

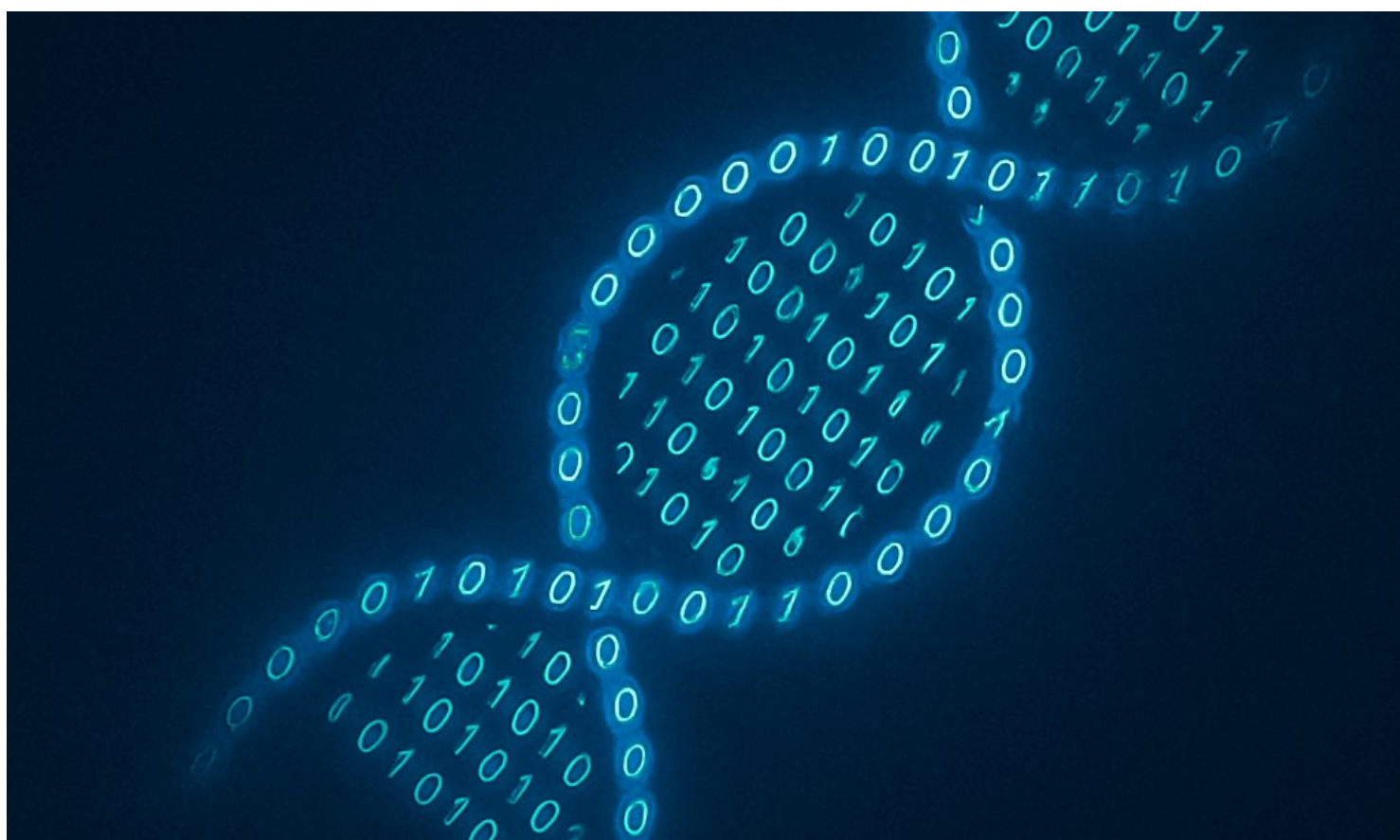
SUMMARIZED BY NOAM KIMHI
noam.kimhi@mail.huji.ac.il

ALGORITHMS IN COMPUTATIONAL BIOLOGY

YEAR 3 – SEMESTER A

COURSE NO. 76558

2025-2026



סיכום הרצאות הקורס

מרצה: ניר פרידמן

| | |
|---|-----------|
| שבוע 1 – הרצאה 1 | 3 |
| הקדמה – ביולוגיה מולקולרית | 3 |
| הדוגמה המרכזית של הביולוגיה המולקולרית – Central Dogma of Molecular Biology | 4 |
| קידוד לחומצות אמינו | 5 |
| פיתוח אלגוריתם לבעיה ביולוגית | 6 |
| שבוע 1 – הרצאה 2 | 7 |
| עימוד רצפים – Sequence Alignment | 7 |
| תכנון דינמי – Dynamic Programming | 8 |
| עימוד במקום לינארי – Linear Space Alignment | 10 |
| עימוד מקומי – Local Alignment | 12 |
| שבוע 2 – הרצאה 3 | 13 |
| Probabilistic Models and Decisions | 13 |
| Two Hypotheses | 13 |
| למת ניימן-פירסון (1933) | 16 |
| One Hypothesis | 17 |
| שבוע 2 – הרצאה 4 | 19 |
| למידה על הסתברות מנתונים | 19 |
| דרך שיערוך בייסיאנית | 20 |
| שימוש בהסתברות בעימוד רצפים | 21 |
| שבוע 3 – הרצאה 5 (הרצאת אורח) | 23 |
| הקדמה | 23 |
| מידול אינטגרטיבי של מערכות ביולוגיות דינמיות | 24 |
| שלבים ביצירת מודל | 25 |
| שימוש במודל | 26 |
| שבוע 3 – הרצאה 6 | 27 |
| היוריסטיקות לעימוד רצפים | 27 |
| שרשראות מרקוב | 28 |
| שבוע 4 – הרצאה 7 | 31 |
| מודלים מרקוביים חבויים – Hidden Markov Model | 31 |
| שבוע 4 – הרצאה 8 | 36 |
| בעיית השחזור – 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 | | הגיבום וברומטידים |
| 53 | | הרצאה 11 |
| 53 | | סימונים כימיים על החלבון |
| 54 | | ChromHMM |

הקדמה – ביולוגיה מולקולרית

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

[מהי ביולוגיה מולקולרית?](#)

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

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

| | | |
|-------------------------|-----------------|---------------------|
| | Population | All the cats in TLV |
| $\approx 10^0 - 10^1 m$ | Organism | Cat |
| | Organs/Tissues | Heart |
| { ביולוגיה מולקולרית | Cells | |
| | Organelles | Mitochondria |
| | Macro-molecules | DNA/Protein |
| | Small molecules | Amino Acid |
| | Atoms | H, O |

מדובר על 9 (!) סדרי גודל בין המולקולה הקטנה לאורגניזם.

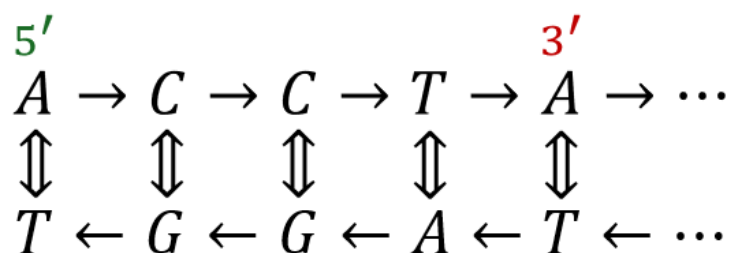
מקרו מולקולות – הרבה מולקולות שמורכבות יחד מסט מוגבל של יחידות, ואוצר די קטן של אבני בניין שבאמצעותן ניתן ליצור מבנים מסובכים.

דוגמאות:

| | | |
|-------------------|---|---------|
| { כאן נתמקד בקורס | נוקלאוטידים (T,G,C,A) | DNA |
| | נוקלאוטידים (U,G,C,A) | RNA |
| | חומצות אמינו (יש 20 קבוצות שונות) | חלבון |
| | מונוסכרידים (לא הוזכר בהרצאה) | סוכרים |
| | גליצרול וחומצות שומן (לא הוזכר בהרצאה) | ליפידים |
| | יוצרים מחיצות בין אזורים נוזליים הודות להידרופוביות | |

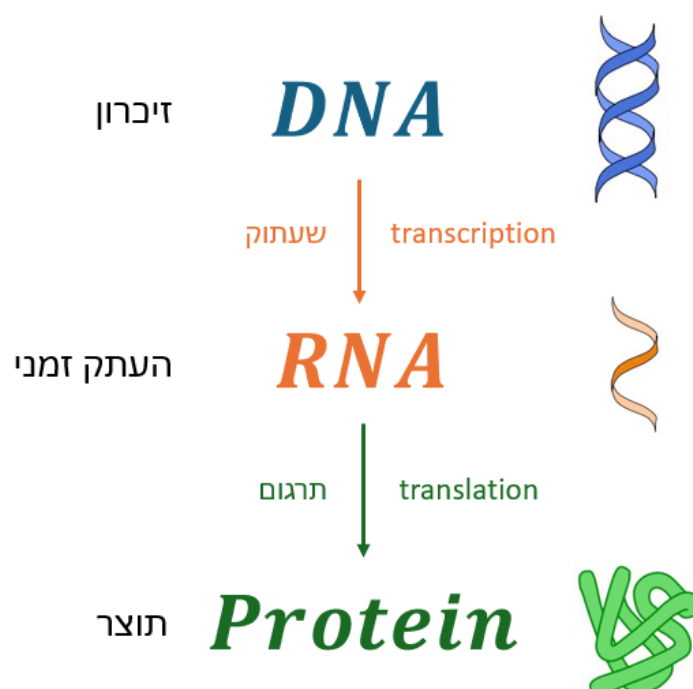
נוקלאוטידים של DNA ושל RNA מעט שונים מבחינה כימית.

DNA זו מולקולה גדולה השומרת מידע. ה-DNA הוא דו-גדילי. הנוקלאוטידים בגדילים שונים מצוותים בזוגות כך: $A \leftrightarrow T, C \leftrightarrow G$. בכל גדיל יש קשר חזק בין הנוקלאוטידים באותו גדיל (לא ייפרד בהרתחה), ובין הגדילים קשר כימי חלש יחסית (ייפרד בהרתחה, אבל לאור המשיכה החזקה – שני הגדילים ימצאו אלו את אלו). נהוג לסמן כך:



מרצף אחד ניתן לשחזר את האחר, כך אפשר לייצר 2 העתקים של הרצף. תכונה זו עונה על שאלה בסיסית של העברת אינפורמציה – תורשה גנטית. ה-DNA לבדו נטול ערך, מדובר במאגר מידע ללא פונקציונליות משל עצמו. ל-DNA אין נטייה להתחבר למולקולות אחרות (ובכך להיהרס). לתמונה נכנס RNA ששונה בהעתקת הבסיסים: $A \leftrightarrow U, C \leftrightarrow G$. ה-RNA יותר פעיל כימית, וממנו מתרגמים את המידע לחלבון בעל פונקציונליות ומטרה.

הדוגמה המרכזית של הביולוגיה המולקולרית – Central Dogma of Molecular Biology

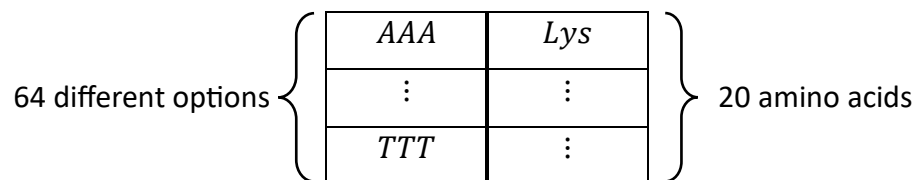


מהתוצר – החלבון, אפשר להרכיב דברים בתא, לייצר מנועים לכיווץ שריר, להעביר סיגנלים, מולקולות וכו'. ייצור החלבון מוכתב לפי הרצף ב-DNA, אשר יעבור גם לדור הבא, והדרך לייצר עוד חלבון היא לבצע קריאה נוספת מה-DNA.

קידוד לחומצות אמינו

מדובר בקידוד משפה של 4 אותיות (A,T,C,G) לשפה של 20 אותיות (חומצות אמינו). מזכיר לנו מעט תרגום בינארי. יש התאמה ברורה בין מילה באורך 4 לספרה בין 0-15 (בסיס הקסדצימלי). כשעוברים לממדים של 4 אותיות אל 20 אותיות, היחסים האלו לא נשמרים – הביולוגיה בזבזנית במובן הזה, כי משפה של 4 אותיות התאמות מדויקות יהיו ברצף באורך 4, 16, 64 – אבל 16 מעט מדי, ו-64 יותר מדי.

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



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

... CA ATG TAA CTG GTT UGA CC ... ATG ...
start
stop
start

הרצף ATG יכול להופיע גם באמצע רצף ולקודד למתינוין, נאמר כי זו חומצת אמינו נדירה יחסית, ואולי זו הסיבה שקיבלה את התפקיד הנוסף.

