$ נגידיר 4 פרמטרים:

$$\vec{\theta} = (\theta_A, \theta_T, \theta_C, \theta_G) \in [0,1]^4 \quad s.t. \theta_A + \theta_T + \theta_C + \theta_G = 1$$

(היה אפשר להגיד באמצעות 3 פרמטרים אחד שהוא המשלים, אבל לצורך הדוגמה נניחו נם 4)

$$\mathbb{P}_{\theta}^{\rightarrow}(A) = \theta_A ; \mathbb{P}_{\theta}^{\rightarrow}(T) = \theta_T ; \mathbb{P}_{\theta}^{\rightarrow}(C) = \theta_C ; \mathbb{P}_{\theta}^{\rightarrow}(G) = \theta_G$$

$$L(\vec{\theta}) = \prod_n \mathbb{P}_{\theta}(x[n]) = \theta_A^{N_A} \theta_T^{N_T} \theta_C^{N_C} \theta_G^{N_G} \quad \ell'(\vec{\theta}) = N_A \log \theta_A + \dots + N_T \log \theta_T$$

חקר הפונקציות מסתבר – כי מדובר בעצם בפונקציה רב-מדנית. צריך שהגזרת החלקית בכל ביון תהיה 0:

$$\frac{\partial}{\partial \theta_A} \ell(\vec{\theta}) = 0, \dots, \frac{\partial}{\partial \theta_T} \ell(\vec{\theta}) = 0$$

עלינו לפתור את הבעיה $\max_{\theta} f(\vec{\theta}) \quad s.t. g(\vec{\theta}) = 0$ והדרך dazu היא באמצעות כופלי לגראנדי:

$$J(\vec{\theta}, \vec{\lambda}) = f(\vec{\theta}) - \vec{\lambda} g(\vec{\theta}) = N_A \log \theta_A + \dots + N_T \log \theta_T - \lambda(\theta_A + \dots + \theta_T - 1)$$

עבור למשל $\hat{\theta}$ קיבל: $\frac{\partial}{\partial \theta_A} J = \frac{N_A}{\theta_A} - \lambda \Rightarrow \hat{\theta}_A = \frac{N_A}{N}$ וכן גם עבור השאר

$$\frac{\partial}{\partial \lambda} J = \theta_A + \dots + \theta_T - 1$$

בשעושים אירע מולטינומי, הסטטיסטים עומדים להיות במא פעמים ראייתי כל תוצאה, ולכן הנראות המקסימלית תהיה באופן הבא:

$$\hat{\theta} = \left(\frac{N_A}{N}, \frac{N_T}{N}, \frac{N_C}{N}, \frac{N_G}{N} \right)$$

עימוד רצפים

ברצוננו לשערק 2 התפלגיות P_0, P_1, P_2 . עבור:

$$P_0(x) \quad x \in \{A, T, C, G\} \text{ or } x \in \{\text{Amino acids}\}$$

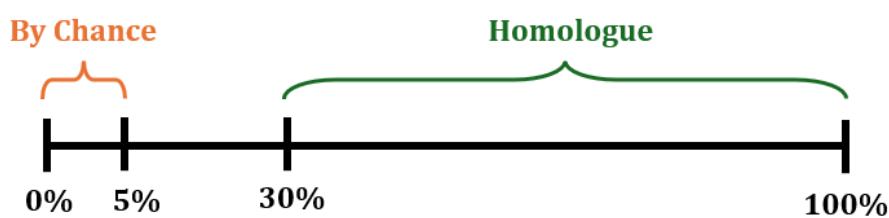
בhinint 20 רצפים והנחה שכל אות נדגמה באופן בלתי תלוי. לשערק נראות מקסימלית משמעotta למספר כמה פעמים כל אות הופיע. אילו רצפים נכניס לסט האימון?

במקרה של (y, x) במקום למדוד 4 אותיות נלמד 16 זוגות. איפה זה מסתבך? בעית העימוד נולדה כשהיו רק 20 רצפים של חלבוניים ומעט מאוד DATA לאטיה לגבייהם. היום, במקרים מסוימים שקיבלנו מהעולם מסביב היא בקנה מידה עצום.

עבור P_0 נוכל פשוט לספר מתוך הדאטה, אבל הדאטה שלנו לא בצורה שירוטית. כך למשל, חלבון ספציפי יכול להופיע בדאטה פעמים רבות בויריאנטים שונים מהסיבה הפשוטה שהחלבון הספציפי הוא מוקד מחקר חשוב (למשל בחידוך איקולי). חלבון זה יקבל יותר משקל.

הינו רוצה להיפטר מדברים שהם פשוט הכפלה – זה תחום שנקרא **ביו-איןפורטטיקה**, התחום מתעסק בעובדה על ממד נתונים וניתוח החזרות מהם כדי שנישאר עם ממד נתונים קרוב ככל הניתן *i.e.*

מה לגבי למידת P_1 ? איך נדע שני רצפים בעלי אב קדמון משותף?



אם יש לנו לפחות 30% זהות אפשר להגיד שהחלבוניים **הומולוגים**. יש הרבה זוגות הומולוגים (מקור אבולוציוני משותף) גם בטווח שמתוחת ל-30%, אבל לא נוכל להבדיל אותם מהאחרים שאין ביניהם כל קשר. אם ניקח טווח בסקירה של אחוזי דמיון, למשל הטווח 60%-50% אנחנו די בטוחים שהם מקור משותף – אך הם לא זיהים. נוכל לחפש אותם, להגדיר באמצעותם את P_1 ופונקציית ציון, ואז להתחיל לחפש רצפים באחוזים נמוכים יותר של דמיון.

עימוד לפי עמדה בחלבון

רוב החלבוניים בגוף יכולים להיות ב-2 מדויומים שונים: מדיום נזלי (דם, ציטופלזמה), או דרך מברכנות. הסבירה משפיעה על מבנה החלבן ועל סדר הרצפים. מכאן נוצרת גישה נוספת לפיה אפשר למדוד מטריצה שמתיחסת לחלק ספציפי בחלבון, ועימוד רצפים באמצעות $(t, s)_\sigma$ לפי עמדה.

הערה: עד עכשוו הטעלמנו מ-indel. אפשר להתייחס לכך לפי כמה פעמים ראיינו – מול אותן. בספרות בדרך כלל נהוג לתת קנס שונה בין ה-indel הראשון לבין המשכת רצף של indels.

שבוע 3 – הרצאה 5 (הרצאת אורח)

הרצאת אורח: ד"ר ברק רווה, מידול אינטגרטיבי של מערכות ביולוגיות

הקדמה

מושגים בסיסיים

מערכת – אוסף רכיבים הקשורים ביניהם וمتפקידים יחד למען מטרת משותפת.

מערכת דינמית – מורכבת מרכיבים, כוחות וחוקים שמתארים איך הכוחות משתנים לאורך זמן בעקבות אינטראקציות בין החלקים.

דוגמה קלאסית לדינמיקה היא חוקי ניוטון.

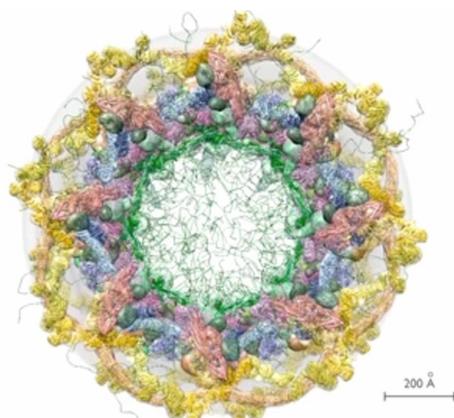
דיון – מהם חיים?

רעונות שהולו בכיתה: כוחות שמתקנדים לסביבה, שימור סביבה פנימית, שכפול ורביה, מוכנות היישרות, אדפטיביות, הנעה עצמית (אולי לوكאלית).

דעה מקובלת במדע: חיים – אוסף של חלקים הפועלים ייחודי על מנת לקיים את יסודות החיים:

- לאסוף מידע, אנרגיה, חומרי בניין
- לקבל החלטות
- לבצע תוכניות סדרות
- לאלתר + לבצע התאמות
- לגודל ולהתרבות
- להישמר מהסביבה

מערכות ביולוגיות הן מורכבות ודינמיות בהגדתן, ולכן נרצה למדל (ליצור "זגט") של מערכת ביולוגית דינמית. במהלך הרצאה נעבד עם דוגמה של מבנה הקאים בגרעין התא ומהו מסקן בין הגרעין לציטופלזמה – טרנספורט מולקולרי.



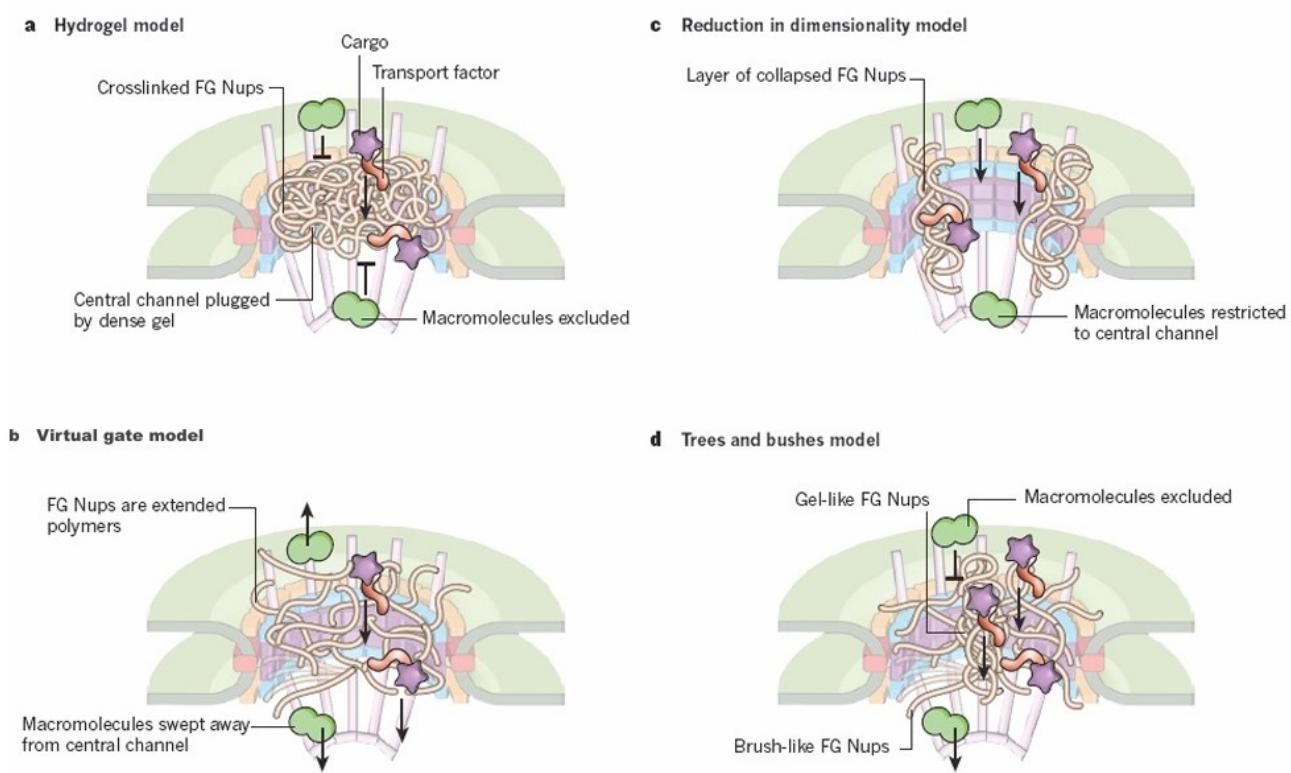
למה צריך מודלים והדמיות? לחיזות להסביר תכונות, להכריע בין היפותזות, לשולט בתכונות המערכת, לזהות פעuri מידע ולאתר מידע חסר.

מודול אינטגרטיבי של מערכות ביולוגיות דינמיות

מידול = בניית מודלים. נרצה לפשט את המודל, אבל לא להגיע למצב של פישוט יתר (מודל פרות כזריות שנענות בוואקום, הוא רחוק מדי מהמציאות). נדרש תיאור היררכי.

נבוד לארוך ההרצאה עם דוגמה – מודל של תנוצה מולקולרית דרך חירויות הגרעין. החירויות מאשרות מעבר מהיר,iesel וסלקטיבי בין חלקו התא. באפשרותו מתבצע מעבר פסיבי של מולקולות קטנות (יונים, מים) במקביל לסינון של מולקולות גדולות שלא נקשרות ל-FG Repeats, ואלו יישארו מחוץ לגרעין. סינון זה מאפשר תקשורת בין ה-DNA בגרעין לבין שאר התא.