מסגרת קריאה/reading frame – השלשות שמתרגמים ברצף. הזזה בודדת תוביל לתוצר שונה לחלוטין. רצף DNA יכול להכיל 3 מסגרות קריאה בכל כיוון של הרצף, לכן בסך הכל יש 6 אפשרויות לקריאה.

ריצוף/sequencing – התהליך של פיענוח סדר הנוקלאוטידים בקטע DNA לרצף אותיות. כיום ניתן לרצף גנום של אורגניזמים שלמים, ומדובר בכ- $10^9 \times 3$ אותיות.

פיתוח אלגוריתם לבעיה ביולוגית

לא נתמקד רק באלגוריתם, אלא גם בלמה פיתחנו אותו, ועל מה הוא מנסה לענות. התהליך: (דוגמה בתכלת)

1. **שאלה בביולוגיה** – החלק שאותו אנו מעוניינים לחקור.
מה זה רצף החלבון שמולנו? מה מטרתו?
2. **אינפורמציה רלוונטית** – כשבאים לענות על השאלה יש הרבה כיוונים לחקור.
האם אנחנו מעוניינים לבנות מודל כימי של החלבון? לצורך הדוגמה, נבקש לחקור מהכיוון של דמיון בין רצפים נובע ממקור אבולוציוני דומה, ולכן גם מטרה ופונקציונליות דומות.
 $sequence\ similarity \rightarrow common\ ancestor \rightarrow same\ function$
3. **דאטה** – עבור הרעיון צריך בסיס נתונים, כזה שעבר בדיקות והוא מהימן.
Protein database sequence function
4. **שאלה אלגוריתמית/מתמטית** –
בהינתן הדאטה, מתן מענה לשאלה.
Sequence comparison/query
בשלב הזה גם משיהו שלא מבין בביולוגיה יכול להיות מסוגל לסייע על ידי כך שניצור הגדרה לרעיון.
Similarity is...
- - עד פה הגדרנו רק את מטרת האלגוריתם, ולא מה הוא עושה - -
5. **אלגוריתם** – מציאת האלגוריתם לפתרון הבעיה.
למצוא את המרחק הקטן ביותר..
6. **בחירת פרמטרים** – מדובר בעולם דאטה, האם אפשר להשתמש בה כדי להגדיר את מחיר הטעות?
להשתמש במידע של פונקציונליות חלבונים, נכנס כאן אלמנט למידה. כאן זה מסתבך – אפשר להגדיר רמות קירוב ומה טווח הטעות שנרצה. ההגדרה שיצרנו בשלב 4 צריכה להיות מגובה בפרמטרים: "דמיון בין שני רצפים לפי טבלת 20×20 שבה המידע עד כמה כל אות קרובה לאחרת". זו טבלה שאנחנו נספק ליוצר האלגוריתם.
7. **סטטיסטיקה** – נרצה לגבות את התוצאה שלנו תוך השוואה סטטיסטית לתוצאה מקרית, זה ייתן לנו ודאות על נכונות התוצאה.
מה היה קורה אם לא היה חלבון דומה ב-DB? מה אז הייתה התשובה? נשווה את התוצאה לרצף רנדומלי של אותיות, האם קיבלנו בערך את אותה תוצאה? "Similar by chance".
8. **ויזואליזציה** – איך הגענו למסקנה הסופית? מה תומך בה?
סיפוק של בסיס שיתמוך במסקנות שלנו.

עימוד רצפים – Sequence Alignment

הרבה מאקרו-מולקולות שראינו מתוארות על ידי רצפים של אותיות, והרבה מהביולוגיה 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: $s = \text{AACT}$ $t = \text{AGT}$

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

קודם נתבונן בכמה דוגמאות ונדון בדמיון בין רצפים:

CTAACTG

GAGTG

GACTG

שאלת העימוד – איך ניקח 2 רצפים ונדיז אותם כך שיהיו דומים? לשם כך נגדיר את **חוקי המשחק**:

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

נגדיר א"ב מורחב: $s^*, t^* = \{A, T, C, G, -\}$ ונמצא שני רצפים כך שמתקיים $|s^*| = |t^*|$ וכך שמתקיים $s^* = s, t^* = t$ כאשר הפונקציה $remove - (s^*) = s$ 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 אותה אות.

באפור – רואים indel¹ כי מופיעה אות מול רווח.

בכתום – מופיעה אי-התאמה/mismatch כי אין שוויון באותיות.

הערה: לא נרשה עימוד של – מול – מאחר שעימוד כזה הוא חסר משמעות.

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

| | | | | | | | | | | | | | | | | | | | |
|-------|--|---|---|---|---|--|---|---|---|---|---|--|---|---|---|---|---|---|---|
| s^* | | A | A | C | T | | A | A | - | C | T | | A | A | C | T | - | - | - |
| t^* | | A | - | G | T | | A | - | G | - | T | | - | - | - | - | A | G | T |

¹ indel – הלחם של המילים של **insert** ושל **deletion**, מאחר שעבור הרצף העליון עלינו לבצע $delete(T)$ ועבור הרצף השני עלינו לבצע $insert(T)$.

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

נרצה להגדיר פונקציית ציון σ עבור רמת קירבה לרצף נתון. פונקציית ציון כזו יכולה להיות למשל:

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

עבור הדוגמה שראינו מקודם, נראה עכשיו 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 |
| | | | -1 | | | | -4 | | | | | | | -14 | | |

כעת יש בידינו את הכלים להגדיר בצורה טובה את הבעיה:

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

הערה: תמיד קיים משהו טוב יותר מהעימוד הגרוע ביותר (יש דוגמה בטבלה לעיל, כל התאים indel). כמו כן, יתכן מצב בו יותר מעימוד אחד מקבל את אותו הציון, לכן יכול להיות יותר מעימוד אחד עם ציון מקסימלי.

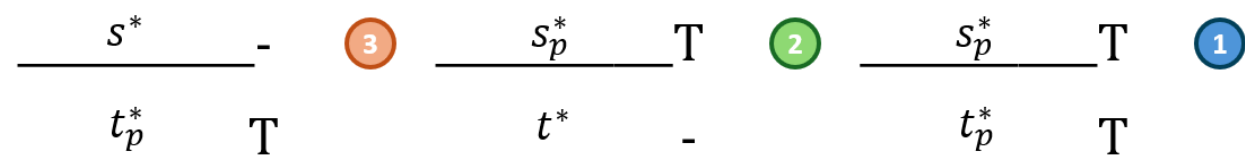
תכנון דינמי – Dynamic Programming

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

בהינתן שני רצפים s, t שמסתיימים באות T נסתכל עליהם כך:

$$\begin{array}{c} * \\ \hline t_{\text{prefix}} \end{array} T \qquad \begin{array}{c} * \\ \hline s_{\text{prefix}} \end{array} T$$

יש למעשה שלוש אפשרויות למקם את הרצפים זה מול זה:



ולכן:

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

נגדיר $|s| = n, |t| = m$ וכן: $V[i:j] = \max(s[1:i], t[1:j])$

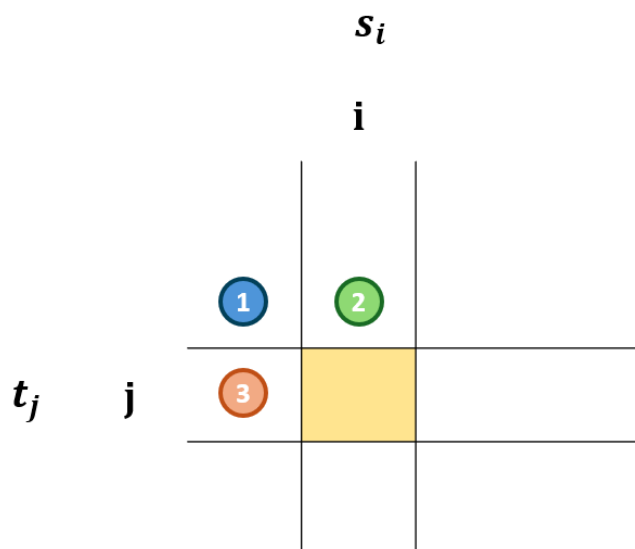
מקרי בסיס: $V[-1, *] = \infty \quad V[*, -1] = \infty \quad V[0, 0] = 0$

הגדרת הרקורסיה:

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

מילוי הטבלה:

באופן כללי אם נתבונן בתא בטבלה, אלו הכללים למילוי:



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

שיטה זו מציעה לנו גם דרך לשחזר את העימוד. נסמן בכל שלב אילו מ-3 האופציות בחרנו כשמילאנו את התא. נגיע למילוי באופן הבא:

| | | | | | |
|----------|---|----------|----|----|----|
| | | <i>t</i> | | | |
| | | A G T | | | |
| | | 0 | 1 | 2 | 3 |
| <i>s</i> | A | 0 | -2 | -4 | -6 |
| | A | -2 | 1 | -1 | -3 |
| | A | -4 | -1 | 0 | -2 |
| | C | -6 | -3 | -2 | -1 |
| | T | -8 | -5 | -4 | -1 |

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

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

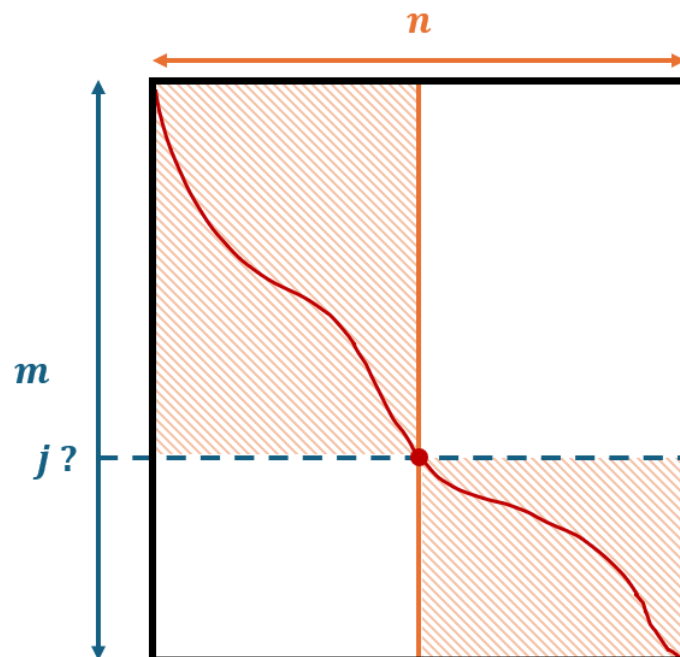
הזמן שקיבלנו הוא פולינומי, אבל אם מבצעים עימוד גנים של אדם מול גנים של עכבר, עדיין מדובר בסדרי גודל עצומים ומתחילה גם בעיית מקום ומוגבלות hardware. כיצד נתמודד? כדי לחשב את הציון בלבד, אין צורך במטריצה כולה, אבל עבור העימוד הסופי עלינו לשמור יותר מידע.

עימוד במקום לינארי – Linear Space Alignment

פתרון נאיבי: יכולנו לנסות בכל שלב להתחשב ב-3 תתי מטריצות עבור 3 האפשרויות. במקרה זמן הזמן יהפוך להיות בשלישית, לא מאוד עוזר.

נחפש פתרון טוב יותר.

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



כך נוכל לפתור כל צד באופן בלתי תלוי. חתכנו את s באמצע, ומשם נפתור שתי בעיות עימוד. העימוד הכי טוב של הרישיות בתוספת העימוד הכי טוב של הסיפות ייתן את העימוד הכולל 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. מחיר מעבר קו המחצית ב- j הוא $V\left[\frac{n}{2}, j\right] + U\left[\frac{n}{2}, j\right]$ ולכן השאלה היא מציאת 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 שמירה של 2 עמודות, והזמן הוא הגודל של הטריטוריה (גודל המטריצה). לכן צעד ה-"Divide" הוא בזמן $O(n \cdot m)$ והמקום הוא $O(n)$. עומק הרקורסיה הוא $\log n$ כי בכל פעם אנחנו חוצים את מספר העמודות ב-2.

באופן נאיבי יש לנו כרגע אלגוריתם שהזמן שלו הוא $O(n \log n)$ אבל זה לא בדיוק המצב, מדוע? כי מדובר במציאת חציון:

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

(שתי המטריצות המקווקוות בכתום באיור למעלה). נבחין שהחלקים שמסומנים במשוואה בירוק מסתכמים יחד ל- m . אם נמשיך את הרקורסיה הלאה, בסיבוב הבא נעבוד על 4 מטריצות שסכומן יחד $\frac{n}{4} \cdot m$, ואז 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 ב- s ותת מחרוזת ב- t שיש ביניהן עימוד טוב? כלומר תת מחרוזת ב- s ותת מחרוזת ב- t שאם נוסיף אליהן עוד אותיות מימין ומשמאל רק יגרעו מהציון.

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 כה היה מוגדר העימוד הכי טוב מהרישה עד לנקודה זו). נגדיר 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. נבחין 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 לדיון 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 אי-ודאות על העולם. איך נמדוד אותה? לפי דגימת תכיפות של אירוע (frequency), לפי תיאור סובייקטיבי של האירוע..

המטרה שלנו בנושא זה היא שנוכל להמשיך לדבר על הדוגמה הרצה מתחילת הקורס: בהינתן שני רצפים נרצה לדעת האם הם מאותו מקור אבולוציוני. אומנם, בשיעור התחלנו להעביר את הנושא באמצעות דוגמה פשוטה יותר: נניח שיש לנו מדידות אובייקט שאפשר לאסוף על הרבה אנשים כגון נתוני בדיקת דם. אפשר לקבל כך שיערוך של איך נראה אדם בריא: $P_H(x)$ כאשר x זו ספירת דם של אובייקט (אלמנט רב-ממדי). בהינתן x נוכל לקבל את ההסתברות ל- x מאוכלוסייה כזו או אחרת.

נדבר על שתי שיטות הכרעה בבחינת היפותזות.

Two Hypotheses

כאן נתעסק במקרה של איך נראים אנשים חולים מול אנשים בריאים, עם 2 התפלגויות:

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

על מנת לקצר, נסמנם ב- $P_H(x)$ וב- $P_F(x)$. כעת נוכל עבור פרט לדבר על הערכים שמתקבלים (התפלגות לפי חולה/בריא). איך נקבע את איכות ההחלטה שקיבלנו על סמך ההתפלגות? באמצעות **כלל החלטה**:

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

נהוג לסדר את סוגי ההחלטות שהתקבלו גם בצורת 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 |

סיוע להתבוננות בטבלה: הצבע של המילה בתוך התאים של הטבלה נגזר מהעמודות. למשל נתבונן בתא השמאלי עליון $True\ Positive$, ה- $Positive$ נגזר מתוך זה שחזינו $Positive$ (לכן ירוק), וה- $True$ נובע מזה שהמצב האמיתי הוא אכן $Positive$ (לכן סגול).

נגדיר $Positive$ כאדם חולה (זה האובייקט אותו אנחנו בודקים). כל חוק הכרעה מתחשב בטעויות שנעשות, אם נקבע שכולם חולים יהיה הרבה FP (טעיתי בכך שקבעתי שהם $Positive$ למחלה, כי הם למעשה $Negative$). לעומת זאת, אם נגיד שכולם בריאים יהיו הרבה FN (טעיתי בכך שקבעתי שהם $Negative$ למחלה, כי הם למעשה $Positive$).

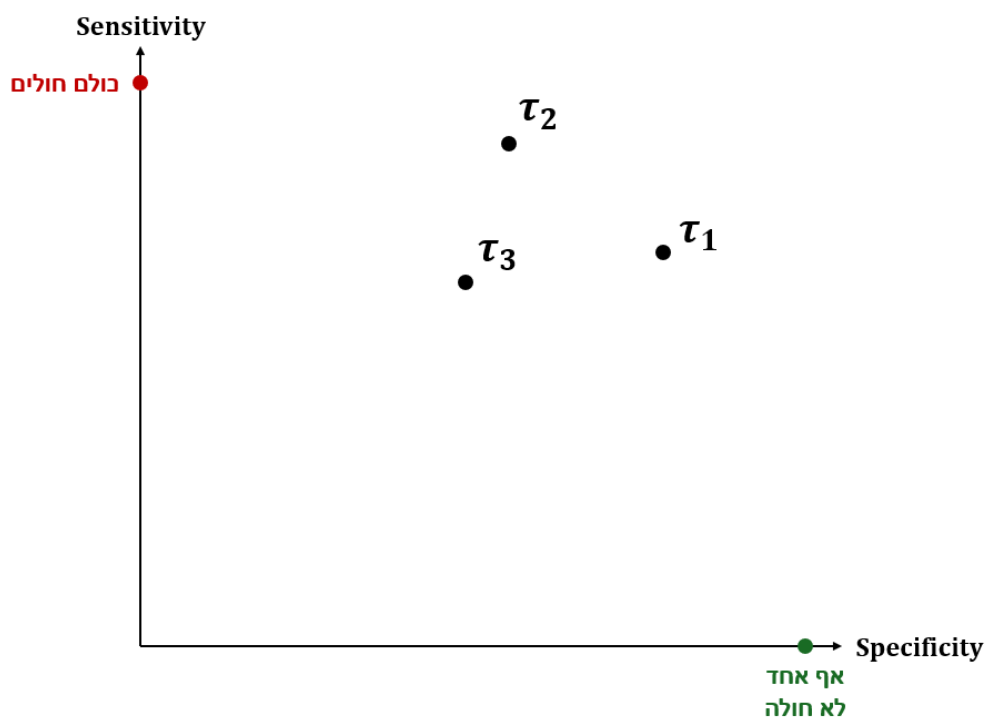
העלינו כמה יחסים מעניינים:

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

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

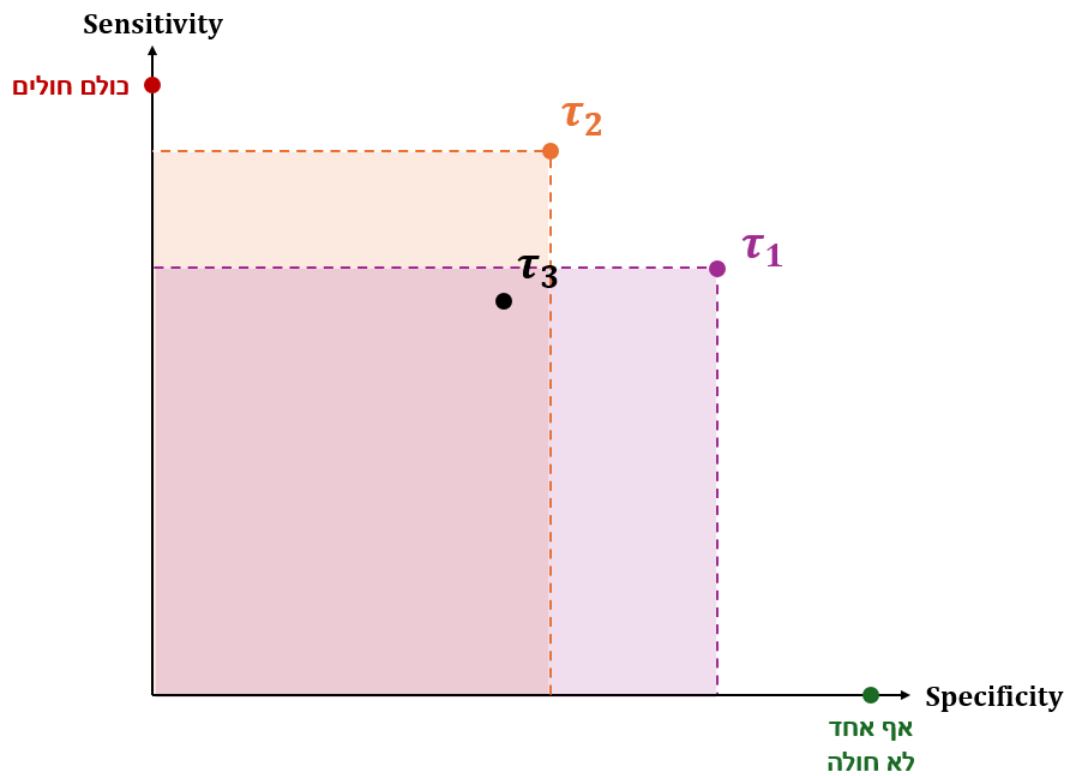
לפירוט המלא: [Sensitivity and specificity - Wikipedia](#)

אפשר להמחיש גם בגרף:

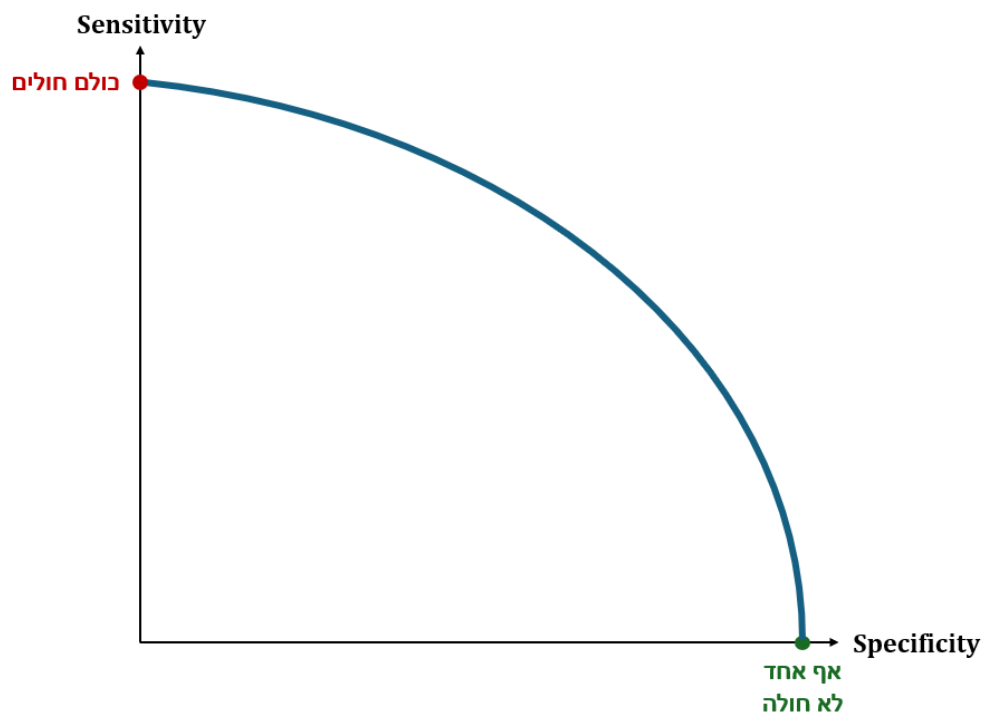


כל חוק החלטה הוא נקודה במרחב הזה (עם ביצועים בשני הממדים). האם נעדיף את τ_1 או את τ_2 ? תלוי במה שנדרש מאיתנו. לעומת זאת, לעולם לא נעדיף את τ_3 על פני אחד מהם, כי הוא גרוע מהם בשני הפרמטרים. בשלב הזה, זה שיקול של המשתמש מה חשוב יותר.

כך למעשה אפשר להציג את זה שכל נקודה במרחב מגדירה אזור (מקווקו) שהיא טובה יותר מכל הנקודות השוכנות בתוכו:



בהנחה שהנתונים שבידינו מאפשרים רציפות נקבל קו קעור. האזור הכלוא מתחתיו ניתן למימוש, אבל כל מה שמעליו בלתי ניתן להשגה:



אם נסתכל על חוק החלטה כלשהו:

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

הזאת t (הסף/threshold) היא משחק בעד כמה מחמירים בדרישה להגיד שמישהו בריא/חולה. הלמה טוענת שמשפחת החוקים הזו מקיימת:

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

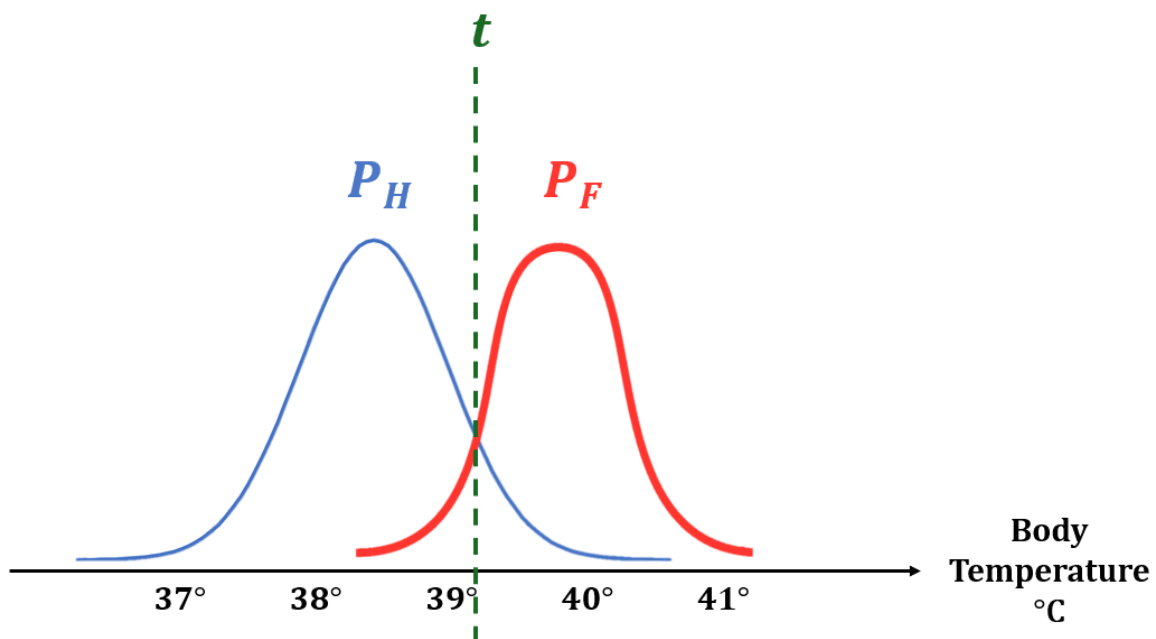
כלומר, החוק מקסימלי אם קיים סף t שעבורו זה מתנהג כמו יחס הנראות. במילים אחרות, אם מציבים את חוקי המשחק (2 הסתברויות) כל החוקים שצריך לשקול בעולם יהיו מהצורה הזו (אם לא, נהיה בתוך האזור שמתחת לקו הכחול שראינו למעלה).