במהלך השנים פותחו כל מיני תאוריות על איך עובד המנגנון, והוצעו היפותזות שונות בנושא:



יש צורך במודל שהוא מעבר לתיאור במילימ', אבל הקושי הוא שאין טכנולוגיה שמאפשרת לנו פשוט להסתכל ולראות איך המנגנון עובד באמת כדי ליצור מודל זהה. מודל אינטגרטיבי לוקח בחשבון סוגים שונים של מידע – ניסוי ותיאורטי, וմבגד אותו לאורך פרקי זמן שונים.

שלבים ביצירת מודל

1. איסוף מידע –

בשלב זה לא נטריד את עצמנו באיר נמדל את המידע שנאנסף. המידע יכול להיות מכמה סוגים: מידע ניסויי – בשלב זה נבחן מקורות רעש, איפה עלולות להיות טעויות, סטיות תקן ומידע ניסויי נוספת. מידע תיאורתי – מודלים שוכנים, אינטואיציות מדעית.

2. ייצוג מתמטי של המידע והמערכות במחשב –

יש צורך לייצג במחשב את הרכיבים, האינטראקציות, והדינמיקה (חישובי כוחות). בשלב זה מבניםים מימד של פשוטות, למשל החלפת אלף אטומים שייצגו ככדור אחד. ייצוג האינטראקציות יכול להתבצע בעזרת פונקציה מתמטית פשוטה. ייצוג דינמיקה (динמיקה ברואנית – איפה המערכת תימצא בשלב הבא). בשלב זה יתכן שנחזר לשלב הקודם ונשלים מידע נוסף שcharה. לבסוף יבוצע טיב הנ吐ונים מהעולם האמיתי לייצוג במחשב.

3. מציאת מודל טוב של המערכת –

סורקים את האופציות של הפרמטרים, משפחות היפומות. בשלב זה נדגום פרמטרים שיביאו לאופטימיזציה של המידע.

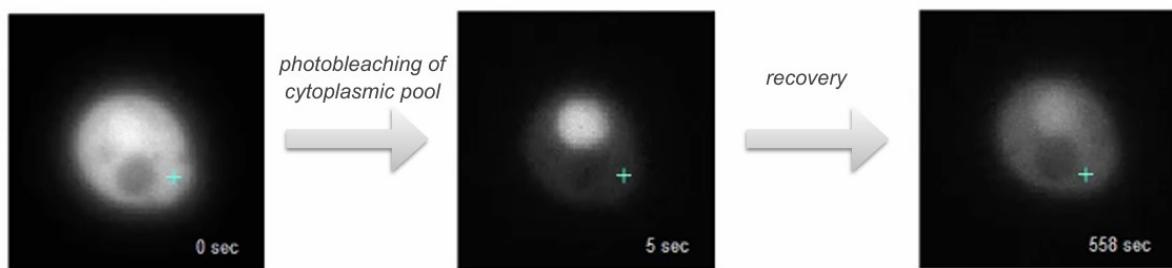
4. אישוש נכונות ואיכות המודל והפקת תובנות חדשות על המערכת –

באמצעות מידע ב"ת תבצע ואילידציה של התוצאות. בשלב זה ננסה להבין אילו אירועים לא טריוויאליים המודל מצליח לחזות. היתרון של מודל כזה על פני מודל נאיבי הוא האינטראקטיביות שלו – זה לא מודל מתמטי או LM-ו טהור (כמו רשת נירוניים) שבו לא ניתן לבצע שינויים שմגדמים את המערכת האמיתיית, ובמודל זהה יוכל לשנות פרמטרים כמו הבנחת מוציאות לחלבונים, למחוק שרשרות. מה שלא אפשרי במודלים אחרים.

שימוש במודל

מה המודל טוען לגבי הסתננות של מולקולות לפי גודן?

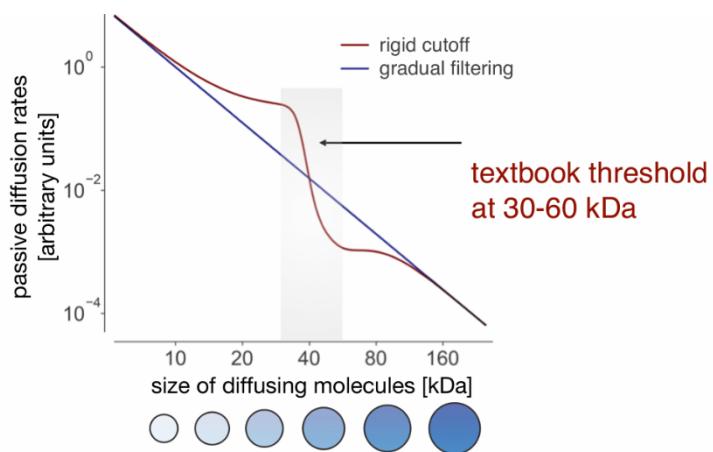
כל מולקולות GFP בцитופלסמה עברו Photo Bleaching ולכן תוך 5 שניות כל היציטופלסמה הפסיקה לזהור:



כעבור מספר דקות (תמונה ימנית) מבוחנים שרמת הזהור בגרעין ירדה, והיציטופלסמה זהרה יותר. זה הוכיח ניסיית מעבר חומרים מהגרעין ליציטופלסמה.

בניסוי גם נמצא שלגודל תא השמר השפעה על קצב חילוף החומרים בתא. האם המודל מצליח לחזות את זה? במודל הסופי היה צורך להכפיל את התוצאות קבוע מסויים, זה פרמטר חופשי ולרוב צריך לתקן את התוצאות כדי להגיע לאותם ערכים מסוימים.

בשיעור דנו בסוגיה איזה קו מגמה מתאר טוב יותר את קצב הפעוף לפי גודל מולקולות:



הניסוי מלמד אותנו שאי Cutoff באוזר kDa 40. מבחינת הסימולציה אנחנו די סומכים על המודל, וambilim שהוא יותר דומה להשערה של שער וירטואלי שהוצגה לעיל. אם מצליחים לתאר את המודל עם כמה שפחות הנחות מפשטות, נראה שהמודל יותר מתקרב למציאות.

לסייע, למדנו שמודול אינטגרטיבי שלוקח כמה סוגים של מידע מאפשר להתמודד עם המורכבות של מערכות ביולוגיות – אבל חייבים לעبور שלבים של אליידציה.

היריסטיות לעימוד רצפים

לעתים הפתרון המוכח הוא יקר, ראיינו שאפשר להגיע לזמן קווואדרטיים בגודל הקלט. חלופה היא מציאת פתרון ביןים – שיטות היוריסטיות שלא מבטיחות לעמוד בכל המקרים, אבל להתקבב לוצאה הטובה ביותר. דיברנו על בעית העימוד הגלובלית, וشكילת כל המסלולים האפשריים. גישה היוריסטית אומרת אחרת – לעימוד "טוב" יש תכונות מסוימות, אפשר לנצל אותן.

גישה ראשונה – FASTA

נסתכל על שתי תכונות של עימוד שאנו מגדירים בעימוד טוב:

1. מזעור במקומות ה-gaps
2. שאיפה לזהות בעמדות ("אימס" בהם יש זהות).

איך ההנחות האלה עוזרות לנו?

למשל בהינתן שני רצפים: $m = |t|, n = |s|$ ורמטר k , נרצה למצוא כמה מילים באורך k משותפות בין הרצפים. שני חסמים על כמות האפשרויות: אם גודל הא"ב הוא Δ אז יש Δ^k אפשרויות לפחות, ואנו גם חסומים בגודל הרצפים (m, n). האם העבודות הללו עוזרות לנו למצמצם בדמן? כן – נוכל להגיע בזמן לינארי על ידי מצמצם כל המילים שמופייעות בשני הרצפים ובאיזה אינדקס הופיעו. נשים לב שהאינדקס בכל אחד מהרצפים j, i (ברצפים s, t , בהתאם) מגדיר לנו $j - i = \Delta$ שמדובר רטט על ה"אלבסון" או "שיפוע" בטבלה של התכון הדינמי. אם סופרים כמה פעמים מצאנו כל Δ , אפשר למצאו את האלבסטונים שדרכם מומלץ להעביר את העימוד.

שיטת **FASTA** דוגלת בהתבוננות על מילים קצרות, חישוב Δ , וה- Δ שמופיעה הרבה תהיה מועמדת טובה לאלבסון שמננו מתחילה.

הדוגמה זו היא בעיקר עבור האינטואיציה. איך נוכל לשפר את התכונות שדרשו למעלה?

BLAST

נרצה לשנות מעט את התנאי השני: שאיפה **להזות לדמיון** בעמדות.

ב- k קטע ב-FASTA נמצא גם הרבה רעשאים אקראיים. נרצה מילים ארוכות יותר ולכן את דרישת הזהות בדרישת הדמיון. גם את התנאי החדש אפשר לשפר.

במקום דרישת לדמיון (מושג עמוס מעט), נרצה שאיפה **לציוון טוב על פי פונקציית score**. נרצה לקבל התאמות שיקבלו ציון גבוה ונוכל להשתכנע שיש בינהן קשר אמיתי.

לפעמים שינאים לא עולים הרבה ולכן הבחירה להחליף את דרישת identity בדרישת high score. איך האלגוריתם שלנו ישתנה? למשל אם z מילה שנרצה למצוא לה עימוד מול מיל מסד הנתונים אז ניקח k ונשאלו מהן המילים שהעימוד בגודל k מקבל ערך ציון גבוה.

אנחנו עתדים למצוא מילים שיכולים למצוא ציון עימוד גבוה מספיק. שיטה זו מצמצמת את כמות המילים שאנחנו מעוניינים בהן למספר הרבה יותר كثير לעומת FASTA – כי הן יחסית מדיות. לאחר מכן, כאשר בידינו הקטיעים הרלוונטיים נוכל לבצע **עימוד לokaל** החל מהנקודה שלהם (השלמת שאר הרצף).

שרשות מركוב

$$\text{עד עבשו דיברנו על מודל רצפים שנחשב יחסית "בנאלית" - } \mathbb{P}(x_1, \dots, x_n) = \prod_i \mathbb{P}_0(x_i | x_{i-1}, \dots, x_1)$$

יש כאן הנחה שככל האותיות נציגות התפלגות ובאופן בלתי תלוי. לפעמים נרצה ליצור מודלים מורכבים יותר – והדרך לשם לאורך הקורס תהיה בשלבים: ככל שנרצה מודלים יותר מורכבים, כך נדרש לעבוד קשה יותר כדי לבסס אותם.

רצף של משתנים מקרים/אותיות/אירועים על פני זמן – נהוג לתאר בעזרת **אירועים מركובים**.

שרשת מרכיב – בדרך כלל מוגדרת על ידי תכונת מרכיב:

$$\forall i \quad \mathbb{P}(x_{i+1} | x_1, \dots, x_i) = \mathbb{P}(x_{i+1} | x_i)$$

העריכים x_{-1}, \dots, x_1 לא משפיעים על המאורע ה- i . אם המאורע ה- $i+1$ מופיע אוטומטית, כל ערך שהוא תחילה בגודל 1 – i לא רלוונטי כל עוד יש בידנו את הערך ה- i . שימוש נפוץ הוא למשל בהילוך שיבוב.

שרשת מרכיב אחידה – מכילה את התכונה הבאה:

$$\forall i, j \quad \mathbb{P}(x_{i+1} = a | x_i = b) = \mathbb{P}(x_{j+1} | x_j = b) \stackrel{*}{=} \tau[a, b]$$

במילים פשוטות, הכוונה היא שהסבירו של השיבוב לוז מקומות אחד לשני אינם תלויים באורך הדרך שעבר (הוא לא "מתעייף").

עם שתי ההנחהות האלה, נחזור לרצף שלם (x_1, \dots, x_n) ונטען כי:

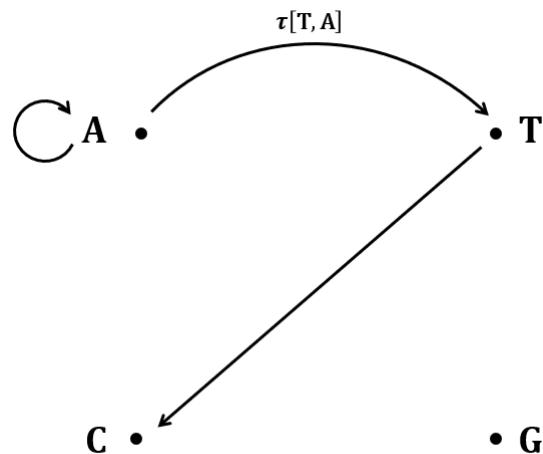
$$\mathbb{P}(x_1, \dots, x_n) \stackrel{\text{שרשת}}{=} \mathbb{P}(x_1) \cdot \mathbb{P}(x_2 | x_1) \cdot \dots \cdot \mathbb{P}(x_n | x_1, \dots, x_{n-1}) \stackrel{\text{Markov}}{=}$$

$$= \mathbb{P}(x_1) \cdot \prod_{i=1}^{n-1} \mathbb{P}(x_{i+1} | x_i) \stackrel{\begin{array}{c} \text{ההתפלגות אחידה} \\ \text{שינוי בנסיבות} \end{array}}{=} \mathbb{P}_0(x_1) \cdot \prod_{i=1}^{n-1} \tau[x_{i+1}, x_i]$$

מטריצת זוגות אותיות שמנדרה את הסיכוי לאות אחת בהינתן אחרת.