גישה בייזיאנית אומרת שבהינתן תצפית אפשר להכריע איך פרט מתנהג:

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

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

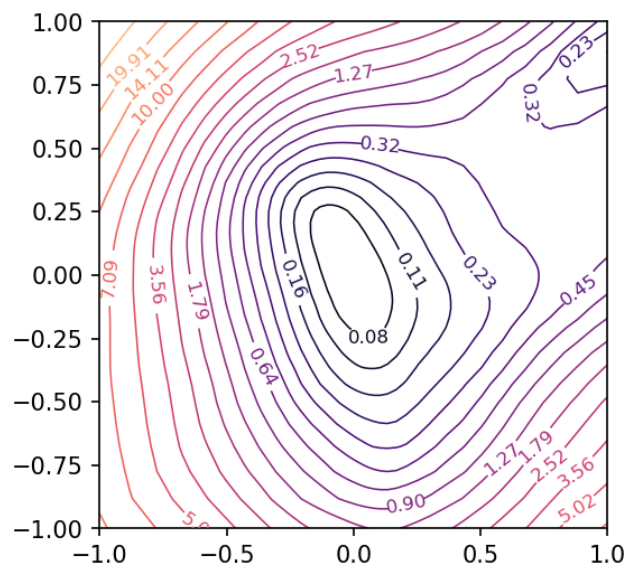
$$\mathbb{P}(x|+) = P_F(x) \quad \mathbb{P}(x|-) = P_H(x) \quad \mathbb{P}(+), \mathbb{P}(-) = \text{Prio (ידע מקדים)}$$



One Hypothesis

נתון לנו $P_H(x)$ ואנחנו רוצים לבחון השערות: H_0 – בריא, H_1 – לא H_0 . המבחנים האלו לא נותנים לנו אלטרנטיבה, אלא רק הכרעה האם מי שמולנו עונה על השערת האפס או לא. במקרה הזה, יהיה לנו רק את P_H בדיאגרמה לעיל.

עכשיו נוכל להחליט לדחות את השערת האפס בהתאם לנתונים, למשל $\overbrace{P_H(temp > 38.5^\circ)}^{P \text{ value}} = 0.0001$. נוח לדבר במונחים של הסתברות המשלים. העניין מסתבך כאשר הבדיקה שלנו היא ביותר מממד אחד (לא רק חום גוף, למשל רמת בולסטרול, לחץ דם..). איך אז ייראה הגרף?



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

חזרה לדוגמת הרצפים

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

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

סימונים:

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

$$p_0, p_1 \text{ תהיה דגימה של אות} \quad -$$

$$s_i \text{ ו-} t_i \text{ האות ה-} i \text{ ברצף.} \quad -$$

הנחות:

- אורך כלשהו, לא נרשה עימוד מסוג indel.
- הרצפים הם $i.i.d$, כלומר כל אות ברצף נדגמה באקראי ובאופן ב"ת. ההסתברות של רצף היא דגימת אותיות מאותו "סל".

לכן תחת ההנחות האלו:

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

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

הערה: המודל הזה כמובן לא מציאותי, אבל הוא יכול להוביל לתוצאות משמעותיות ואז אפשר לחזור להנחות המקלות שלקחנו.

עבור P_1 נניח שמתקיים:

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

הנחת אי-תלות ודגימה באקראי מתקיימות, אבל s_i, t_i בן תלויות ונדגמו יחד. נוכל למשל לקבוע:

| p_1 | A | C | G | T | |
|-------|-----|-----|-----|-----|-----|
| A | | | | | 0.3 |
| C | | | | | 0.2 |
| G | | | | | 0.2 |
| T | | | | | 0.3 |
| | 0.3 | 0.2 | 0.2 | 0.3 | 1 |

| | p_0 |
|---|-------|
| A | 0.3 |
| C | 0.2 |
| G | 0.2 |
| T | 0.3 |
| | 1 |

מה שחשוב להדגיש 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 . כעת נוכל להגיע לפיתוח המרכזי:

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

מזכיר 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

בהינתן דאטה נרצה ללמוד על התפלגות. הדאטה מורכבת מווקטורים של תצפיות. מספר נוציות:

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

המימד ה- n בדגימה ה- i : $\vec{x}_i[n]$

מה זה אומר שהצלחנו ללמוד טוב הסתברות? שאם נקבל עוד דאטה מאותו מקור נצליח עליו בצורה טובה.

[הטלת נעצים](#)

דומה להטלת מטבע, בנטרול הדעה הקדומה שיש לנו על מטבעות.



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

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

הפרמטרים: נגדיר θ באופן הבא:

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

אפשר להגדיר פרמטר גם בצורה 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 . נתחיל בהגדרת פונקציה שמתארת איך הפרמטר מושפע על ידי הדאטה

שלנו. הפונקציה בה נשתמש 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 $\{H, T, T\}$:

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

אנחנו מעוניינים **בנראות המיטבית**, עקרון לפיו כאשר אנחנו משערכים פרמטרים, נבחר את אלו שמביאים

$$\hat{\theta} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \ell(\theta) \quad \text{כלומר:}$$

נסמן $MLE := \text{Maximum Likelihood Estimation}$. נבחין 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 סטטיסט – פונקציה שמתארת את הדאטה.

במקרה הזה, שתיהן יחד 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 . נגזור 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 הראינו את השלבים האלו? מאחר שבמידול של בעיות מורכבות יותר, עוברים דרך שלבים דומים.

דרך שיעור בייסיאנית

ל-maximum likelihood יש יתרונות שונים: תמיד אפשר להפעיל אותם, ה-MLE אינו תלוי בפרמטר (היה אפשר לעבוד גם עם η), הוא אינו מוטה (אם ניקח הרבה datasets – הממוצע עליהם יהיה הנכון) והוא גם מדויק אסימפטוטית. **איפה הבעיה?**

נחזור למקרה הטלת הנעץ ונניח שבידינו דאטה מהצורה $\{T, T, T\}$ השערוך בדרך שראינו יעריך שהסיכוי לקבל H הוא 0, אין פה מקום לביטוי שלנו על ידע קודם. שיטת השיעור הזאת תיתן את אותו הערך גם לדאטה מהצורה $\{T, T, \dots, T\}$ כך 100 פעמים – ברור לנו ששתי קבוצות הדאטה שונות זו מזו.

השיטה הבייסיאנית

בשיטה זו אנחנו לא מתעניינים ב- θ האמיתית של הנעץ, אלא מה אני **חושב** ש- θ הנכון. מתחילים עם מידע קודם על מהו θ ואז מחשבים אותו בהינתן ההנחה הזו.

הדאטה שלנו יהיה מהצורה $x[1], \dots, x[N]$ כאשר $x[i] \in \{A, T, C, G\}$. נגדיר 4 פרמטרים:

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

(היה אפשר להגדיר באמצעות 3 פרמטרים ואחד שהוא המשלים, אבל לצורך הדוגמה נעבוד עם 4)

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

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

חקר הפונקציות מסתבך – כי מדובר עכשיו בפונקציה רב-ממדית. צריך שהנגזרת החלקית בכל כיוון 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 $\max_{\vec{\theta}} f(\vec{\theta}) \quad s.t. g(\vec{\theta}) = 0$ והדרך לזו היא באמצעות כופלי לגראנז':

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

עבור למשל $\hat{\theta}_A$ נקבל: $\hat{\theta}_A = \frac{N_A}{N} \Rightarrow \frac{\partial}{\partial \theta_A} J = \frac{N_A}{\theta_A} - \lambda \Rightarrow \hat{\theta}_A = \frac{N_A}{N}$, וכך גם עבור השאר.

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

בשעושים אירוע מולטינומי, הסטטיסטים עומדים להיות כמה פעמים ראיתי כל תוצאה, ולכן הנראות המקסימלית תהיה באופן הבא:

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

עימוד רצפים

ברצוננו לשערך 2 התפלגויות P_0, P_1 . עבור P_0 :

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

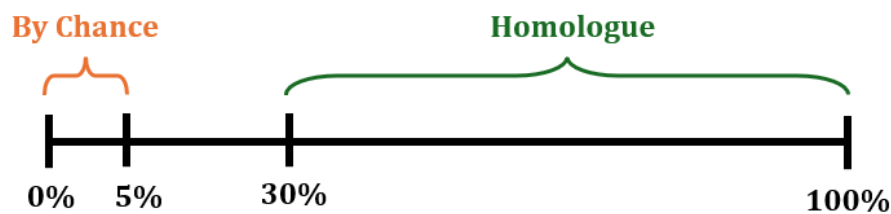
בהינתן 20 רצפים וההנחה שכל אות נדגמה באופן בלתי תלוי. לשערך נראות מקסימלית משמעותה לספור כמה פעמים כל אות הופיעה. אילו רצפים נבניס לסט האימון?

במקרה של $P_1(x, y)$ במקום ללמוד 4 אותיות נלמד 16 זוגות. איפה זה מסתבך? בעיית העימוד נולדה כשהיו רק 20 רצפים של חלבונים ומעט מאוד דאטה לגביהם. היום, כמות הרצפים שקיבלנו מהעולם מסביב היא בקנה מידה עצום.

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

היינו רוצים להיפטר מדברים שהם פשוט הכפלה – זה תחום שנקרא **ביו-אינפורמטיקה**, התחום מתעסק בעבודה על מסדי נתונים וניקוי החזרות מהם כדי שנישאר עם מסד נתונים קרוב ככל הניתן ל-*i. i. d.*

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



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

[עימוד לפי עמדה בחלבון](#)

רוב החלבונים בגוף יכולים להימצא ב-2 מדיומים שונים: מדיום נוזלי (דם, ציטופלזמה), או דרך ממברנות. הסביבה משפיעה על מבנה החלבון ועל סדר הרצפים. מכאן נוצרת גישה נוספת לפיה אפשר ללמוד מטריצה שמתייחסת לחלק ספציפי בחלבון, ועימוד רצפים באמצעות $\sigma(s, t)$ לפי עמדה.

הערה: עד עכשיו התעלמנו מ-indel. אפשר להתייחס לציון לפי כמה פעמים ראינו – מול אות. 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 רוז, מידול אינטגרטיבי של מערכות ביולוגיות

הקדמה

[מושגים בסיסיים](#)

מערכת – אוסף רכיבים הקשורים ביניהם ומתפקדים יחד למען מטרה משותפת.

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

דוגמה קלאסית לדינמיקה היא חוקי ניוטון.

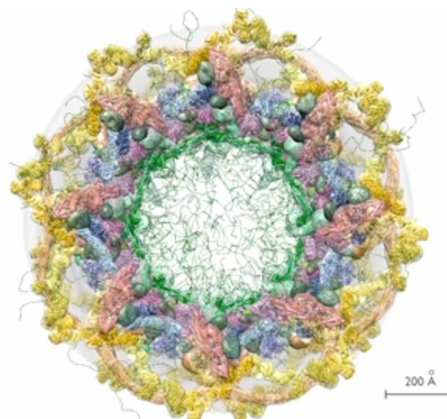
[דיון – מהם חיים?](#)

רעיונות שהועלו בכיתה: כוחות שמתנגדים לסביבה, שימור סביבה פנימית, שכפול ורבייה, מוכוונות הישרדות, אדפטיביות, הנעה עצמית (אולי לוקאלית).

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

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

מערכות ביולוגיות הן מורכבות ודינאמיות בהגדרתן, ולכן נרצה למדל (ליצור "דגם") של מערכת ביולוגית דינאמית. במהלך ההרצאה נעבוד עם דוגמה של מבנה הקיים בגרעין התא ומהווה מסנן בין הגרעין לציטופלסמה – טרנספורט מולקולרי.



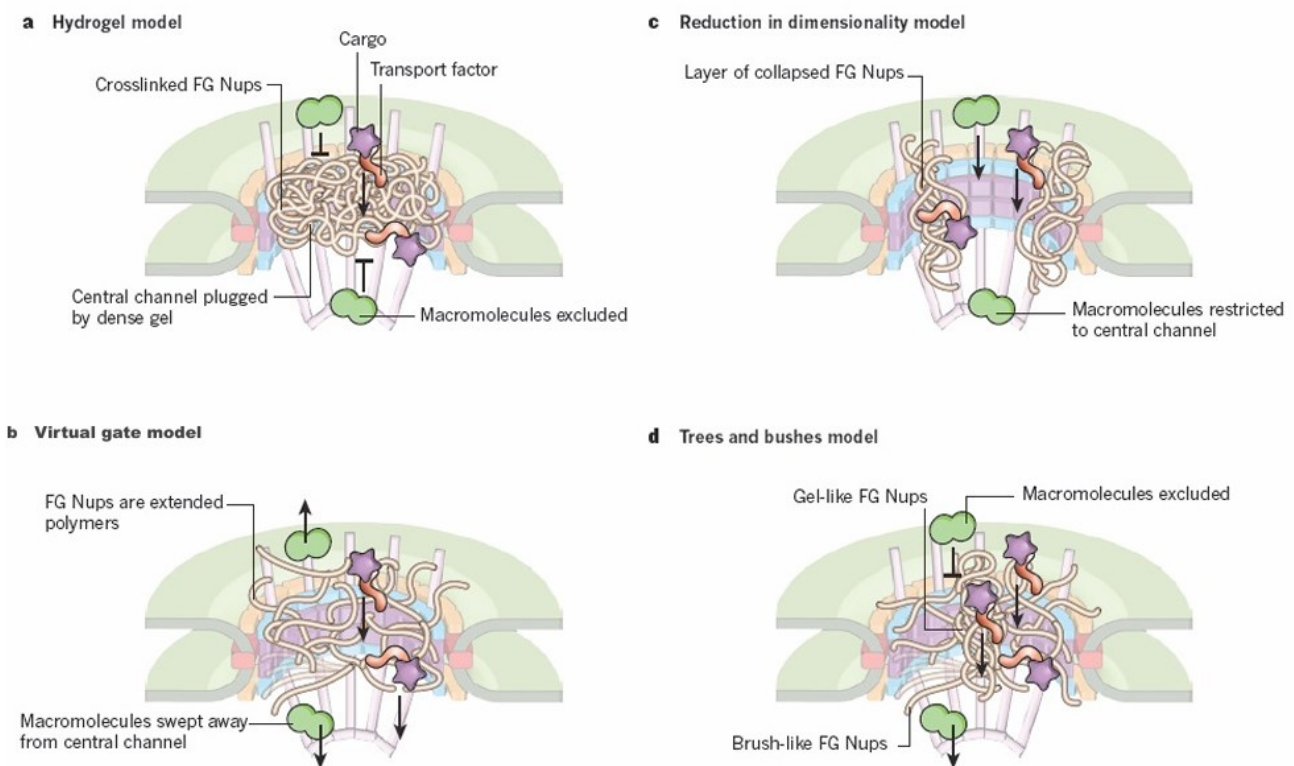
למה צריך מודלים והדמיות? לחזות להסביר תכונות, להכריע בין היפותזות, לשלוט בתכונות המערכת, לזהות פערי מידע ולאתר מידע חסר.

מידול אינטגרטיבי של מערכות ביולוגיות דינמיות

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

נעבוד לאורך ההרצאה עם דוגמה – מודל של תנועה מולקולרית דרך חרירות הגרעין. החרירות מאשרות מעבר מהיר, יעיל וסלקטיבי בין חלקי 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, והועלו היפותזות שונות בנושא:



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

1. איסוף מידע –

בשלב זה לא נטריד את עצמינו באיך נמדל את המידע שנאסף. המידע יכול להיות מכמה סוגים: מידע ניסויי - בשלב זה נבחן מקורות רעש, איפה עלולות להיות טעויות, סטיות תקן ומידע ניסויי נוסף. מידע תיאורטי – מודלים שונים, אינטואיציות מדעיות.

2. ייצוג מתמטי של המידע והמערכות במחשב –

יש צורך לייצג במחשב את הרכיבים, האינטרקציות, והדינמיקה (חישובי כוחות). בשלב זה מכניסים מימד של פשטות, למשל החלפת אלפי אטומים שייצגו ככדור אחד. ייצוג האינטרקציות יכול להתבצע בעזרת פונקציה מתמטית פשוטה. ייצוג דינמיקה (דינמיקה בראונית – איפה המערכת תימצא בשלב הבא). בשלב זה ייתכן שנחזור לשלב הקודם ונשלים מידע נוסף שחסר. לבסוף יתבצע טיוב הנתונים מהעולם האמיתי לייצוג במחשב.

3. מציאת מודל טוב של המערכת –

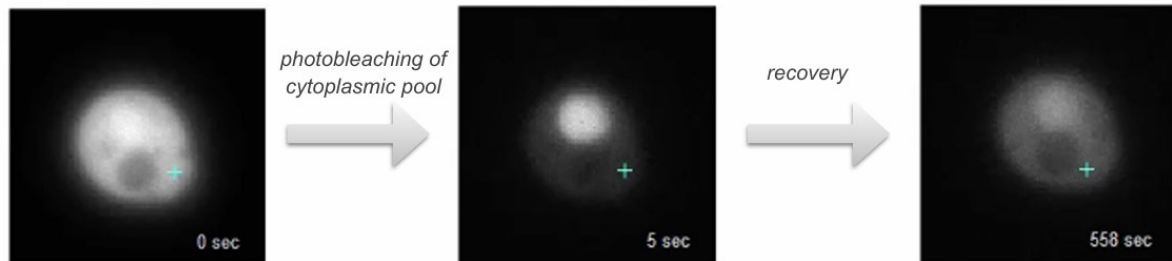
סורקים את האופציות של הפרמטרים, משפחות היפותזות. בשלב זה נדגום פרמטרים שיביאו לאופטימיזציה של המידע.

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

באמצעות מידע ב"ת תבוצע ואלידציה של התוצאות. בשלב זה ננסה להבין אילו אירועים לא טריוויאליים המודל מצליח לחזות. היתרון של מודל כזה על פני מודל נאיבי הוא האינטרפרטביליות שלו – זה לא מודל מתמטי או ML-י טהור (כמו רשת נוירונים) שבו לא ניתן לבצע שינויים שמדמים את המערכת האמיתית, ובמודל הזה נוכל לשנות פרמטרים כמו הכנסת מוטציות לחלבונים, למחוק שרשראות.. מה שלא אפשרי במודלים אחרים.

מה המודל טוען לגבי הסתננות של מולקולות לפי גודלן?

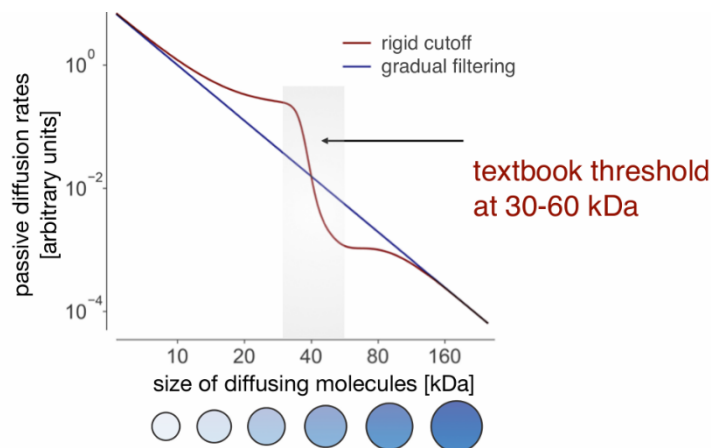
כל מולקולות ה-GFP בציטופלסמה עברו Photo Bleaching ולכן תוך 5 שניות כל הציטופלסמה הפסיקה לזהור:



כעבור מספר דקות (תמונה ימנית) מבחינים שרמת הזוהר בגרעין ירדה, והציטופלסמה זהרה יותר. זוהי הוכחה ניסויית למעבר חומרים מהגרעין לציטופלסמה.

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 בקבוע מסוים, זה פרמטר חופשי ולרוב צריך לתקן את התוצאות כדי להגיע לאותם ערכים מספריים.

בשיעור דנו בסוגייה איזה קו מגמה מתאר טוב יותר את קצב הפעפוע לפי גודל מולקולה:



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

לסיכום, למדנו שמידול אינטגרטיבי שלוקח כמה סוגים של מידע מאפשר להתמודד עם המורכבות של מערכות ביולוגיות – אבל חייבים לעבור שלבים של ולידציה.

היוריסטיקות לעימוד רצפים

לעתים הפתרון המוכח הוא יקר, ראינו שאפשר להגיע לזמנים קוואדראטיים בגודל הקלט. חלופה היא מציאת פתרון ביניים – שיטות היוריסטיות שלא מבטיחות לעמוד בכל המקרים, אבל להתקרב לתוצאה הטובה ביותר. דיברנו על בעיית העימוד הגלובלית, ושקילת כל המסלולים האפשריים. גישה היוריסטית אומרת אחרת – לעימוד "טוב" יש תכונות מסוימות, אפשר לנצל אותן.

גישה ראשונה - FASTA

נסתכל על שתי תכונות של עימוד שאנחנו מגדירים כעימוד טוב:

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

איך ההנחות האלו עוזרות לנו?

למשל בהינתן שני רצפים: $|s| = n, |t| = m$ ופרמטר k , נרצה למצוא כמה מילים באורך k משותפות בין הרצפים. שני חסמים על כמות האפשרויות: אם גודל הא"ב הוא Ω אז יש Ω^k אפשרויות כאלו, ואנחנו גם חסומים בגודל הרצפים (n, m) . האם העובדות האלו עוזרות לנו לצמצם בזמן? כן – נוכל להגיע לזמן לינארי על ידי צמצום כל המילים שמופיעות בשני הרצפים ובאיזה אינדקס הופיעו. נשים לב שהאינדקס בכל אחד מהרצפים i, j (ברצפים t, s בהתאמה) מגדיר לנו $\Delta = i - j$ שמגדיר פרמטר על ה"אלכסון" או "שיפוע" בטבלה של התכנון הדינמי. אם סופרים כמה פעמים מצאנו כל Δ , אפשר למצוא את האלכסונים שדרכם מומלץ להעביר את העימוד.

שיטת **FASTA** דוגלת בהתבוננות על מילים קצרות, חישוב Δ , וה- Δ שמופיעה הרבה תהיה מועמדת טובה לאלכסון שממנו מתחילים.

הדוגמה הזו היא בעיקר עבור האינטואיציה. איך נוכל לשפר את התכונות שדרשנו למעלה?

BLAST

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

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

במקום דרישה לדמיון (מושג עמום מעט), נרצה שאיפה **לציון טוב** על פי פונקציית score. נרצה לקבל התאמות שיקבלו ציון גבוה ונוכל להשתכנע שיש ביניהן קשר אמיתי.

לפעמים שינויים לא עולים הרבה ולכן הבחירה להחליף את דרישת identity בדרישת high score. איך האלגוריתם שלנו ישתנה? למשל אם s זו מילה שנרצה למצוא לה עימוד מול מסד הנתונים אז ניקח k ונשאל מהן המילים שהעימוד בגודל k מקבל עבורן ציון גבוה.

אנחנו עתידים למצוא מילים שיכולות למצוא ציון עימוד גבוה מספיק. שיטה זו מצמצמת את כמות המילים שאנחנו מעוניינים בהן למספר הרבה יותר קטן לעומת FASTA – כי הן יחסית נדירות. לאחר מכן, כאשר בידינו הקטעים הרלוונטיים נוכל לבצע עימוד לוקאלי החל מהנקודה שלהם (השלמת שאר הרצף).

שרשראות מרקוב

עד עכשיו דיברנו על מודל רצפים שנחשב יחסית "בנאלי" – $\mathbb{P}(x_1, \dots, x_n) = \prod_i \mathbb{P}_0(x_i)$

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

רצף של משתנים מקריים/אותיות/אירועים על פני ציר זמן – נהוג לתאר בעזרת **אירועים מרקוביים**.

שרשרת מרקוב – בדרך כלל מוגדרת על ידי תכונת מרקוב:

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

הערכים x_1, \dots, x_{i-1} לא משפיעים על המאורע ה- i . אם המאורע ה- $i+1$ מעניין אותנו, כל ערך שהיה לתחילת בגודל $i-1$ לא רלוונטי כל עוד יש בידנו את הערך ה- i . שימוש נפוץ הוא למשל בהילוך שיכור.