State Diagram

גרף שהקודוקדים בו הם אותיות, ותהייה קשת בין קודוקדים אם יש סיכוי לעبور ביניהם.



בנוקלאוטידים זה בнерאה פחות מעניין, אבל יהו סיטואציות בהן בן יהיה בכר שימוש.

איפה שרשראות מركוב יישמשו אותנו? אפשר להסתכל על אבולוציה בתהיליר מרקובי באשר המצב בכל נקודה הוא מה הרץ שיש לו עד בה. בדרך כלל התובנות אותן אחת אחרה בחלון נותנת יחסית מעט מידע, אנחנו משתמש בשרשראות מרקוב לא ברמת האות הבודדת.

רקע ביולוגי להמשך

בחלקים נרחבים בגנים יש אזורים שלא מקודדים לחלבון והם נחשבים לאזרוי בקרה. בהמשך תהיה הרצאת אורח שטטרטה להסביר מתי להשתמש באזורי בגנים. תופעה שנקראת מטילציה של DNA – שינוי שקרה באאות C שהופכת ל- C^* . אחד הדברים שקרו באבולוציה נבעה מה הצורך להעביר את המידע לדור הבא. בגדייל DNA מול C יש G, אבל נרצה שגם את C^* תהיה לנו דרך להעיבוד. הפתרון האבולוציוני הוא להסתכל על C שאחריו מגיעה G ואז בכיוון הפוך הרץ יזהה בצורה תקינה – האנדים ישים את הסימון על האות C גם בגדייל המועתק. תופעה אבולוציונית היא שאנדימים שמסמנים מטילציה ההתחלו להעיף C שאחריהם מגיעה G, ולסמן אוטומטית.

אם מה, בשיטמים את הסימון משנים מעט את C ובכך חושפים אותו יותר למוטציות – גובר הסיכון שנאנבד את C, מה שהוא שמנוביל לאיבוד fitness. נוצר מצב עם אזורי בגנים בהם הסיכון שאחרי C באה G הוא גדול – להם נקרא **עשירים ב-G**, ומנגד אזורי בהם הסיכון שאחרי C באה G הוא דל ולחם נקרא **עניים ב-G**.

איך הינו בונים מטריצת מעבר לפי האוזור עשיר/עני ב-Gcp? נגידושתי מטריצות שיבדלן ביניהן בתא

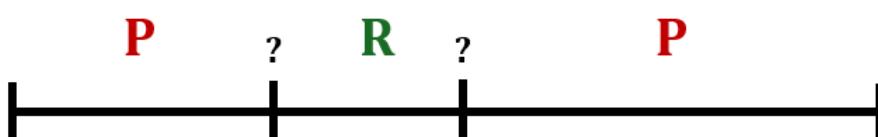
המסומן:

τ	A	C	G	T	
A					
C			█		
G					
T					

לפעמים האוזור המסומן יהיה נמוך, ולפעמים קצר יותר גבוה מ- $\frac{1}{4}$.

איך נשתמש בזה? אפשר לנקח רצף ולחשב לפי 2 שרשרות מركובות שונות. נשתמש בסוף גראות כדי להחליט איזה אוזור יותר סביר יהיה (עשיר/עני ב-Gcp).

צריך לדעת להגדיר/لت חום את האוזור – בתקווה באמצעות מודל הסתברותי שיבצע זאת אוטומטית:



מודל אחד שנדבר עליו יהיה מודל **תהליך חבוי**:

תהליך חבוי – נניח שנרצה להוסיף סיוג לכל עמדה ברצף x_1, \dots, x_n משתנה מקרי חבוי H_1, \dots, H_n שמדווח על סטוס העמדות.

Poor/Reach	P	P	P	P	R	R	R	R	...
	C	T	A	A	C	G	C	G	...

מודל מركובי חבוי הוא זהה שבו נניח שמתיקיימת שרשרת מركוב: $(H_1, \dots, H_n) \mathbb{P}$ כלומר שההתווויות החבויות מתוארכות על ידי שרשרת מركוב. סביר שנשאר באותו מצב $R \setminus C$ ואת המעברים בהםם אפשר לשירות לארוך אופייני של אוזור עשיר או עני.

יהיה מודל שיודיע להגדיר $(H_n, \dots, H_1 | x_1, \dots, x_n) \mathbb{P}$. ברגע שיש הסתברות משותפת, תיארנו מה קורה בעולם. נשים לב שהאובייקט מורכב ولكن נרצה להניח עליו הנחות:

$$\mathbb{P}(x_1, \dots, x_n | H_1, \dots, H_n) = \prod_{i=1}^n \mathbb{P}(x_{i+1} | x_i, H_i)$$

עדין יש סימני שאלה – איך משתמשים במודל זהה?

מודלים מركוביים חבויים – Hidden Markov Model

כפי שראינו בשיעור הקודם, מודל מركובי חבוי זה סט של משתנים/ שרשרת תצפויות $\mathbb{P}(H_1, \dots, H_n | x_1, \dots, x_n)$, אשר ההנחה היא שגם שרשרת מרקוב: $\mathbb{P}(x_1, \dots, x_n | H_1, \dots, H_n)$

$$\mathbb{P}(x_1, \dots, x_n | H_1, \dots, H_n) = \prod_{i=1}^n \mathbb{P}(x_i | H_i)$$

ויזואלית, זה אומר הממצבים החבויים הם שרשרת מרקוב שמתקדמת בזמן, בכל שלב ההתקדמות תליה רק במצב האחרון (ולא איך הגיענו אליו), אפשר להתייחס לה באוטומט סופי דטרמיניסטי שمبוסס הסתороות. התצפויות (α -ים) הן תצפויות שתלויות במצב הנוכחי. אפשר לחשב על זה כך שהמצב החבוי זה מצב כל השחקנים במשחק, והתצפויות זה רק מה שהשחקן מגלה לנו באותו רגע על המסך (תלוי במצב, אך לא גלי לנו). לעיתים התצפויות רוחשות וכן התצפית היא בהינתן המצב החבוי.

דוגמה

משחק הימורים מול שחזור שטוטיל קובייה, והוא טוען שיש לו קובייה חוקית. בלי שנשים לב הוא מחליף את הקובייה לקובייה מוגה, והוא לא תמיד יכול לבצע את ההפחה (רק בשאנחנו לא שמים לב). זה "hidden" כי אנחנו לא יודעים מה הקובייה שהיריב מחזיק, אלא רק סדרת תצפויות.

סימונים:

$$obs \in [1, \dots, 6], loaded - L, fair - F, hidden - H$$

הגדרת המודל: $\hat{\mathbb{P}}_0(H_1, \dots, H_n | x_1, \dots, x_n) = \frac{\star}{\mathbb{P}_0(H_1) [\prod_{i=1}^{n+1} \tau(H_{i+1}, H_i)] \cdot \prod_i^n \pi(x_i, H_i)}$. מתאר משחק הסתברותי שימושיו מקדם את ה-*hidden state* *hidden state* *hidden state* *hidden state*.

שימושים במודל:

1. בהינתן תצפויות, מה היה המצב בזמן i ? $\mathbb{P}(H_i | x_1, \dots, x_n)$
2. מהו $\max_{h_1, \dots, h_n} \mathbb{P}(H_1 = h_1, \dots, H_n = h_n | x_1, \dots, x_n)$? $\arg \max_{h_1, \dots, h_n} \mathbb{P}(H_1 = h_1, \dots, H_n = h_n | x_1, \dots, x_n)$ בambilם אחרות, מהו רצף הממצבים החבויים הכי סביר בהינתן התצפויות? זה עית שחזו. בשאלת 2 מענין אותנו רק המצב החבוי הנוכחי.
3. מהו הסיכוי לראות את התצפויות הללו? $\mathbb{P}(x_1, \dots, x_n | H_1, \dots, H_n)$. בתת-בעיה של בעיה 2 נהיה חייבים לחשב את הנראות.
4. האם אפשר למצוא את Π ואת τ ? נראה בשיעור הבא

נתחיל בשאלת התצפית (3). לפי הסתברות שלמה:

$$\mathbb{P}(x_1, \dots, x_n) = \sum_{h_1} \dots \sum_{h_n} \mathbb{P}(H_1 = h_1, \dots, H_n = h_n | x_1, \dots, x_n)$$

הבעיה היא שיש כאן סכום על n ערכים, כלומר מספר אקספוננציאלי $-n$ של מקרים. אחת הדריכים לפטור את זה היא לנקח את הביטוי שסימנו ב-*, ולהתחל לփועל חוקים אלגבריים. אנחנו נציג דרך אחרת באמצעות תבנון דינامي. לשם כך, נכתב את הבעיה ללא הסכום האקספוננציאלי.

$$\mathbb{P}(x_1, \dots, x_n) \stackrel{\text{הסתברות}}{=} \sum_{h_n} \mathbb{P}(H_n = h_n, x_1, \dots, x_n)$$

$$\mathbb{P}(H_n = h_n, x_1, \dots, x_n) \stackrel{\text{הסתברות}}{=} \sum_{h_{n-1}} \mathbb{P}(H_n = h_n, H_{n-1} = h_{n-1}, x_1, \dots, x_{n-1})$$

$$\begin{aligned} &= \sum_{h_{n-1}} \mathbb{P}(x_n | H_n = h_n, H_{n-1} = h_{n-1}, x_1, \dots, x_{n-1}) \\ &\quad \cdot \mathbb{P}(H_n = h_n | H_{n-1} = h_{n-1}, x_1, \dots, x_{n-1}) \cdot \mathbb{P}(H_{n-1} = h_{n-1}, x_1, \dots, x_{n-1}) \end{aligned}$$

נשים לב **שהביטויים בכחולים** מזכירים זה את זה, פחות או יותר אחד. זה אומר שםណו נדע לפטור את יהה פתרון גם לשנייה. נזכיר שאנו בשרות מרקוב, כלומר בהינתן המצב הנוכחי H_n לא תלוי במה שקרה קודם, כולל תוצאות של המצבים ולבן בשאר עם:

$$\begin{aligned} &\sum_{h_{n-1}} \mathbb{P}(x_n | H_n = h_n) \cdot \mathbb{P}(H_n = h_n | H_{n-1} = h_{n-1}) \cdot \mathbb{P}(H_{n-1} = h_{n-1}, x_1, \dots, x_{n-1}) \\ &= \sum_{h_{n-1}} \pi[x_n, h_n] \tau[h_n, h_{n-1}] \cdot \mathbb{P}(H_{n-1} = h_{n-1}, x_1, \dots, x_{n-1}) \end{aligned}$$

נגדיר אובייקט: $F[i] = \mathbb{P}(H_i = s, x_1, \dots, x_i)$. הפיתוח מראה שאפשר לחשב (שינוי נוטציה):

$$F_i[s] = \pi[x_i, s] \sum_t \tau[s, t] \cdot F_{i-1}[t]$$

בפועל יש לנו את המטריצה הבאה:

	0			i		n
F				↗		
L				↘		

בשביל למלא את ההסתברות במצב i אפשר לשאול על המצב $(i-1)$.

קיבלנו פונקציה רקורסיבית אבל טרם הגדרנו תנאי התחלה. נגידו:

$$F_1[s] = \mathbb{P}(H_1 = s, x_1) = \mathbb{P}_0(s)\pi[x_1, s]$$

נinth דוגמה מספרית:

		From		π	F	L
		F	L			
τ_0	F	0.99	0.05			
	L	0.01	0.95			

π	F	L
1	$1/6$	$1/10$
2	$1/6$	$1/10$
3	$1/6$	$1/10$
4	$1/6$	$1/10$
5	$1/6$	$1/10$
6	$1/6$	$1/2$

עם התוצאות: $sdp = [2,6,6,6,6,6]$. אנחנו מניחים שאנו לא יודעים איזו קובייה היריב מחזיק בהתחלה.

		1	2	3
τ_0	F	$\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$	♣	
	L	$\frac{1}{2} \cdot \frac{1}{10} = \frac{1}{20}$		

נחשב את ♣:

$$\clubsuit = \frac{1}{6} \cdot \left[\frac{99}{100} \cdot \frac{1}{12} + \frac{5}{100} \cdot \frac{1}{20} \right] = 0.014$$

באופן דומה נחשב את השאר:

	1	2	3
F	$\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$	0.014	
L	$\frac{1}{2} \cdot \frac{1}{10} = \frac{1}{20}$	0.024	

נשים לב שאחרי 2 תצפויות מאחר שההטלה השנייה הייתה 6, קיבלנו עלייה בסבירות שמדובר בקוביה המוטה (שכן בה ההסתברות לקבל 6 היא 0.5).

נמשיך עם החישובים:

	1	2	3
F	$\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$	0.014	♠
L	$\frac{1}{2} \cdot \frac{1}{10} = \frac{1}{20}$	0.024	★

$$\spadesuit = \frac{1}{6} \cdot \left[\frac{99}{100} \cdot 0.014 + \frac{5}{100} \cdot 0.024 \right] = 0.0025$$

$$\star = \frac{1}{2} \cdot \left[\frac{1}{100} \cdot 0.014 + \frac{95}{100} \cdot 0.024 \right] = 0.01147$$

ה-likelihood של התצפויות עומדת להיות סכום $\spadesuit + \star = 0.01397$.

מה הסיבוכיות? a כפול גודל ה- a "ב של המצביעים, כלומר לינארי באורך הקלט. ועדיין, יש כאן בעיה פרקטית של דעיכה כי אנחנו בכל צעד מכפילים בהסתברות, והרי פונקציית ההסתברות לא מביאה ערכאים גדולים מ-1. כאשר נזין את המספרים במחשב, נקבל underflow.

איך נטפל בעיה? נשים לב שבנוסחה יש $\sum c_i \cdot b = a$, ובמקרים כאלה נפוץ לשמר log representation של b . נבחין כי $\log \sum_i e^{c_i} = \tilde{a} + \log b = \log(e^{\tilde{a}} \sum_i e^{\tilde{c}_i})$, ועוד $\log b = \tilde{x}$, ונשתמש בתכונה של לוגים ונקבל:

$$\log \sum_i e^{\tilde{c}_i} = \max_i \tilde{c}_i + \log \sum_j e^{\tilde{c}_j - \max_i \tilde{c}_i}$$

את כל המספרים שעולים לגרום לunderflow הוצאנו מחוץ לסדרה.

השאלה השנייה ששאלתנו היא מהי $\mathbb{P}(H_i = s | x_1, \dots, x_n)$. נטען כי עבור $i > n$ כבר פתרנו את זה, כי הסיכוי שבמצב האخرון יהיה s אחריו שראינו את המצב האחרון היה:

$$\mathbb{P}(H_n = s | x_1, \dots, x_n) = \frac{\mathbb{P}(H_n = s, x_1, \dots, x_n)}{\mathbb{P}(x_1, \dots, x_n)} = \frac{F_n[s]}{\sum_t F_n[t]}$$

כלומר הסיכוי להיות ב- s הוא פרופורציונלי ל- F_n אבל צריך לנормל כדי שיסכם ל-1.

אבל, לפעמים נרצה לדעת מה קרה קודם לכן:

$$\mathbb{P}(H_i = s | x_1, \dots, x_n) = \frac{\mathbb{P}(H_i = s, x_1, \dots, x_n)}{\mathbb{P}(x_1, \dots, x_n)}$$

$$\mathbb{P}(H_i = s, x_1, \dots, x_n) = \mathbb{P}(H_i = s, x_1, \dots, x_i) \cdot \mathbb{P}(x_{i+1}, \dots, x_n | H_i = s, x_1, \dots, x_i)$$

$$= \underset{\substack{\text{שרשרת} \\ \text{מרקוב}}}{\mathbb{P}(H_i = s, x_1, \dots, x_i)} \cdot \mathbb{P}(x_{i+1}, \dots, x_n | H_i = s)$$

$$= \underset{\substack{\text{סימוי}}}{F_i[s]} \cdot B_i[s]$$

$$\therefore \mathbb{P}(H_i = s | x_1, \dots, x_n) = F_i[s] \cdot B_i[s] \quad \text{לסיכום:}$$

$$B_i[s] = \mathbb{P}(x_{i+1}, \dots, x_n | H_i = s) = \underset{\substack{\text{סתירות} \\ \text{שלמה}}}{\sum_t} \mathbb{P}(x_{i+1}, \dots, x_n, H_{i+1} = t | H_i = s)$$

$$= \underset{\substack{\text{כלי} \\ \text{השרותת}}}{\sum_t} \mathbb{P}(x_{i+2}, \dots, x_n | H_i = s, H_{i+1} = t, x_{i+1}) \cdot \mathbb{P}(x_{i+1} | H_{i+1} = t, H_i = s) \cdot \mathbb{P}(H_{i+1} = t | H_i = s)$$

נבחן שאות כל החלקים המסומנים באדום אפשר להוריד בזכות התכונות של שרשרת מרקוב. נשארנו עם:

$$B_i[s] = \sum_t B_{i+1}[t] \pi[x_{i+1}, t] \tau[t, s] \quad B_n[s] = 1$$

ולכן אפשר לענות על $\mathbb{P}(H_i = s | x_1, \dots, x_n)$ גם על $\mathbf{Forward}$ ב-(n) O ו- $\mathbf{Backward}$ ב-(n) O , כלומר אפשר לענות על שאלות מהסוג $\mathbb{P}(H_i = s | x_1, \dots, x_n)$ לבלי שנרצה, אבל צריך לשמור את כל המטריצה ולא רק לזכור את סופה. במחיר של $2n$ כפול מספר המצבים נוכל לספק תשובה. **לסיכום**, קיבלנו אלגוריתם Forward-Backward ששתיים בלתי תלויים אחד בשני, אבל שילובם יספק מענה לשאלה שענינה אותנו.

$$\mathbb{P}(H_i = s | x_1, \dots, x_n) = \frac{F_i[s] B_i[s]}{\sum_t F_i[t] B_i[t]}$$

לכל מקום במטריצה המכפלה של העמודה ה- i של F והעמודה ה- i של B תסתכם להיות הנראות.

תזכורת וסיכום השיעור הקודם

$$\mathbb{P}(x_1, \dots, x_n, H_1, \dots, H_n) = \mathbb{P}_0(H_0) \cdot \prod_i \mathbb{P}(H_{i+1} | H_i) \cdot \prod_i \mathbb{P}(x_i | H_i)$$

$$\mathbb{P}(x_1, \dots, x_n) = \sum_s F_n[s] = \sum_s B_1[s] \cdot \mathbb{P}_1(s)$$

$$\mathbb{P}(H_1 = s | x_1, \dots, x_n) \propto F_1[s] \cdot B_1[s]$$

בעיית השחזור – Maximum Probability Reconstruction

בעיה של מציאת ה- h -ים שגורמים לתוצאות שלמו להיות הći סבירות:

$$\arg \max_{h_1, \dots, h_n} \{\mathbb{P}(H_1 = h_1, \dots, H_n = h_n, x_1, \dots, x_n)\}$$

הערה: אם נתנה בתוצאות, בלומר נמקסם את $(x_n | H_1 = h_1, \dots, H_n = h_n | x_1, \dots, x_n)$ נקבל את אותה הבעה, אבל יותר קל לעבד ללא ההנחה.

כמו בעיות רבות בקורס, נפתרו גם את הבעיה זו באמצעות תכנון דינמי. תת' הבעיות:

$$\max_{h_1, \dots, h_n} \{\mathbb{P}(H_1 = h_1, \dots, H_n = h_n, x_1, \dots, x_i)\}$$

בביטוי זהה אפשר לשאול מה המקסימום על המשטנה האחרון:

$$\max_{h_1, \dots, h_{i-1}} \left\{ \max_{h_i} \{\mathbb{P}(H_1 = h_1, \dots, H_{i-1} = h_{i-1}, x_1, \dots, x_{i-1}) \cdot \tau[h_i, h_{i-1}] \cdot \pi[x_i, h_i]\} \right\}$$

$$\mathbb{P}(H_1, \dots, H_i, x_1, \dots, x_i) =$$

$$= \mathbb{P}(H_1, \dots, H_{i-1}, x_1, \dots, x_{i-1}) \cdot \mathbb{P}(H_i | H_1, \dots, H_{i-1}, x_1, \dots, x_n) \cdot \mathbb{P}(x_i, H_1, \dots, H_{i-1}, x_1, \dots, x_{i-1})$$

וכזיא מטור max הפנימי את הביטויים שלא קשורים ל- h_i ונשאר עם:

$$\max_{h_1, \dots, h_{i-1}} \left\{ \mathbb{P}(H_1 = h_1, \dots, H_{i-1} = h_{i-1}, x_1, \dots, x_{i-1}) \cdot \max_{h_i} \{\tau[h_i, h_{i-1}] \cdot \pi[x_i, h_i]\} \right\}$$

נדיר:

$$V_i[s] = \max_{h_1, \dots, h_{i-1}} \{\mathbb{P}(H_1 = h_1, \dots, H_{i-1} = h_{i-1}, H_i = s, x_1, \dots, x_i)\}$$

$$V_i[s] = \max_t \{V_{i-1}[t] \cdot \tau[s, t] \cdot \pi[x_i, s]\}$$

כלומר, כאשר אנחנו שואלים מה שרשרת המצבים הטובה ביותר ביחס למשתנה s עם ערך $s = H_i$, הדריך לחשב על זה היא להסתבל על כל המצבים עד $i - 1$ ולבסוף אחד מהם צריך להתייחס באילו החל ממה ממשיכים ל- s והתצפויות הרלוונטיות בהינתן s .

אלגוריתם זה קוראים **אלגוריתם Viterbi**.

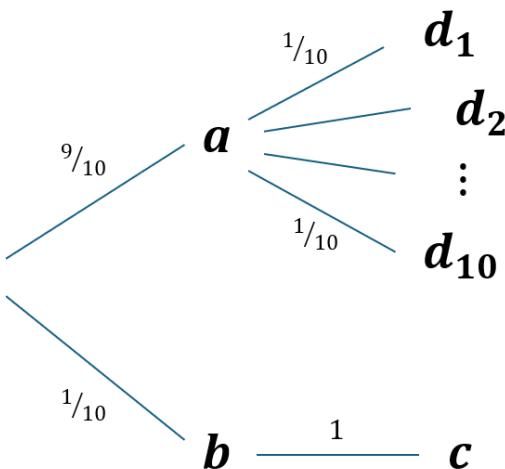
שאלה הבנה

נניח שקיבלנו רצף x_1, \dots, x_n , הרצינו אלגוריתם Viterbi וקיבלנו רצף של המצבים הסבירים ביותר בהינתן התציפות $\hat{h}_1, \dots, \hat{h}_n$.

$$\text{נגדיר } \tilde{h}_i = \arg \max_{h_i} \mathbb{P}(H_i = h_i | x_1, \dots, x_n).$$

מה אפשר להגיד על הערך $(\tilde{h}_i | x_1, \dots, x_n)$ אם \hat{h}_i זה הערך שהאלגוריתם החזיר? יכול להיות שהאלגוריתם במצב i -בחר מצב מאד לא סביר, כי בכל שאר המצבים המצב סביר.

האם $\hat{h}_i = \tilde{h}_i$? אינטואיטיבית היינו רוצים שהוא יהיה המצב, אך זה לא כן. אחת הדרכים היא ליצור דוגמה מנומנת לחולון שעושה את זה. נבנה שרשרת מركוב שבה שרשרת הכיוון סבירה אינה סביר:



כאן הערך הכיוון סביר בזמן 1 הוא במצב a , אבל שרשרת הכיוון סבירה היא $c \rightarrow b \rightarrow a$.

מהחר שאנחנו צריכים להתחייב על כל מצב בשרשראת כאשר מבצעים שימוש באלגוריתם Viterbi, אנחנו צריכים במסלול שעובר דרך a לעבור דרך b במסלול שמוביל בתוכו אי-ודאות, ולכן הסיכוי נמוך. i למשל, יכול להביא רצף כמו $c \rightarrow a \rightarrow b$ שאינו אפשרי בסדרת המצבים שהגדכנו.

כאשר הגדרנו את $[s]_i$ קיבלנו רק את הערך המקסימלי ולא את השרשראת שהובילו אליו. גם כאן אנחנו נתקלים בבעיה דומה לביעית העימוד, ונרצה לשמר במטריצה שאנו ממלאים גם את המסלול, ואז נוכל ללבת אחרת ורשרא.

אנחנו מעוניינים ללמידה H_n , מאייפה לומדים אותו? ללמידה את τ (הסיבוי לעבור בין מצבים) וגם את π (הסיבוי לכל מצב).