שרשרת מרקוב אחידה – מכילה את התכונה הבאה:

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

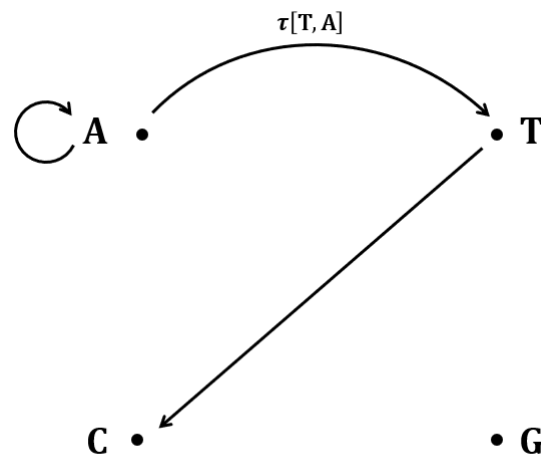
במילים פשוטות, הכוונה היא שהסיכוי של השיכור לזוז ממקום אחד לשני אינו תלוי באורך הדרך שעבר (הוא לא "מתעייף").

עם שתי ההנחות האלה, נחזור לרצף שלנו (x_1, \dots, x_n) ונטען כי:

$$\begin{aligned} \mathbb{P}(x_1, \dots, x_n) &\stackrel{\text{שרשרת}}{=} \mathbb{P}(x_1) \cdot \mathbb{P}(x_2 | x_1) \cdot \dots \cdot \mathbb{P}(x_n | x_1, \dots, x_{n-1}) \stackrel{\text{Markov}}{=} \\ &= \mathbb{P}(x_1) \cdot \prod_{i=1}^{n-1} \mathbb{P}(x_{i+1} | x_i) \stackrel{\text{התפלגות אחידה}}{=} \mathbb{P}_0(x_1) \cdot \prod_{i=1}^{n-1} \tau[x_{i+1}, x_i] \end{aligned}$$

מטריצת זוגות אותיות שמגדירה את הסיכוי לאות אחת בהינתן אחרת.

גרף שהקודקודים בו הם אותיות, ותהיה קשת בין קודקודים אם יש סיכוי לעבור ביניהם.



בנוקלאוטידים זה כנראה פחות מעניין, אבל יהיו סיטואציות בהן כן יהיה בכך שימוש.

איפה שרשראות מרקוב ישמשו אותנו? אפשר להסתכל על אבולוציה כתהליך מרקובי כאשר המצב בכל נקודה הוא מה הרצף שיש לי עד כה. בדרך כלל התבוננות אות אחת אחורה בחלבון נותנת יחסית מעט מידע, אנחנו נשתמש בשרשראות מרקוב לא ברמת האות הבודדת.

רקע ביולוגי להמשך

בחלקים נרחבים בגנום יש אזורים שלא מקודדים לחלבון והם נחשבים לאזורי בקרה. בהמשך תהיה הרצאת אורח שמטרתה להסביר מתי להשתמש באזורים בגנום. תופעה שנקראת מטילציה של DNA – שינוי שקורה באות C שהופכת ל-C*. אחד הדברים שקרו באבולוציה נבעה מהצורך להעביר את המידע לדור הבא. בגדיל DNA מול C יש G, אבל נרצה שגם את C* תהיה לנו דרך להעביר. הפתרון האבולוציוני הוא להסתכל על C שאחריו מגיעה G ואז בכיוון ההפוך הרצף יזוהה בצורה תקינה – האנזים ישים את הסימון על האות C גם בגדיל המועתק. תופעה אבולוציונית היא שאנזימים שמסמנים מטילציה התחילו להעדיף C שאחריהם מגיעה G, ולסמנם אוטומטית.

אממה, כששמים את הסימון משנים מעט את C ובכך חושפים אותו יותר למוטציות – גובר הסיכוי שנאבד את C, מה שמוביל לאיבוד fitness. נוצר מצב עם אזורים בגנום בהם הסיכוי שאחרי C באה G הוא גדול – להם נקרא **עשירים ב-CpG**, ומנגד אזורים בהם הסיכוי שאחרי C באה G הוא דל ולהם נקרא **עניים ב-CpG**.

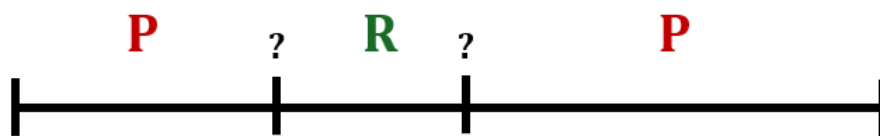
איך היינו בונים מטריצת מעבר לפי האם האזור עשיר/עני ב-CpG? נגדיר שתי מטריצות שיבדלו ביניהן בתא המסומן:

| τ | A | C | G | T |
|--------|---|---|---|---|
| A | | | | |
| C | | | | |
| G | | | | |
| T | | | | |

לפעמים האזור המסומן יהיו נמוך, ולפעמים קצת יותר גבוה מ- $1/4$.

איך נשתמש בזה? אפשר לקחת רצף ולחשב לפי 2 שרשראות מרקוב שונות. נשתמש בסף נראות כדי להחליט איזה אזור יותר סביר שזה יהיה (עשיר/עני ב-CpG).

צריך לדעת להגדיר/לתחום את האזור – בתקווה באמצעות מודל הסתברותי שיבצע זאת אוטומטית:



מודל אחד שנדבר עליו יהיה מודל תהליך חבוי:

תהליך חבוי – נניח שנרצה להוסיף סיווג לכל עמדה ברצף x_1, \dots, x_n ע"י משתנה מקרי חבוי H_1, \dots, H_n שמדווח על סטטוס העמדות.

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

מודל מרקובי חבוי הוא כזה שבו נניח שמתקיימת שרשרת מרקוב: $\mathbb{P}(H_1, \dots, H_n)$ כלומר שהתוויות החבויות מתוארות על ידי שרשרת מרקוב. סביר שנשאר באותו מצב $P \setminus R$ ואת המעברים ביניהם אפשר לשייך לאורך אופייני של Strehl של אזור עשיר או עני.

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

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

עדיין יש סימני שאלה – איך משתמשים במודל כזה?

מודלים מרקוביים חבויים – Hidden Markov Model

כפי שראינו בשיעור הקודם, מודל מרקובי חבוי זה סט של משתנים / שרשרת תצפיות $\mathbb{P}(x_1, \dots, x_n | H_1, \dots, H_n)$, כאשר ההנחה היא שזו שרשרת מרקוב: $\mathbb{P}(H_1, \dots, H_n)$ כלומר שמתקיים גם:

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

ויזואלית, זה אומר המצבים החבויים הם שרשרת מרקוב שמתקדמת בזמן, בכל שלב ההתקדמות תלויה רק במצב האחרון (ולא איך הגענו אליו), אפשר להתייחס לזה כאוטומט סופי דטרמיניסטי שמבוסס הסתברות. התצפיות (x -ים) הן תצפיות שתלויות במצב הנוכחי. אפשר לחשוב על זה כך שהמצב החבוי זה מצב כל השחקנים במשחק, והתצפיות זה רק מה שהמשחק מגלה לנו באותו רגע על המסך (תלוי במצב, אך לא גלוי לנו). לפעמים התצפיות רועשות ולכן התצפית היא בהינתן המצב החבוי.

דוגמה

משחק הימורים מול שחקן שמטיל קובייה, והוא טוען שיש לו קובייה חוקית. בלי שנשים לב הוא מחליף את הקובייה לקובייה מוטה, והוא לא תמיד יכול לבצע את ההחלפה (רק כשאנחנו לא שמים לב). זה "hidden" כי אנחנו לא יודעים מה הקובייה שהיריב מחזיק, אלא רק סדרת תצפיות.

סימונים:

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

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

שימושים במודל:

1. בהינתן תצפיות, מה היה המצב בזמן i ? $\mathbb{P}(H_i | x_1, \dots, x_n)$
2. מהו $\arg \max_{h_1, \dots, h_n} \mathbb{P}(H_1 = h_1, \dots, H_n = h_n | x_1, \dots, x_n)$ במילים אחרות, מהו רצף המצבים החבויים הכי סביר בהינתן התצפיות? זוהי בעיית שחזור. בשאלה 2 מעניין אותנו רק המצב החבוי הנוכחי.
3. מהו הסיכוי לראות את התצפיות שלי? $\mathbb{P}(x_1, \dots, x_n)$, $likelihood$. כתת-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:

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

הבעיה היא שיש כאן סכום על n ערכים, כלומר מספר אקספוננציאלי ב- n של מקרים. אחת הדרכים לפתור את זה היא לקחת את הביטוי שסימנו ב-*, ולהתחיל להפעיל חוקים אלגבריים. אנחנו נציג דרך אחרת באמצעות תכנון דינאמי. לשם כך, נכתוב את הבעיה ללא הסכום האקספוננציאלי.

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

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

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

נשים לב שהביטויים **בכחול** מזכירים זה את זה, פחות אינדקס אחד. זה אומר שאם נדע לפתור אחת יהיה פתרון גם לשנייה. נזכיר שאנחנו בשרשרת מרקוב, כלומר בהינתן המצב הנוכחי H_n לא תלוי במה שקרה קודם, כולל תצפיות של המצבים ולכן נשאר עם:

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

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

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

בעצם יש לנו את המטריצה הבאה:

| | | | | | | |
|----------|---|--|--|---|--|---|
| | 0 | | | i | | n |
| F | | | | | | |
| L | | | | | | |

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

קיבלנו פונקציה רקורסיבית אבל טרם הגדרנו תנאי התחלה. נגדיר:

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

ניקח דוגמה מספרית:

| | | From | |
|----|----------|------|------|
| | | F | L |
| To | F | 0.99 | 0.05 |
| | L | 0.01 | 0.95 |

| π | F | L |
|-------|---------------|----------------|
| 1 | $\frac{1}{6}$ | $\frac{1}{10}$ |
| 2 | $\frac{1}{6}$ | $\frac{1}{10}$ |
| 3 | $\frac{1}{6}$ | $\frac{1}{10}$ |
| 4 | $\frac{1}{6}$ | $\frac{1}{10}$ |
| 5 | $\frac{1}{6}$ | $\frac{1}{10}$ |
| 6 | $\frac{1}{6}$ | $\frac{1}{2}$ |

עם התצפיות: $obs = [2, 6, 6]$. אנחנו מניחים שאנחנו לא יודעים איזו קובייה היריב מחזיק בהתחלה.

| | | | |
|----------|---|---|---|
| | 1 | 2 | 3 |
| 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 תצפיות מאחר שההטלה השנייה תוצאתה הייתה 6, קיבלנו עלייה בסבירות שמדובר בקובייה 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 התצפיות עומד להיות סכום $\spadesuit + \star = 0.01397$.

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

איך נטפל בבעיה? נשים לב שבנוסחה יש $b = a \cdot \sum c_i$, ובמקרים כאלה נפוץ לשמור log representation כלומר $\tilde{x} = \log b$ ואז $\log b = \log(e^{\tilde{a}} \sum_i e^{\tilde{c}_i})$. נבחין כי $\tilde{b} = \tilde{a} + \log \sum_i e^{\tilde{c}_i}$. נשתמש 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 שעלולים לגרום ל-underflow הוצאנו מחוץ לסכימה.

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

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

כלומר הסיכוי להיות ב- s הוא פרופורציונלי ל- F_n אבל צריך לנרמל כדי שישכם ל-1.

אבל, לפעמים נרצה לדעת מה קרה קודם לכן:

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

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

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

$$\stackrel{\text{סימון}}{=} F_i[s] \cdot B_i[s]$$

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

$$B_i[s] = \mathbb{P}(x_{i+1}, \dots, x_n \mid H_i = s) \stackrel{\text{הסתברות שלמה}}{=} \sum_t \mathbb{P}(x_{i+1}, \dots, x_n, H_{i+1} = t \mid H_i = s)$$

$$\stackrel{\text{כלל השרשרת}}{=} \sum_t \mathbb{P}(x_{i+2}, \dots, x_n \mid H_i = s, H_{i+1} = t, x_{i+1}) \cdot \mathbb{P}(x_{i+1} \mid H_{i+1} = t, H_i = s) \cdot \mathbb{P}(H_{i+1} = t \mid H_i = s)$$

נבחין שאת 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 **Forward** ב- $O(n)$ וגם על **Backward** ב- $O(n)$, כלומר אפשר לענות על שאלות מהסוג $\mathbb{P}(H_i = s \mid x_1, \dots, x_n)$ לכל i 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 ששניהם בלתי תלויים אחד בשני, אבל שילובם יספק מענה לשאלה שעניינה אותנו.

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

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

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

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

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

Maximum Probability Reconstruction – בעיית השחזור

בעיה של מציאת ה-h-ים שגורמים לתצפיות שלנו להיות הכי 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 בתצפיות, כלומר נמקסם את $\mathbb{P}(H_1 = h_1, \dots, H_n = h_n | x_1, \dots, x_n)$ נקבל את אותה הבעיה, אבל יותר קל לעבוד ללא ההתניה.