נתחילה בבעיה הקלה שנקרה לה Full Data. במקרה זה התחפויות שלנו מורכבות מזוגות של רצפים:

$$1) \begin{bmatrix} h_1 & h_2 & \dots & h_{n_1} \\ x_1 & x_2 & \dots & x_{n_1} \end{bmatrix}$$

$$2) \begin{bmatrix} h_1 & h_2 & \dots & h_{n_2} \\ x_1 & x_2 & \dots & x_{n_2} \end{bmatrix}$$

⋮

$$M) \begin{bmatrix} h_1 & h_2 & \dots & h_{n_m} \\ x_1 & x_2 & \dots & x_{n_m} \end{bmatrix}$$

זה"ב M דוגמאות שכל אחת מהן היא רצף של תחפויות ורצף של מצבים שפלטו אותן. נספר כמה פעמים כל אירוע קורה, נבחן תדירות (maximum likelihood). האם יש מצבים בחיים בהם נוכל להשיג data שנקרה ככה? כן למשל התחפויות הן חלבוניים (רצפים של τ . אמינו), ובתוור hidden state מידע על המבנה השינויי שלהם. במצב כזה, בהינתן רצף חדש אם נריץ עלייו Viterbi הוא יבנה את המבנה השינויי הסביר ביותר. רוב המקרים בחיים אינם כאלה.

ונסה לבחון את הנראות המרבית:

$$\mathbb{P}(\text{Data}) \stackrel{i.i.d}{=} \prod_m \mathbb{P}(h_1[m], \dots, h_{n_m}[m], x_1[m], \dots, x_{n_m}[m]) =$$

כלומר ההסתברות של בל ה-data שיש לנו. נמשיך:

$$= \prod_m \left[\mathbb{P}_0(h_1[m]) \cdot \prod_{i=1}^{n_m-1} \mathbb{P}(H_i = h_i[m] | H_{i-1} = h_{i-1}[m]) \cdot \prod_{i=1}^{n_m} \mathbb{P}(x_i[m] | H_i = h_i[m]) \right] =$$

$$= \prod_m \left[\mathbb{P}_0(h_1[m]) \cdot \prod_{i=1}^{n_m-1} \tau[h_i[m], h_{i-1}[m]] \cdot \prod_{i=1}^{n_m} \pi[x_i[m], h_i[m]] \right]$$

마חר שיש לנו רק מכפלות ללא סכומים, הכל קומוטטיבי ואפשר לסדר באיזו דרך שנרצה:

$$= \left[\prod_m \mathbb{P}_0[h_1[m]] \right] \cdot [\Pi \tau[\dots]] \cdot [\Pi \pi[\dots]]$$

יש לנו M תצפויות. אפשר לחלק אותן לפי כמה פעמים מופיע כל אחד מהמצבים. את החלק **המסומן** אפשר לכתוב באופן הבא:

$$\prod_{s \in S} \mathbb{P}_0(s)^{N_{0,s}}$$

כאשר $\{s = s\} = \sum_m \mathbb{1}\{h_1[m] = s\}$, סטטיסט על הדאטה כמה פעמים התחלתי במצב s .

האם אפשר לעשות טרנספורמציה זהה עם שאר הביטויים?

$$\mathbb{P}(\text{Data}) \stackrel{i.i.d}{=} \left[\prod_{s \in S} \mathbb{P}_0(s)^{N_{0,s}} \right] \cdot \left[\prod_{t \in S} \prod_{s \in S} \tau[s, t]^{N_{t,s}} \right] \cdot \left[\prod_{s \in S} \prod_{x \in O} \pi[x, s]^{N_{s,x}} \right]$$

כאשר $\{s = s\} = N_{t,s} = \sum_m \sum_i \mathbb{1}\{h_{i-1}[m] = t, h_i[m] = s\}$ (מספר הפעמים שראיתי מעבר מ- t ל- s), ונגיד $\{x = x\} = N_{s,x} = \sum_m \sum_i \mathbb{1}\{h_i[m] = s, x_i[m] = x\}$

עבדיו נרצה לעשות מקסימום לביטוי. הדרך הנוחה תהיה להשתמש בבעיה שכבר פתרנו – אין אילוץ שמקשר בין \mathbb{P} לבין τ למשל, 3 בעיות שונות (מסומנים לעיל ב-3 סוגרים [] שונים). משתמש בתוכנה הבאה:

$$\max_x \max_y f(x)g(y) \stackrel{\text{non negative}}{=} \left[\max_x f(x) \right] \left[\max_y g(y) \right]$$

הבעיה **הכתומה** זהה למידת קובייה:

$$\text{MLE: } \hat{\mathbb{P}}_0 = \frac{N_{0,s}}{\sum_t N_{0,t}}$$

הבעיה **הירוקה**: אנחנו יודעים על τ שבහינתן t סכום כל ה- s בעמודה של ה- t צריכים להסכם ל-1. בעצם זו לא קובייה אחת, אלא הרבה קוביות שיש קשר ביניהן:

$$\text{MLE: } \hat{\tau}[s, t] = \frac{N_{t,s}}{\sum_u N_{t,u}}$$

באופן דומה הבעיה **הסגולה**:

$$\text{MLE: } \hat{\pi}[s, x] = \frac{N_{s,x}}{\sum_y N_{s,y}}$$

סיכום:

MLE: $\hat{\mathbb{P}}_0 = \frac{N_{0,s}}{\sum_t N_{0,t}}$;	$\hat{\tau}[s, t] = \frac{N_{t,s}}{\sum_u N_{t,u}}$;	$\hat{\pi}[s, x] = \frac{N_{s,x}}{\sum_y N_{s,y}}$
--	---	---	---	--

לכן כדי למשוך את זה בפועל, צריך לאסוף את הסטטיסטיים לכל s לכל t , ולכל x, s מהדата, ואז השמה לתוך המשוואות שסומנו ברייבוע.

הבעיה הקשה

במקרה הזה ה-*hidden state* לא גלווי לנו. כלומר הדטה הוא מהצורה:

$$1) \left[x_1[1], \dots, x_{n_1}[1] \right]$$

$$2) \left[x_1[2], \dots, x_{n_2}[2] \right]$$

:

גם כאן נתחיל מהנראות:

$$\mathbb{P}(\text{Data}) \stackrel{i.i.d}{=} \mathbb{P}(x_1[m], \dots, x_{n_m}[m])$$

הבעיה בצעד הבא היא שאם אנחנו רוצים להגיע לפרמטרים שלנו (במנוחים של π - α) נצטרך לפרק ביטוי שבתוכו יש סכומים – העARBוב שלהם קשה יותר, ולא מזכה לקומוטטיביות שהייתה במקרה הקל. אם נשילר את הבעיה לבעיית Machine Learning, אפשר לראות קווי דמיון בעיות supervised learning בהן שיכנו דטה לתוך clusters מבלי שנותנות לנו תוויות על הדטה.

הפתרון שנראה לבעיה זו מזכיר מאוד את אלגוריתם K-Means, חלוקת כל הנקודות לקבוצות וחישוב המרכז, העברת נקודות בין הקבוצות עד אשר לא חל שינוי במרכזי הקבוצות. נחלק את הדטה שלנו לקבוצות שונות של hidden values, ועוד נשתמש בפרמטרים כדי לבצע ניחוש מושכל יותר של החלוקה.

האלגוריתם יחקק בין הפרמטרים π , לבין ה-*Hidden Values*. אם נדע את הערכים הנסתורים נוכל לחשב את הפרמטרים, ואם נדע את הפרמטרים אולי נוכל לשפר את הניחוש של הערכים הנסתורים.

שבוע 5 – הרצאה 9

תזכורת

אנחנו נמצאים במצב שבו הדאטה שלנו נראה בצורה זו:

$$\begin{aligned} 1) & \left[x_1[1], \dots, x_{n_1}[1] \right] \\ 2) & \left[x_1[2], \dots, x_{n_2}[2] \right] \\ & \vdots \\ M) & \left[x_M[M], \dots, x_{n_M}[M] \right] \end{aligned}$$

סט הפרמטרים שאנו ממעוניינים ללמידה הוא:

$$\Theta = \langle \mathbb{P}_0, \tau, \pi \rangle$$

הסטרטגיה שדברנו עליה היא שם נוכל לשער פרמטרים (Θ) אז נוכל לשער גם סטטיסטים (N), וגם ההפנ. נתחל עם ניחוש התחלתי של פרמטרים. הניחוש ההתחלתי משפיע על התוצאה הסופית, אבל נשאיר את הדיון על בר המשך.

סקיצה של האלגוריתם:

```
 $\Theta^1 \leftarrow \text{Guess}$ 
For j = 1, ...
 $N^j = ???$ 
 $\Theta^{j+1} \leftarrow \text{MLE} [N^j]$ 
```

נותר לדבר על אורך מתחלים, על מתי עוצרים ועל אורך קובעים את N . את E MLE בהינתן סטטיסטים רואינו בשיעור שעבר.

יש שני וריאנטים של האלגוריתם זהה: הראשון נקרא Max-Max (נקרא בר ב' עושים מקסימום ל-MLP ומקסימום לסטטיסטים). השני נקרא (Exp-Max) Expectation-Maximization.

(MM) Max-Max

מה יהיה Max Step באלגוריתם?

$$\vec{H}^j \leftarrow \arg \max_H \mathbb{P}_{\Theta^j} (\vec{H}^j | \text{Data})$$

$$\text{וכן: } N^j \leftarrow \text{count on } \vec{H}^j$$

כלומר בהינתן ה-Data קובעים פרמטרים, ורואינו אלגוריתם שモצא את המצביעים החבויים היכי סבירים – ויתרובי.

כל מואוד להוכיח שהאלגוריתם משפר בכל צעד, כי כאשר מעדכנים את \vec{H}^j אנחנו מוצאים את H שמנמקסם את הנראות, ובצד של ה-likelihood בחרנו פרמטרים שמנמקסים את הנראות. בסופו של דבר, ממקסם את הפונקציה $\mathbb{P}(H, X, \Theta)$.

אלגוריתם Exp-Max

אנחנו נתמקד באלגוריתם זהה. נגיד: $\vec{N}^j = \mathbb{E}[\vec{N} | \vec{X}, \Theta]$. מכאן השם expectation, כי אנחנו לא מחשבים את הספירות על הדאטה החבוי סביר, אלא למצע על כל ה-Hidden States. מתמטית, חושבים על זה בסכום הבא: $\sum_{\vec{H}} \mathbb{P}(H | \vec{X}, \Theta)$, הרבה פחות נוח לעובד עם הביטוי הזה מבחינה קומבינטורית – לוקחים את כל ההשומות האפשרות ל- H ומשקללים אותן. נשים לב שלאורך הדרך אנחנו עובדים עם מספרים שלמים, אבל ברגע שנשקלל תוחלת נקבל מספרים רציפים, ולכן נוכל לדבר על התוצאות גם בצורה שברית (לעומת Max-Max שבו מטילים על אוסף התוצאות הדיסקרטיות שאפשר לקבל).

чисוב עבר $N_{0,x}$

$$\begin{aligned} \mathbb{E}[N_{0,x} | \vec{X}, \vec{\Theta}] &= \sum_m \mathbb{E}\left[1\{H_1[m] = s\} | \vec{X}, \vec{\Theta}\right] = \sum_m \mathbb{P}(H_1[m] \\ &= s | \vec{X}, \vec{\Theta}) \stackrel{i.i.d}{=} \sum_m \mathbb{P}(H_1[m] = s | \vec{X}[M], \vec{\Theta}) = \end{aligned}$$

אם נתעלם לרגע מהאינדקס, למדנו בשיעור על Forward Backward איך לחשב את ההסתברות ש- H במשהו שווה לערך מסוים בהינתן כל התוצאות בסדרה, ולכן נקבל:

$$\mathbb{E}[N_{0,x} | \vec{X}, \vec{\Theta}] = \sum_m \frac{F_1^M[s] \cdot B_1^M[s]}{\mathbb{P}(\vec{X}[m] | \Theta)}$$

чисוב עבר $N_{s,x}$

$$\begin{aligned} \mathbb{E}[N_{s,x} | \vec{X}, \vec{\Theta}] &= \sum_m \sum_i \mathbb{E}\left[\{x_i[m] = x, H_i[m] = s\} | \vec{X}, \vec{\Theta}\right] \\ &= \sum_m \sum_i \mathbb{P}(x_i[m] = x, H_i[m] = s) \end{aligned}$$

גם כאן מעוניין אותנו רק האירוע $-m$ -י, וגם שאנחנו צופים את האיקסים, כלומר כאשר לא מתקיים המקרה $x_i[m] = x$, ההסתברות היא 0. נקבל:

$$\sum_m \sum_{i \text{ s.t. } x_i[m]=x} \mathbb{P}\left(H_i[m] = s \mid \vec{X}[m], \vec{\Theta}\right) = \sum_m \sum_{i \text{ s.t. } x_i[m]=x} \frac{F_i^m[s] \cdot B_i^m[s]}{\mathbb{P}\left(\vec{X}[m]\right)}$$

חישוב עבור $N_{s,t}$

$$\begin{aligned} \mathbb{E}\left[N_{s,t} \mid \vec{X}, \vec{\Theta}\right] &= \sum_m \sum_i^{n_m-1} \mathbb{E}\left(1\{H_i[m] = s, H_{i+1}[m] = t\} \mid \vec{X}, \vec{\Theta}\right) \\ &= \sum_m \sum_i^{n_m-1} \mathbb{P}\left(H_i[m] = s, H_{i+1}[m] = t \mid \vec{X}[m], \vec{\Theta}\right) \end{aligned}$$

חזרה לסטטיסטים

בשידרנו על סטטיסטים הייתה שאלת מה הערך של:

$$\mathbb{P}\left(H_i = s, H_{i+1} = t, \vec{X}\right) \stackrel{\text{כלי}}{=} \text{השערות}$$

- $\mathbb{P}(x_1, \dots, x_i, H_i = s)$
- $\mathbb{P}(x_{i+2}, \dots, x_n \mid H_{i+1} = t, \textcolor{red}{x}_1, \dots, \textcolor{red}{x}_i, H_i = s)$
- $\mathbb{P}(H_{i+1} = t \mid H_i = s, \textcolor{red}{x}_1, \dots, \textcolor{red}{x}_i)$
- $\mathbb{P}(x_{i+1} \mid H_{i+1} = t, H_i = s, \textcolor{red}{x}_1, \dots, \textcolor{red}{x}_i, x_{i+2}, \dots, \textcolor{red}{x}_n)$

מכל מה **שמסום באדום** אפשר להיפטר הודות לתכונות השרשרת. נקבל באמצעות שימוש בנסיבות:

$$\mathbb{P}\left(H_i = s, H_{i+1} = t, \vec{X}\right) = F_i[s] \cdot B_{i+1}[t] \cdot \tau[t, s] \cdot \pi[x_{i+1}, t]$$

הביטויים מופיעים לפי הסדר.

פירוט האיטרציה

ובשיו אפשר לעדכן האלגוריתם. מה עושים בכל איטרציה? עוברים על כל הסדרות, מחשבים-Forward ואות בטהוטיסטים. דרך אחת לעשות את זה היא:

```

 $\theta^1 \leftarrow \text{Guess}$ 
For j = 1, ...
    Run Forward/Backward for all m
    Collect  $E[N]$  with F/B
     $\theta^{j+1} \leftarrow \text{MLE}\left(\vec{N}^j\right)$ 

```

מבחןת סיבוכיות זמן איטרציה אחת עשוה B/F על כל רצף באינפוט (זמן לינארי). גרסה מעט עיליה יותר משלבת את שני השלבים הראשונים באיטרציה ומבצעת במקביל (אין צורך לשמר B/F של כל ה-m-ים אם אוספים את הטוטיסטים בעבר). גם מבחינת מקום, נקבל סאב-לינארי בגודל האינפוט.

בתוכר לוואי אפשר לחשב גם את $P(\vec{X}, \theta^j)$ ב-scale לוגריתמי – נדבר בהמשך על החשיבות של כך.

משפט

1. $P(\vec{X}, \theta^{j+1}) \geq P(\vec{X}, \theta^j)$
2. אם $\theta^{j+1} = \theta^j$ אז θ^j הוא מקסימום מקומי (ליתר דיוק, מקום בו הנגדת מתאפסת).

כלומר כל איטרציה כזו משפרת פונקציית נראות של התצפיות, ואם הוכנסנו אז הגענו למקסימום מקומי

$$\text{של הפונקציה } P(\vec{\theta}) = P(\vec{X}, \vec{\theta}).$$

תנאי העזירה

לפי המשפט שראינו $P(\theta^{j+1}) \geq P(\theta^j)$ ולכן $\ell(\theta^{j+1}) \geq \ell(\theta^j)$ (שכן הגדרנו את ℓ להיות $\log P(\vec{X}, \vec{\theta})$). כלומר אפשר לעקוב אחרי ההתקדמות, ואם אנחנו רואים שמתחלת דעיכה בнерאה שנתבננס. אנחנו יודעים:

$$\ell(\theta^j) \xrightarrow{j \rightarrow \infty} \ell(\theta^*)$$

כאשר θ^* הוא מקסימום מקומי, אבל הקצב הוא חשוב, ושיטה מקובלת לקביעת עזירה של אלגוריתם היא עד כמה מהר מתכנסים.

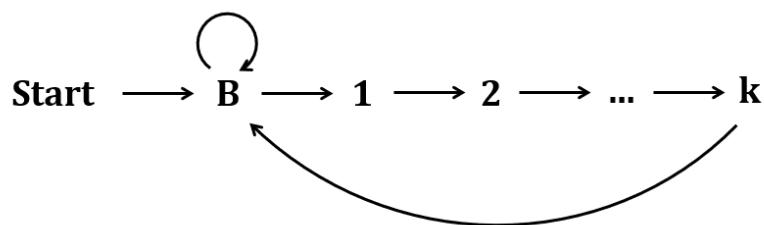
מבט מלמעלה

ראשית, אנחנו שאנחנו יודעים כמה מצבים חביבים יש לנו במערכת. נניח ברגע שיש לנו k מצבים חביבים. אנחנו צריכים לבחור אתחול אקריאי, אבל קיים חשש שנתבננס לאזור בעייתי. איך בוחרים נקודת התחלה טובה?

דבר על מושג שנקרא להגדר "structure". דוגמה – אם אנחנו רוצים להסתכל על רצף של DNA עם מחזוריות שנובעת מאייזו תכונה, כמו למשל שיש דפוסי מעבר מסוים לנצח. באשר נבצע EM, מה יקרה בכלל המעברים הלא אפשריים (אלו שההסתברות שיקרו היא 0)? אנחנו עושים תוחלת על כל ה-hidden states בהינתן פרמטרים. כאשר התחלנו עם מצב שבו חלק מהמעברים הם בהסתברות 0, כל סדרת מצבים שיש בה מעבר עם הסתברות 0 תקבל גם היא הסתברות 0. לכן, אם התחלנו עם $[s^1 \dots s^j] = 0$ אז גם $[t^1 \dots t^j] = 0$ לכל האיטרציות. בר גם עם מטריצת הפלטיהם.

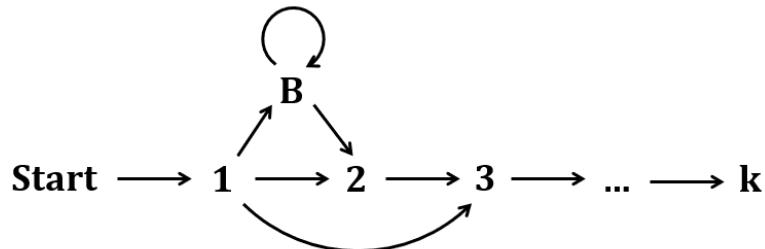
מודלים לדוגמה

1. מילה באורך k



אילו רצפים של hidden states הם אפשריים? מהסוגנו $B, B, B, 1, 2, \dots, k, B, 1, 2, \dots, k$ בתוכו הרצף. בך החבאים **מילה באורך k** בתוך הרצף.

2. מידת רצף חלבוניים כולל שינויים



לומד עמדות חשובות בחלבוניים דומים, ומה הסיבוי ל-insertion או deletion בכל עמדה. בהינתן רצף חלבוניים חדש, נוכל ללמד האם סביר שהגיע מאותו אב קדמון.

שבוע 5 – הרצאה 10

הגנים וכרומטידים

שיעור העשרה – נספר היום על הבiology, ובמקביל על שאלות חישוביות הקשורות בשאנהנו חוקרים את biology.

גינום – רצף של DNA שנמצא בתאים. למשל גנים אנושי, גנים של בקטריה וכדומה.

אם נפשת מבחינת מדעי המחשב, מדובר בקובץ שמכיל רצף של המון אותיות המתאר את ה-DNA. בהרבה מקרים אפשר לחלק אותו לכמה רצפים. בשיעור ההקדמה דיברנו על אחרים שנקרוים גנים. לעומת זאת, ה-RNA שמכיל את הגן הזה (תחילת החץ והנקודה שבסוףו):



מבחן כימית מדובר במולקולה ארוכה, הסימוניים והנטציות האלו הם על סמך תוצאות שראים (החל מנקודה מסוימת מתחיל להיווצר RNA, זה לא כתוב עליו בשום צורה). זה מעלה שאלות, איך התא שבתוכו נמצא ה-DNA יודע להשתמש בקטע הנכון?

תשובה אחת היא שגן מתחיל ברצף ATG ונגמר בקודון STOP, דרך אחת לדאוג שרק דברים שהם גנים יהיו בשימוש זה לדאוג שלא יהיו קומבינציות באלו במקומות אחרים – אבל זו דרישת קשה.

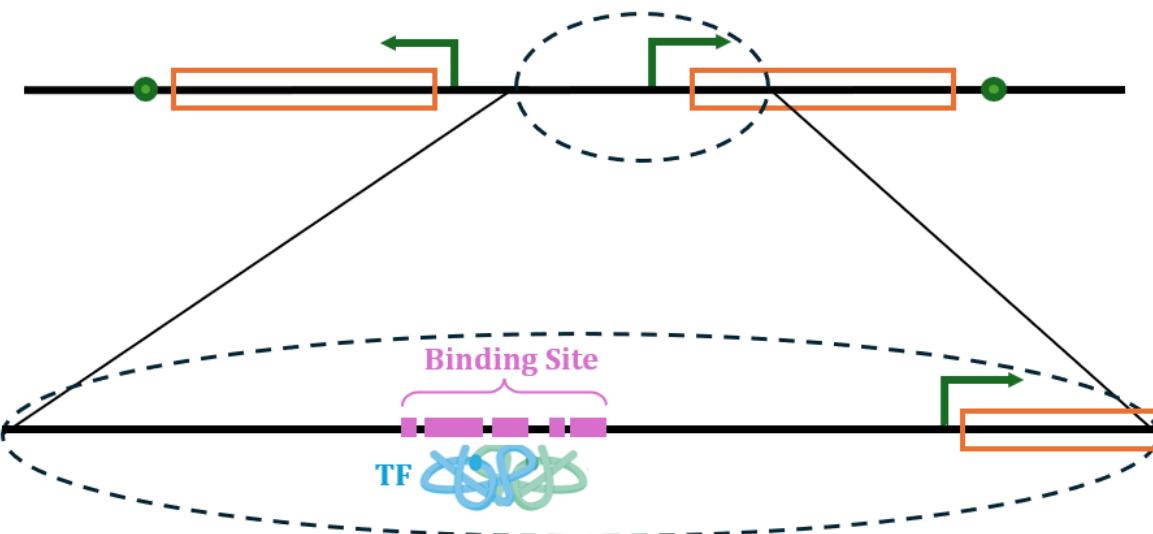
יש עוד אינפורמציה במהלך הרצף שמכונת את התאים להשתמש ב-DNA בצורה ה"נכונה". השאלה היא מהי האינפורמציה זו ואיך ניתן אליה. אם מתחיל בשאלת שnicbt בפni יוצר פשטוט כמו חידך, רוב ה-DNA שלו הם גנים והוא מצליח בהם נכון בעזרת שימוש בעובדה שבאזור שלפני החץ הירוק יש מילה או כמה מילים שמורות שמופיעות בנקודה שמננה אפשר להתחיל לייצר RNA. אותה מילה בדיק רואים באותו רוחים בדיק אותה מילה, והיא מסמנת להתחלה שעתוק.

איך מגלים מילוט מפתח בטקסט?

אם היינו צריכים לקבל קוד למשל של שפת C והיינו צריכים לזהות איפה נמצאת הפונקציה הראשית, אחרי כמה דוגמאות היינו מזהים שפונקציית `main` תמלא את התפקיד הזה.

למה biology משתמש בדברים כאלו? יש משפחות חלבוניים ש מבחינה מבניתם בעלי יכולת להיקשר ל-DNA עם העדפה למילים מסוימות. נניח שיש חלבון שאוהב להיקשר למשל לרצף TATAG. כשהחלבון רואה את הרצף הזה הוא נקשר אליו, ובהעדרו יכולת החצמדות שלו ל-DNA נפגעת והוא לא ישאר שם זמן רב.

לחלבונים כאלו קוראים DNA-binding proteins, אבל יכול להיות גם סתם חלבון זהה – אנחנו נתמקד באחד עם sequence preference, בולם זהה שיש לו העדפה לרצפים מסוימים. נקרא לחלבונים האלו ²Transcription Factors הנובחות של החלבון הזה משפיעה על תהליך השעתוק. לאחר שאליו הוא נקשר אנחנו נקרא binding site.



יש מספר חלבונים כאלו, כל אחד עם העדפות שלו, וهم נוכחים בתא אז הם יקשרו ל-DNA ובמקומות מסוימים יפעילו החלבונים המתאיםים. אם נוכל להבין את הדפוס של איזה חלבון נקשר לאן, היינו יכולים ליצור מיפוי של איזה חלבון ישב איפה. בגנים האדם יש כמה מאות ואולי אלפיים של חלבוני TF (קשרי DNA), רובם עם העדפות קישות, עדין לא כולן ידועים לנו.