כמו בעיות רבות בקורס, נפתור גם את הבעיה הזו באמצעות תכנון דינמי. תתי הבעיות:

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

בביטוי הזה אפשר לשאול מה המקסימום על המשתנה האחרון:

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

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

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

נוציא מתוך ה-max הפנימי את הביטויים שלא קשורים ל- h_i ונשאר עם:

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

נגדיר:

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

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

כלומר, כשאנחנו שואלים מה שרשרת המצבים הטובה ביותר שמסיימת בעמדה i עם ערך $H_i = s$, הדרך לחשוב על זה היא להסתכל על כל המצבים עד $i - 1$ ולכל אחד מהם צריך להתייחס כאילו החל מהם ממשיכים ל- s והתצפית הרלוונטית בהינתן s .

לאלגוריתם זה קוראים **אלגוריתם Viterbi**.

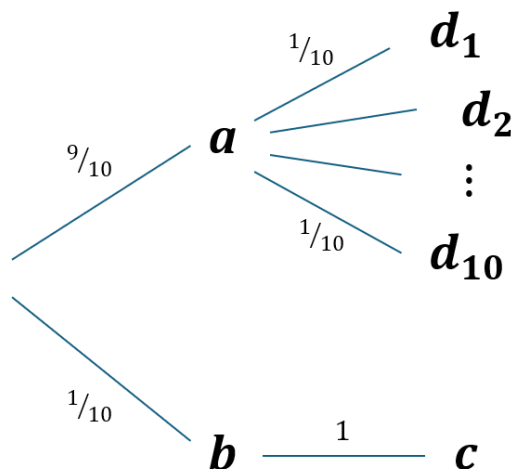
שאלת הבנה

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

נגדיר $\tilde{h}_i = \arg \max_{h_i} \mathbb{P}(H_i = h_i | x_1, \dots, x_n)$ כלומר עם הסיכוי הגבוה ביותר בזמן i .

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

האם $\hat{h}_i \stackrel{?}{=} \tilde{h}_i$? אינטואיטיבית היינו רוצים שזה יהיה המצב, אך זה לא כך. אחת הדרכים היא לייצר דוגמה מנוונת לחלוטין שעושה את זה. בבנה שרשרת מרקוב שבה השרשרת הכי סבירה אינה הערך הכי סביר:



כאן הערך הכי סביר בזמן 1 הוא כמובן a , אבל השרשרת הכי סבירה היא $b \rightarrow c$.

מאחר שאנחנו צריכים להתחייב על כל מצב בשרשרת כאשר מבצעים שימוש באלגוריתם Viterbi, אנחנו צריכים במסלול שעובר דרך a לעבור דרך מסלול שמכיל בתוכו אי-ודאות, ולכן הסיכוי נמוך. \tilde{h}_i למשל, יכול להביא רצף כמו $a \rightarrow c$ שאינו אפשרי בסדרת המצבים שהגדרנו.

כאשר הגדרנו את $V_i[s]$ קיבלנו רק את הערך המקסימלי ולא את השרשרת שהובילה אליו. גם כאן אנחנו נתקלים בבעיה דומה לבעיית העימוד, ונרצה לשמור במטריצה שאנחנו ממלאים גם את המסלול, ואז נוכל ללכת אחורה ולשרשר.

אנחנו מעוניינים ללמוד H_n , מאיפה לומדים אותו? ללמוד את τ (הסיכוי לעבור בין מצבים) וגם את π (הסיכוי לכל מצב).

נתחיל בבעיה הקלה שנקרא לה Full Data. במקרה זה התצפיות שלנו מורכבות מזוגות של רצפים:

$$\begin{aligned} 1) & \begin{bmatrix} h_1 & h_2 & \cdots & h_{n_1} \\ x_1 & x_2 & \cdots & x_{n_1} \end{bmatrix} \\ 2) & \begin{bmatrix} h_1 & h_2 & \cdots & h_{n_2} \\ x_1 & x_2 & \cdots & x_{n_2} \end{bmatrix} \\ & \vdots \\ M) & \begin{bmatrix} h_1 & h_2 & \cdots & h_{n_m} \\ x_1 & x_2 & \cdots & x_{n_m} \end{bmatrix} \end{aligned}$$

סה"כ M דוגמאות שכל אחת מהן היא רצף של תצפיות ורצף של מצבים שפלטו אותן. נספור כמה פעמים כל אירוע קורה, נבחן תדירות (maximum likelihood). האם יש מצבים בחיים בהם נוכל להשיג data שנראה ככה? כן למשל התצפיות הן חלבוניים (רצפים של ח. אמינו), ובתור hidden state מידע על המבנה השניוני שלהם. במצב כזה, בהינתן רצף חדש אם נריץ עליו Viterbi הוא יבנה את המבנה השניוני הסביר ביותר. רוב המקרים בחיים אינם כאלה.

ננסה 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 כל ה-data שיש לנו. נמשיך:

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

מאחר שיש כאן רק מכפלות ללא סכומים, הכל קומוטטיבי ואפשר לסדר באיזו דרך שנרצה:

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

יש לנו M תצפיות. 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 $N_{0,s} = \sum_m \mathbb{1}\{h_1[m] = s\}$ 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 $N_{t,s} = \sum_m \sum_i \mathbb{1}\{h_{i-1}[m] = t, h_i[m] = s\}$ (מספר הפעמים שראיתי מעבר מ- t ל- s), ונגדיר $N_{s,x} = \sum_m \sum_i \mathbb{1}\{h_i[m] = s, x_i[m] = x\}$ (כמה פעמים ראיתי את התופעה שאני במצב s וקיבלתי x).

עבשיו נרצה לעשות מקסימום לביטוי. הדרך הנוחה תהיה להשתמש בבעיה שכבר פתרנו – אין אילוץ שמקשר בין \mathbb{P}_0 לבין τ למשל, 3 בעיות שונות (מסומנים לעיל ב-3 סוגריים [] שונים). נשתמש בתכונה 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 ללמידת קובייה:

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

הבעיה **הירוקה**: אנחנו יודעים על τ שבהינתן 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:

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

לכן כדי לממש את זה בפועל, צריך לאסוף את הסטטיסטיים לכל s לכל t, s ולכל x, s מהדאטה, ואז השמה לתוך המשוואות שסומנו בריבוע.

הבעיה הקשה

במקרה הזה ה- $hidden\ state$ לא גלוי לנו. כלומר הדאטה הוא מהצורה:

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

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

⋮

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 (במונחים של τ ו- π) נצטרך לפרק ביטוי שבתוכו יש סכומים – הערבוב שלהם קשה יותר, ולא נזכה לקומוטטיביות שהייתה במקרה הקל. אם נשליך את הבעיה לבעיית Machine Learning, אפשר לראות קווי דמיון לבעיות unsupervised בהן שייכנו דאטה לתוך clusters מבלי שנתונות לנו תוויות על הדאטה.

הפתרון שנראה לבעיה הזו 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 הדאטה שלנו לקבוצות שונות של hidden values, ואז נשמש בפרמטרים כדי לבצע ניחוש מושכל יותר של החלוקה.

האלגוריתם ישחק בין הפרמטרים τ, π לבין ה- $Hidden\ Values$. אם נדע את הערכים הנסתרים נוכל לחשב את הפרמטרים, ואם נדע את הפרמטרים אולי נוכל לשפר את הניחוש של הערכים הנסתרים.

אנחנו נמצאים במצב שבו הדאטה שלנו נראה בצורה הזו:

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

סט הפרמטרים שאנחנו מעוניינים ללמוד הוא:

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

האסטרטגיה שדיברנו עליה היא שאם נוכל לשערך פרמטרים (Θ) אז נוכל לשערך גם סטטיסטיים $(N_{0,s}, N_{s,t}, N_{s,x})$, וגם ההפך. נתחיל עם ניחוש התחלתי של פרמטרים. הניחוש ההתחלתי משפיע על התוצאה הסופית, אבל נשאיר את הדיון על כך להמשך.

סקיצה של האלגוריתם:

$$\begin{aligned} \Theta^1 & \leftarrow \text{Guess} \\ \text{For } j & = 1, \dots \\ N^j & = ??? \\ \Theta^{j+1} & \leftarrow \text{MLE} [N^j] \end{aligned}$$

נותר לדבר על איך מאתחלים, על מתי עוצרים ועל איך קובעים את N^j . את MLE בהינתן סטטיסטיים ראינו בשיעור שעבר.

יש שני ואריאנטים של האלגוריתם הזה: הראשון נקרא Max-Max (נקרא כך כי עושים מקסימום ל-MLE ומקסימום לסטטיסטיים). השני נקרא Expectation-Maximization (Exp-Max).

אלגוריתם Max-Max (MM)

מה יהיה ה-Max Step באלגוריתם?

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

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

כלומר בהינתן ה-Data קובעים פרמטרים, וראינו אלגוריתם שמוצא את המצבים החבויים הכי סבירים – ויטרבי.

קל מאוד להוכיח שהאלגוריתם משפר בכל צעד, כי כאשר מעדכנים את \vec{H}^j אנחנו מוצאים את H שממקסם את הנראות, ובצד של ה-likelihood בחרנו פרמטרים שממקסמים את הנראות. בסופו של דבר, Max-Max ממקסם את הפונקציה $\mathbb{P}(H, X, \theta)$.

אלגוריתם (EM) Exp-Max

אנחנו נתמקד באלגוריתם הזה. נגדיר: $\vec{N}^j = \mathbb{E}[\vec{N} | \vec{X}, \theta^j]$. מכאן השם Expectation, כי אנחנו לא מחשבים את הספירות על הדאטה החבוי הכי סביר, אלא למצע על כל ה-Hidden States. מתמטית, חושבים על זה כסכום הבא: $\sum_{\vec{H}} \mathbb{P}(H | \vec{X}, \theta^j) \cdot \vec{N}(H)$. הרבה פחות נוח לעבוד עם הביטוי הזה מבחינה קומבינטורית – לוקחים את כל ההשמות האפשריות 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 מטיילים על אוסף התוצאות הדיסקרטיות שאפשר לקבל).

חישוב עבור $N_{0,x}$

$$\begin{aligned} \mathbb{E}[N_{0,x} | \vec{X}, \vec{\theta}] &= \sum_m \mathbb{E}[1\{H_1[n] = s\} | \vec{X}, \vec{\theta}] \stackrel{\text{תוחלת של אינדיקטור בינארי}}{=} \sum_m \mathbb{P}(H_1[m] = s | \vec{X}, \vec{\theta}) \\ &= \sum_m \mathbb{P}(H_1[m] = s | \vec{X}[M], \vec{\theta}) \stackrel{i.i.d}{=} \sum_m \mathbb{P}(H_1[m] = s | \vec{X}[m], \vec{\theta}) \end{aligned}$$

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

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

חישוב עבור $N_{s,x}$

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

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

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

[חישוב עבור \$N_{s,t}\$](#)

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

[חזרה לסטטיסטיים](#)

בשדיברנו על סטטיסטיים הייתה שאלה מה הערך של:

$$\mathbb{P}(H_i = s, H_{i+1} = t, \vec{X}) \stackrel{\text{כלל השרשרת}}{=}$$

$$\begin{aligned} &\cdot \mathbb{P}(x_1, \dots, x_i, H_i = s) \\ &\cdot \mathbb{P}(x_{i+2}, \dots, x_n \mid H_{i+1} = t, x_1, \dots, x_i, H_i = s) \\ &\cdot \mathbb{P}(H_{i+1} = t \mid H_i = s, x_1, \dots, x_i) \\ &\cdot \mathbb{P}(x_{i+1} \mid H_{i+1} = t, H_i = s, x_1, \dots, x_i, x_{i+2}, \dots, x_n) \end{aligned}$$

מכל מה **שמסומן באדום** אפשר להיפטר הודות לתכונות השרשרת. נקבל בעזרת שימוש בנוטציות:

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

הביטויים מופיעים לפי הסדר.

פירוט האיטרציה

עכשיו אפשר לעדכן האלגוריתם. מה עושים בכל איטרציה? עוברים על כל הסדרות, מחשבים Forward-Backward ואת כל הסטטיסטיים. דרך אחת לעשות את זה היא:

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

מבחינת סיבוכיות זמן איטרציה אחת עושה F/B על כל רצף באינפוט (זמן לינארי). גרסה מעט יעילה יותר משלבת את שני השלבים הראשונים באיטרציה ומבצעת במקביל (אין צורך לשמור F/B של כל ה-m-ים אם אוספים את הסטטיסטיים במעבר). גם מבחינת מקום, נקבל סאב-לינארי בגודל האינפוט.