זה שחלבון נקשר ל-DNA לא מבטיח לנו שימושו יקרה בעקבות זאת. בחידקים אנחנו מבינים די טוב את התהליך – יש שני חלבונים שכשיהם נקשרים יחד מגיעה מכמה נוספת נספח ונקשרת אליהם, וזה יש את כל הרכיבים הנדרשים להתחלה השעתוק.

איך עובד המנגנון שעוזר/מפריע לשעתוק?

אנחנו רוצים מנגן שיגיד בסיטואציות מסוימות מה לעשות – למשל בהיעדר סובר להתחיל תהליך שידע להזכיר חלבון שיודע ליצור סובר ממשאים אחרים, או עוזר סובר ואז להפסיק את ייצור החלבונים. חלבונים אחרים שנקשרים לרצף יכולים לעזור או להפריע, וזה תהליך פשוט יותר – פשוט ניקשר לאזור הרלוונטי ונפריע להיקשרות של חלבון שאמור לבצע מטרה מסוימת.

²משמעות השם: Factor מאהר שמדובר בישות שמשפיעה על תהליכי, ובשילוב המילה transcription אנחנו מתכוונים לתהליכי השעתוק (DNA הופך ל-RNA).

איך הסיפור הביולוגי שתיארנו מיתרגם לשאלות חישוביות?

נתחילה שאלת פשוטה – יש הרבה ניסויים כדי להבין העדפות קישור של חלבונים. אחת מהשיטות היא Immuno Precipitation (precipitation שמשמעותו להזנת גורמים זרים ולעוזד התקפה עליהם) הידועה לזהות מטרות בעזרת רצף של חלבון. מערכת החיסון הלומדת מנסה רצפים רבים של חלבונים עד שמדוברים אחד שנקשר לגורם הביעיתי. אפשר לגדר חלבונים נגד חלבון מטרה, ואנחנו עונדים להשתמש בתכונת הנוגדן כדי לבקש נוגדן נגד TF.

אפשר לחתך תא שיש בו גנים ומקומות בהם TF מופיע ונקשר לגנים, אם יש לנו נוגדן יש לנו דרך "לשגר" מטען לבדוק מקום שבו החלבון מופיע. התהיליך של Immuno Precipitation הוא לשבר בצורה פיזית את הגנים לפרגמנטים, והנוגדן יקשר רק לפרגמנטים שקשרו אליהם גם TF. אם יש לנו דרך "לדוע" את הנוגדן, נשאר רק עם החתיכות שבahan יש TF.

עבשו אפשר לשאול על הפרגמנטים שבידנו שאלות. למשל, אפשר להתחיל לרצף את הפרגמנטים (לקחת מבנה עם חתיכות DNA ולהפוך אותו לקובץ במחשב של הרצפים שלנו). נתחיל ממילionario תאים ונקבל רצפים אותם אפשר למפות בחזרה לגנים. אפשר ליצור איפה הרצפים שרצפנו יושבים על הגנים. בעיה חישובית שלא נتمكن בה עבשו היא בעית העימוד, איך עושים עימוד של המון רצפים קצריים מול גנים.

בעית Peak Calling

שברנו את h-DNA ומצאנו את הפקטו. אם יש site bindingamiyi נקבל רצפים שמכילים אותו אבל יכול להישבר בכל מיני מקומות סביבו. אחת השאלות שאפשר לחזור היא למצוא את כל המקומות שסביר שקיים בהן אתר (מופיע בהרבה רצפים). אם היינו אוסףים כמה רצפים מכסים כל קטע היינו מקבלים פסגה (הם נערמים) ומכאן השם Peak Calling, כי בשאר המקומות בכל שטחן מהאתר יש פחות רצפים. מדובר בקונבולוציה שמתארת את השבירה. נציין שם יש כמה אתרים זה לצד זה העניין עלול להסתבר.

אחרי שמצאנו את h-peaks בדרך כלל peak caller יסמן לנו אזורים בגנים בהם אנחנו דיבוטחים שיש אתר קשייה. הבiology מתחכמת בנושא זה – לעיתים החלבון יהיה צמוד ל-h-DNA אבל לא בגל שהוא זיהה את h-DNA אלא בגל שיש לו חלבון אחר שזיהה את h-DNA והוא אוהב להיקשר אליו, ונקבל אתרים שמכילים אתר קשייה של חלבון אחר.

בעית Find Enriched Words

נניח שנשאר בעולם הפשט בו חלבון נקשר רק לאתר שלו – שאלת חישובית שקשרורה לבך היא Find Enriched Words. בהינתן אוסף מילים בטקסט שברוב הפסקות מופיעה מילה אחת, לא ידוע איפה, ואנחנו

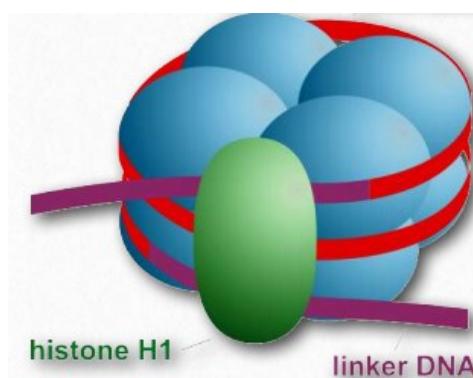
³ Immuno – תלוי במערכת החיסון, Precipitation – השקעה.

נדשים למצוא אותה. כדי שלא נמצא מילים שכיות אנחנו נקבל גם את יתר הגנים שבו המילה צריכה להיות יותר נדירה (בכך נוריד את הסיכוי לקבל מילים שהן פשוט נפוצות כמו "the" באנגלית, שסביר שתהיה נפוצה בכל הטקסט). הבעיה של למצוא מילים כאלה דומה לעביה בעקסט מעלה א"ב, רק שהא"ב יותר מנון מאשר באנגלית, וגם המילים אין שמורות (מדובר בענין של מילים דומות עם יכולת החלפה, אותן השניה למשל חייבות להיות נבונה אבל את השלישית אפשר להחליף וכו'). אפשר לקחת את התוצאות של ניסוי זה ולחזות 어디י קשירה שסביר ומה זיהינו אתרים מסוימים ולא אחרים עד כדיTeVויות מדידה וכדומה.

אפשר לבדוק את הניסוי גם על אותה חיה בתנאי גידול שונים כדי לבדוק האם הפקטור משנה את 어디י הקישור שלו בהתאם למצב.

חידקים הם יצורים שעבורו אופטימיזציות כדי למנוע את מספר הרכיבים הפועלים בהם כדי לקבל מערכת שעובדת היטב ללא חלקים מיותרים. בקיפה האבולוציונית מחידקים לצורים עם גרעין קריה תהילך שאפשר לצורים אלו לבצע פונקציות יותר מורכבות, ועמו הגיע גם בקריה. בצורה מופשטת, כל חתיכה של DNA בחידקים יכולה לעבור שעתוק. ביצורים אוקריוטים (בעלי גרעין) ה-DNA נראה מעט שונה מזה של חידקים, ועטופים בחלבונים בצורה יותר אדוקה מאשר אצל חידקים.

מעבר לשאלה האם החלבן נקשר או לא נקשר, נסיף ממד שמייחס לקישור בירית מחדל של חלבונים שנקראים היסוטנים שיוצרים קומפלקס (מצבור) של חלבונים שבסביבתו מולופfic DNA דמוי דיסקית. אפשר לבצע 2 ליפופים של ה-DNA סביב הדיסקית. הדיסקיות מאוד דיבוקות ל-DNA (נדשת הפעלת אנרגיה כדי להפריד אותו). לקומפלקס כולל (היסוטנים -DNA) קוראים נוקלאוזום.



(התמונה היא תוספת שלי ורק להמחשת המבנה, לא החזרנו מהו DNA)

התכונה הבסיסית של יצורים אוקריוטים זה שיש "מכונות" שכאשר הן נתקלות ב-DNA פנוי הן עוטפות אותו לכדי מבנה של נוקלאוזום. בירית המחדל עברה מבקטריה שבה צריך בצורה מתוכנת להפריע כדי לשעתך, לבירית מחדל שבה כל DNA פנוי יתפס.

השימוש של TF הרבה פעמים הוא רק להיקשר ל-DNA ולהפריע לקישור נוקלאוזמים – בגלל שהם נקשרים שם או אפשר להרכיב שם נוקלאוזם.

לפעמים פקטורי השעטוק מטרתם לסמן אזורים שאחמנם לא רוצים שיהיו עטופים בנוקלאוזם – מוצר משחק של להחביא את הגנים או לא. הרבה פעמים נדבר על האזורים שלא עטופים בנוקלאוזמים בתור החלק הנגיש של הגנים (כל מה שנקשר ל-DNA יכול לעשות זאת). זה שימושי כי אם נוכל לבקר על אילו אזורים חשובים, על אותו רצף גנומי נוכל להציג קומבינציות שונות (במה גנים שונים בתוך רצף אחד).

בשיצורים הפכו לרוב-תאים זה עבר עוד שלב של התמחות (בתאCBD אנחנו רוצים תכונות ספציפיות שלא רלוונטיות לנירון למשל). הדרך לעשות זאת היא להסתיר את רוב הגנים חוץ מהחלקים שרלוונטיים לפעולות התא.

מה קורה עם האזורים הפנויים שלא נחסמו? אפשר ליצור צ'יפ נגד חלבונים שנקשרים באזורים פתוחים. התאים שומרם את זהות התא על ידי כך שמסמנים מה בגנים נשאר נגish.

מה אפשר לבדוק?

נובל לבדוק לאן TF נקשרים בהתאם לשוניים. דבר נוסף הוא לבדוק איפה אין קישורים (ניסוי פופולארי לאחרונה הוא qATAC-Seq שבගדול מפעיל אנדים שחוטך DNA ומסמן אותו – בשיש שני אירושים חיתוך קרובים נצליח לרצף, והאזורים האלה יקרו בדר"כ באזור פתוח). כך מקבלים תמונה של חוסר נוכחות.

אנחנו רק צריכים שהתחלה השעטוק תהיה נגישה – המכונה שעושה שעטוק באוקריוטים יודעת לעבר דרך אזור שארוע על ידי נוקלאוזם.

בגנים אוקריוטי הגנים בדרך כלל חוקיים אחד מהשני (לא קיים לחץ אבולוציוני לצמצם את אורק הגנים לעומת זאת-תאים למשל). הגנים למרות שמקודדים לחלבון קצר יכולים להיות מאוד ארוכים. גם בתוך הגן גם מחוצה לו יכול להיות ארוז, והאזורים הנגישים צריכים להיות אזורים ליד תחילת השעטוק (Transcription start TSS – sites או promoter sites) ולפעמים חוקיים יותר (נקראים enhancers או regulator regions) ובדרך זו משפיעים על הפעולות.

בעיה חישובית

אין הסבר טוב לפטור את בעיית decoding. נניח שיש לנו גישה ל-DB שיש בו את רוב קישורי DNA האנושיים ואילו העדפות קישור יש להם. יש לנו DB שמתאר אילו גנים מתבטאים באילו סוג תאים. אפשר להגיד מה-DB של TF האנושיים רק ה-30 הללו בכלל נמצאים בהתאם, ויש בהחלט שלא מייצרים את החלבון בתא זה לא רלוונטי איפה הוא יכול להיקשר – כי הוא לא ייקשר.

אם אפשר להסביר אילו גנים יתבטאו בתא? אפשר לתאר את האזורים הנגישים ולהגדיר גן א' יבוטא, גן ב' לא יבוטא.. אם נבון את הבעיה זו יהיה צעד אחד קידמה לҚراות להסביר מדוע גנים מסוימים (למשל מחלת)

מתבגרות בשלב מסוים בהתקפות ובדומה. לבאורה יש לנו את כל אבני הבניין לענות על השאלה, אבל ההתקדמות בתחום היא חלקית.

תחנות עגינה

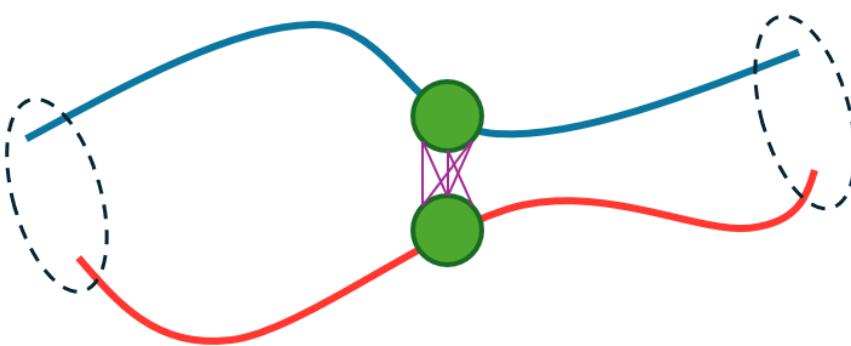
כשבילוג מסמן אזור כ-promotor יש 3 תכניות שיבולות לגרום לכך:

1. בעזרת מסואג בלשחו (שיוקלים שלא נכנס אליהם)
2. נמצא התחלה שעתוק של RNA لكن מתייחל שם שעתוק
3. לקחנו חתיכת h-DNA ושםנו אותה לפני גן מדוח (reporter) שאינו לו פרומוטור ופתאום הופיע ביטוי לגן.

בעזרת גנטיקה (שינו באופן טבעי/מלאכווי את h-DNA) אנשים יכולים שבעת שינוי מספר בסיסים משפייע על גן שמקומו רחוק מאוד. כשהוציאו את הרצף הזה גלו שהגן אכן לא התבטא. מה שנחננו כן יודעים היום זה שיש הרבה אתרים שהם enhancers. מכל מיני ניסויים (צ'יפ, סימון חלבון בעזרת מיקרוסקופיה) אנחנו יודעים שגם enhancementים עובדים על ידי זה שהם מתקבבים לפראמווטור שהם פועלם עלי. זה מעלה שאלות – איך הוא יודע למי להתחבר? מה גורם לו לזהות אותו?

צריך שחלבונים מסוימים יקשרו ל-enhancers ואחרים יקשרו לאזורי Promotor TSS. זה לא עונה על השאלה איך הה-enhancer יודיע לגן להיקשה. התשובה הקצרה – יש שילוב של חלבונים שמשזהים את האתרים עם נטייה מסוימת להתחבר, והגנים ארוז בצורה תלת-ממדית מעניינת, יש אתרי עגינה ספציפיים שэмפוזרים בגנים, אבל בחלקם הם צפופים, ויש חלבונים שקשורים בין אתרי עגינה שונים. הלולאות האלה גורמות למצבי שאטר עגינה שונים רוחקים בגנים אבל קרוביים פיזית, וכן יש מכונות שתופסות DNA ומנסות להעביר אותו דרך מעין טבעת. הטבעת תתקע בנקודות העגינה שתיארנו ועוד או שהטבעת תישאר (ועוד אחת תגע), והטבעות יערמו – נכווץ את כל החוט ביחד, נפוץ בתהילך חלוקת תא) או שהיא פשוט טיפול.

משפחה של שיטות ניסיונות (ש-C-Hi) אחת מהן) לוקחת את הגנים ומפעילות חומר משמר (פורמלדהיד) בתור דבק שגורם לחלבונים להידבק אחד לשני. זה יוצר קשרים בין חלבונים – תהליך שנקרא קיבוע, וזה שוררים את הגנים. כך נקבל קטעים שהיו בעבר חתיכת DNA אחת, רק שעכשיו יצמדו אליהן חתיכות גן שהיו בעבר רחוקות. הקסם הוא לבצע ריאקציה שמחברת בין הקצוות (באזור המקוווקו):



ואז נקבל בעצם את הרצף הבא (שלא היה קיים במקור):



אותו אפשר לרצף. זו עדות לכך שני הרצפים האלה היו קרובים. עבשו אפשר להתחילה למלא מפה של קוואורדינטות בגנים מול קוואורדינטות בגנים ולספור כמה פעמים ראיינו דברים שמתחלים באותו מקום אחד ונגמרים באחרה. מטיב הדברים רוב מה שנקלט יהיה על האלבון של הטבלה, אבל ליד האלבון נקבל אזורים צפופים יותר שמעידים שיש הפרדות פיזיות כמו שתיארנו.

בעיה אחת שuele היא שבדי לא יכולים מטריצה בזאת נדרשים למלא את מספר המקומות בגנים **בריבוע** – צריך לרצף המונן כדי לוודא שאירועים לא קרו באקראי.

מסקנות שידועות היום

1. הרבה מנוקודות העגינה שמורות בכל התאים והרבה מהאזורים של הלולאות חוזרים על עצם בהרבה תאים עם הבדלים קטנים בין תאים שונים.

2. יש *enhancers* שהתפקיד שלהם הוא לא להפעיל את הגן ליד אלא לגייס מוכנות שפותחות את הלולאה או אורחות אותו וקוברות אותו. כך אפשר לבנות חתיכות שלמות בגנים.

אותו *enhancer* יכול לפעול על כמה גנים, אותו גן יש כמה *enhancers* שיכולים לפעול עליו. לאבולוציה כל לעשות דברים כשהמערכת לא מבוקרת (פתאום גנים חדשים יכולים לקבל ביטוי).

סימונים כימיים על החלבון

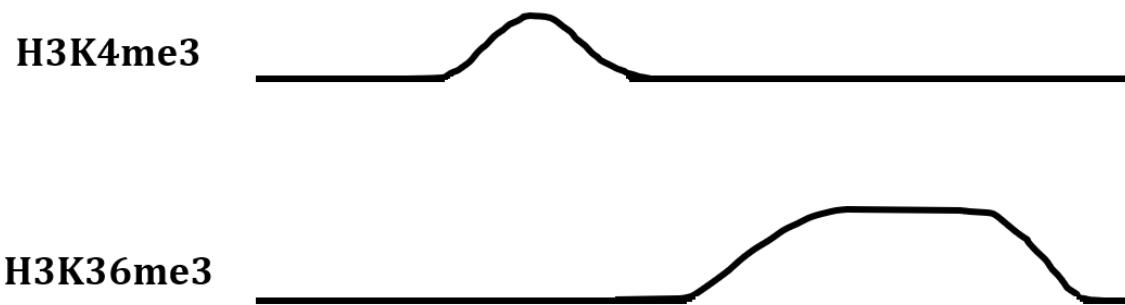
לפני 150 שנה גלו שעל החלבונים שבסביבם ה-DNA ארוז אפשר לסמן סימונים כימיים. הסימונים מאפשרים לסמן חלבונים עם מגוון סימונים כימיים (פחות קרייטי לדין ברגע). מבחינה כימית מדובר בשינויים אחרים התרוגם בצורה של פוספורילציה, מתילציה וכו' שמתאפשרת בעמדות ספציפיות של החלבון.

בשಗילו את זהה, גילו גם אמדים שמסמנים את החלבונים בצורה זו. יש חלבון שההתאים מייצרים שבסייעות מסוימות מבצעים את ההוספה זו. בהמשך גילו שיש חלבונים שנקשרים ל-DNA רק אם הוא עטוף בצורה מסוימת. יש דרך לבתו ויש דרך לקרוא - איז מדובר בזיכרון זמני שמסמן את ה-DNA. זה הביא למאמרי דעה חשובים שיש קוד נוסף ב-DNA - "קוד גנטי נוספת" מעבר לקוד ה-DNA (אפי-גנטי).

מה האינפורמציה השמורה? הדרך שאנשים הגיעו לבעה היא בעזרת מיפוי איפא רואים את הסימנים השונים בגנים בשלבים שונים. איך מmaps? לתפוס את המיקומות בגנים שמחזקים רק את ההיסטון הרלוונטי בעזרת הנוגדן המתאים.

ביצוע מיפוי

היום אנחנו יודעים לבצע מיפויים כאלה בעזרת נוגדים נגד הסימונים. בוגדול, אנשים התחלו לשום לב שכשהם עושים ניסויים כאלה למשל בעזרת H3K4me3 וקיבלו peak במקום מסוים, ואז עם H3K6me3 קיבל אוביזרוציה אחרת.



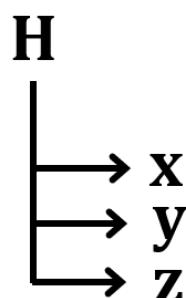
אם יש גן שאנו יודעים שהוא פעיל אז בדרך כלל נראה peak אחריו החץ שמסמן את האזור שבו מתחילה השעתוק, כלומר אפשר להזיהות איפא יש גנים פעילים בחלבון רק על ידי מיפוי הפיקום.

ניסויים הראו שיכל להיות שההנחה שזה קורה רק לפני גנים פעילים לא נכונה, או שיש עוד אלפי גנים שלא ידענו עליהם. התוצאה הוכחה שיש הרבה מקטעים שנראים ומתנהגים כמו גנים, וסומנו בהם אופנים, אבל בפועל לא היו גנים. כך גילו דברים חדשים על הגנים רק ממבט על הסימונים.

מספר שנים אחרי זה יצא פרויקט שמטרתו למצוא את כל המוקומות שימושיים בגנים. אחת הטכנולוגיות הייתה בעזרת צ'יפ, וסימון גנים פעילים ולא פעילים. ביצעו סימונים במספר סוגים תאים וכך חתכו את הגנים במקומות המעניינים. האובייזרוציות של peak מסוג אחד ואחריו peak מסוג אחר מובילים לגן. האם אפשר לבצע מיפוי בצורה יותר אוטומטית?

ChromHMM

מדובר ב-MMM עם ברומטיין - עשו דברים מעט שונים ממה שדיברנו עליו בכיתה. ראיינו שמצב חבוי H מוביל לתצפית x וכן הלאה. התצפית באן במקרה שלו היא רשימה של פרמטרים. באלגוריתם שהם פיתחו הם דיברו על מספר תצפיות:



כלומר כמה תצפיות שונות על אותו זמן (כמו צפיה בסרט עם וידאו וגם אודיו). כל ניסוי צ'יפ זה הוא ערוץ פלט נפרד, והניחו שהוא תלוי תלויה אחרים אם יודעים מה ה-state hidden.

ציר הזמן ובינאריזציה

מה ציר הזמן באן? לגנים אין בדיקת חתיכות לפיהן הוא נשבר. הם החליטו לשבור את הגנים בצורה פשוטה של יצירת חתך באורך 200 בסיסים בכל פעם. אם בתוך ה-200 יש peak מסוים גובה האחת התצפיות אבל בשני לא, היה צריך לבצע ביןאריזציה (להחליט איפה 0 ואיפה 1) - ככלMORE רצף מאוד עשיר לרצף מעט קצר יותר עם 0-1, בערך 15 מיליון מקומות.

בשלב זה היה צריך ללמד את המצבים החבויים והמעבר ביניהם. אפשר ללקח את כל הנ נתונים ולהתעלם מאייהם הם נמצאים ולהסתכל רק על הקואורדינטות - האם בשנייהם נמצא 1 או 0, זה ייתן לנו ניחוש ההתחלתי על המצבים החבויים. בהמשך השתמשו זהה בתוך שנקה לאלגוריתם EM שלמדו בкласс.

Hidden States ו השימוש למחקר ביולוגיה

במודל היה משחק כמה hidden states להבנис וכדומה, אבל בגדול למדנו שכש-MMM רץ על הגנים הוא עושה אנטזיה ואומר באיזה מצב אנחנו בכל מקום בגנים. השאלה היא אם אפשר לחבר את זה לבiology? או שמדובר ברעש?

אפשר להסתכל על המיקומות שבהם אנחנו חושבים שאנחנו יודעים מה קורה, למשל התחלות של הגנים שאנחנו כן מכירים. באופן מפתיע (או לא) ה-hidden state שלהם למדדו היה התחלת גן פעיל והתחלה של גן פעיל נמוך או מוקן לפועלות, גם התחלת גן מושתק - בולם מלמעלה מ-10 מצבים חבויים הגיעו לכ-4. טבלת המעברים הייתה שם נלך משמאלי לימין אם היינו בהתחלה גן סביר שנכנס לגוף הגוף (או מעבר הפור). הבעיה הייתה שכאשר הוא היה באזורי התחלת גן המודול לא זכר האם הוא כבר ראה גן (מאייזה ביון הגע). כשלולים לסוגי תאים אחרים וMRI'ים את אותו מודול, מקבלים אנטציות אחרות (כיו' דברים שונים מתבטאים בהתאם שונים). לוגיקה של מתי גנים מתבטאים שונה בין תא לתא, אבל שיטת הסימון זהה בין תאים שונים. ChromHMM המטריה הייתה לתת ארכיטקטורה של HMM בעדרת שימוש אמיתי בעולם. היום השימוש של HMM משמש גם אזורי בקרה בגנים ואילו אזורי בקרה פעילים. יש סימונים אחרים שביצעו למשל ל-enhancers, לאזורים שעוברים השתקה, אזורי חיבור-L-TF וכדומה. הכל נעשה בעדרת HMM.

למה התוכנה המרקבית מס'ית בآن? לא HMM המסוג היה מצליח לא רע, אבל האינפורמציה שהוא מקבל מ-HMM זה הסתכילות רחבה יותר על הדאטה וזהו שחלק מהמקומות יכולים להיות רועש, אבל בהינתן אינפורמציה נוספת מהסבירה אנחנו יכולים להסיק שהמדובר עליו אנחנו מסתכלים בנסיבות לגיטימי.

איופה זה נופל?

בעיה אחת היא שלפעמים יש תכונות ממתקעים רוחקים יותר, ל-HMM קשה עם זה כי אם נרצה המונע hidden states כדי להתייחס למצבים רוחקים יותר, אנחנו עלולים להגיע למצב של overfit. אם צריך לקלוט מספר אירועים מה עבר, המצב מתחילה להסתבר (הקבלה לחישוביות - האוטומט לא יכול להיות סופי).