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

משפט

$$1. \mathbb{P}(\vec{X}, \theta^{j+1}) \geq \mathbb{P}(\vec{X}, \theta^j)$$

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

כלומר כל איטרציה כזו משפרת פונקציית נראות של התצפיות, ואם התכנסנו אז הגענו למקסימום לוקאלי

$$L(\vec{\theta}) = \mathbb{P}(\vec{X}, \vec{\theta})$$

תנאי העצירה

לפי המשפט שראינו $L(\theta^{j+1}) \geq L(\theta^j)$ ולכן מחוקי לוגריתמים $\ell(\theta^{j+1}) \geq \ell(\theta^j)$ (שכן הגדרנו את ℓ להיות $\ell(\vec{\theta}) = \log \mathbb{P}(\vec{X}, \vec{\theta})$). כלומר אפשר לעקוב אחרי ההתקדמות, ואם אנחנו רואים שמתחילה דעיכה כנראה שנתכנס. אנחנו יודעים:

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

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

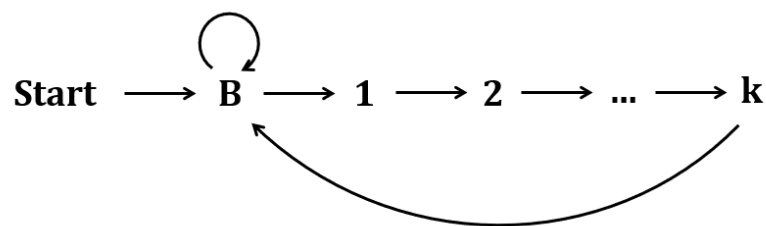
מבט מלמעלה

ראשית, הנחנו שאנחנו יודעים כמה מצבים חבויים יש לנו במערכת. נניח כרגע שיש לנו k מצבים חבויים. אנחנו צריכים לבחור אתחול אקראי, אבל קיים חשש שנתכנס לאזור בעייתי. איך בוחרים נקודת התחלה טובה?

נדבר על מושג שנקרא להגדיר "Fixed structure". דוגמה – אם אנחנו רוצים להסתכל על רצף של DNA עם מחזוריות שנובעת מאיזו תכונה, כלומר שיש דפוס מעבר ממצב למצב. כאשר נבצע EM, מה יקרה לכל המעברים הלא אפשריים (אלו שההסתברות שיקרו היא 0)? אנחנו עושים תוחלת על כל ה-hidden 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 $\tau^1[t, s] = 0$ אז גם $\tau^j[t, s] = 0$ לכל האיטרציות. כך גם עם מטריצת הפלטים.

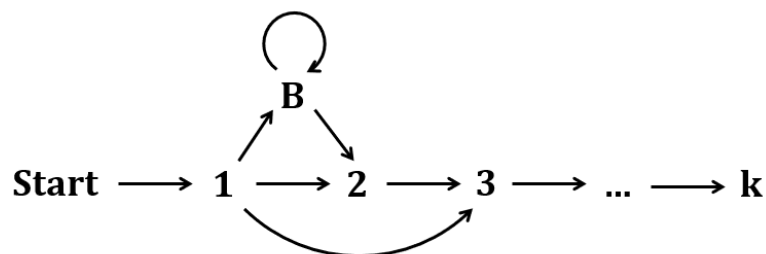
מודלים לדוגמה

1. מילה באורך k



אילו רצפים של hidden states הם אפשריים? מהסגנון B,B,B,1,2,...,k,B,1,2,...,k,B. כך החבאנו מילה באורך k בתוך הרצף.

2. למידת רצף חלבוני כולל שינויים



לומד עמדות חשובות בחלבונים דומים, ומה הסיכוי ל-insertion או deletion בכל עמדה. בהינתן רצף חלבוני חדש, נוכל ללמוד האם סביר שהגיע מאותו אב קדמון.

הגינם וכרומטידים

שיעור העשרה – נספר היום על הביולוגיה, ובמקביל על שאלות חישוביות שעולות כשאנחנו חוקרים את הביולוגיה.

גינם – רצף של DNA שנמצא בתאים. למשל גנום אנושי, גנום של בקטריה וכדומה.

אם נפשט מבחינת מדעי המחשב, מדובר בקובץ שמכיל רצף של המון אותיות המתאר את ה-DNA. בהרבה יצורים אפשר לחלק אותו לכמה רצפים. בשיעור ההקדמה דיברנו על אזורי שנקראים גנים. בעולם הביולוגי לפעמים נצייר בצורה הבאה כדי לסמן איפה מתחיל הגן שיהפוך לחלבון (בתוך המלבן) ואיפה מתחיל להיווצר ה-RNA שמכיל את הגן הזה (תחילת החץ והנקודה שבסופה):



מבחינה כימית מדובר במולקולה ארוכה, הסימונים והנוטציות האלו הם על סמך תוצאות שראינו (החל מנקודה מסוימת מתחיל להיווצר RNA, זה לא כתוב עליו בשום צורה). זה מעלה שאלות, איך התא שבתוכו נמצא ה-DNA יודע להשתמש במקטע 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 כאלו במקומות אחרים – אבל זו דרישה קשה.

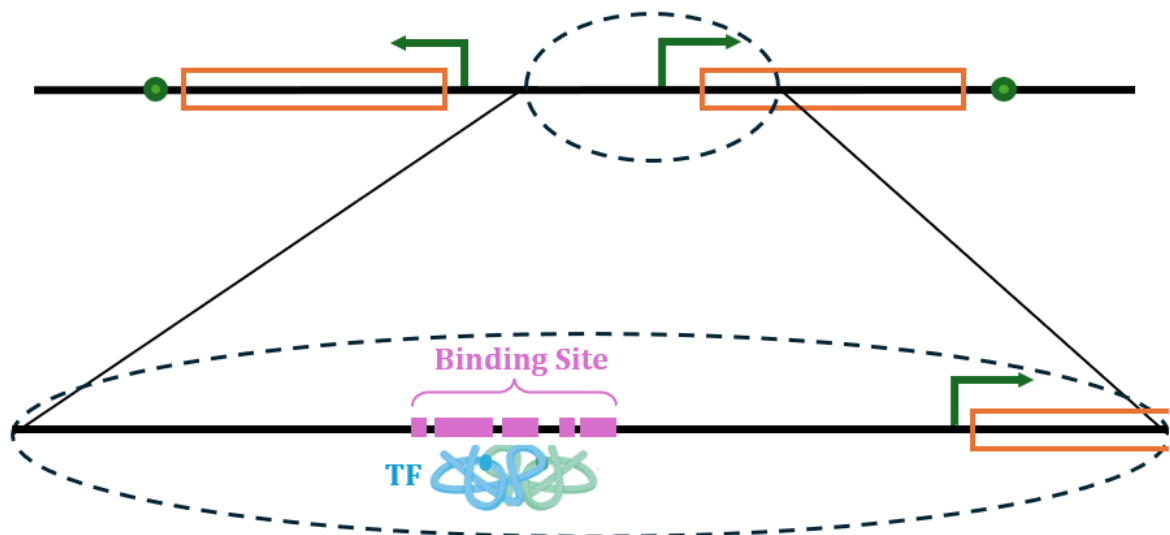
יש עוד אינפורמציה במהלך הרצף שמכוונת את התאים להשתמש ב-DNA בצורה ה"נכונה". השאלה היא מהי האינפורמציה הזו ואיך ניגשים אליה. אם נתחיל בשאלה שניצבת בפני יצור פשוט כמו חיידק, רוב ה-DNA שלו הם גנים והוא מצליח להשתמש בהם נכון בעזרת שימוש בעובדה שבאזור שלפני החץ הירוק יש מילה או כמה מילים שמורות שמופיעות בנקודה שממנה אפשר להתחיל לייצר RNA. אותה מילה בדיוק רואים באותם רווחים בדיוק את אותה מילה, והיא מסמנת להתחלת שעתוק.

איך מגלים מילות מפתח בטקסט?

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

למה הביולוגיה משתמשת בדברים כאלו? יש משפחות חלבונים שמבחינה מבנית הם 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 ההצמדות שלו ל-DNA נפגעת והוא לא יישאר שם זמן רב.

לחלבונים כאלו קוראים DNA-binding proteins, אבל יכול להיות גם סתם חלבון כזה – אנחנו נתמקד באחד עם sequence preference, כלומר כזה שיש לו העדפה לרצפים מסוימים. נקרא לחלבונים האלו Transcription Factors². הנוכחות של החלבון הזה משפיעה על תהליך השעתוק. לאתר שאליו הוא נקשר אנחנו נקרא binding site.



יש מספר חלבונים כאלו, כל אחד עם העדפות שלו, ואם הם נוכחים בתא אז הם ייקשרו ל-DNA ובמקומות מסוימים יופיעו החלבונים המתאימים. אם נוכל להבין את הדפוס של איזה חלבון נקשר לאן, היינו יכולים לצייר מיפוי של איזה חלבון יושב איפה. בגנום האדם יש כמה מאות ואולי אלפים של חלבוני TF (קושרי DNA), רובם עם העדפות קישור, עדיין לא כולם ידועים לנו.

זה שחלבון נקשר ל-DNA לא מבטיח לנו שמשהו יקרה בעקבות זאת. בחיידקים אנחנו מבינים די טוב את התהליך – יש שני חלבונים שכשהם נקשרים יחד מגיעה מכונה נוספת ונקשרת אליהם, ואז יש את כל הרכיבים הנדרשים להתחלת השעתוק.

[איך עובד המנגנון שעוזר/מפריע לשעתוק?](#)

אנחנו רוצים מנגנון שיגיד בסיטואציות מסוימות מה לעשות – למשל בהיעדר סוכר להתחיל תהליך שידע להכין חלבון שיוזע לייצר סוכר ממשאבים אחרים, או עודף סוכר 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.

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

איך הסיפור הביולוגי שתיארנו מיתרגם לשאלות חישוביות?

נתחיל משאלה פשוטה – יש הרבה ניסויים כדי להבין העדפות קישור של חלבונים. אחת מהשיטות היא Immuno Precipitation³. הרעיון הוא שאם יש נוגדן (שמטרתו לזהות גורמים זרים ולעודד התקפה עליהם) הוא יודע לזהות מטרות בעזרת רצף של חלבון. מערכת החיסון הלומדת מנסה רצפים רבים של חלבונים עד שמוצאים אחד שנקשר לגורם הבעייתי. אפשר לגדל חלבונים נגד חלבון מטרה, ואנחנו עומדים להשתמש בתכונת הנוגדן כדי לבקש נוגדן נגד TF.

נוכל לקחת תא שיש בו גנום ומקומות בהם TF מופיע ונקשר לגנום, אם יש לנו נוגדן יש לנו דרך "לשגר" מטען בדיוק למקום שבו החלבון מופיע. התהליך של Immuno Precipitation הוא לשבור בצורה פיזית את הגנום לפרגמנטים, והנוגדן ייקשר רק לפרגמנטים שקשור אליהם גם TF. אם יש לנו דרך "לדוג" את הנוגדן, נשאר רק עם החתיכות שבהן יש TF.

עכשיו אפשר לשאול על הפרגמנטים שבידנו שאלות. למשל, אפשר להתחיל לרצף את הפרגמנטים (לקחת מבחנה עם חתיכות DNA ולהפוך אותם לקובץ במחשב של הרצפים שלנו). נתחיל ממיליוני תאים ונקבל רצפים אותם נוכל למפות בחזרה לגנום. נוכל לצייר איפה הרצפים שריצפנו יושבים על הגנום. בעיה חישובית שלא נתמקד בה עכשיו היא בעיית העימוד, איך עושים עימוד של המון רצפים קצרים מול גנום.

בעיית Peak Calling

שברנו את ה-DNA ומצאנו את הפקטור. אם יש binding site אמיתי נקבל רצפים שמכילים אותו אבל יכלו להישבר בכל מיני מקומות סביבו. אחת השאלות שאפשר לחקור היא למצוא את כל המקומות שסביר שקיים בהן אתר (מופיע בהרבה רצפים). אם היינו אוספים כמה רצפים מכסים כל קטע היינו מקבלים פסגה (הם נערמים) ומכאן השם Peak Calling, כי בשאר המקומות ככל שנתרחק מהאתר יש פחות רצפים. מדובר בקובבולוציה שמתארת את השבירה. נציין שאם יש כמה אתרים זה לצד זה העניין עלול להסתבך.

אחרי שמצאנו את ה-peaks בדרך כלל ה-peak caller יסמן לנו אזורים בגנום בהם אנחנו די בטוחים שיש אתר קשירה. הביולוגיה מתחכמת בנושא הזה – לפעמים החלבון יהיה צמוד ל-DNA אבל לא בגלל שהוא זיהה את ה-DNA אלא בגלל שיש לו חלבון אחר שזיהה את ה-DNA והוא אוהב להיקשר אליו, ונקבל אתרים שמכילים אתר קשירה של חלבון אחר.

בעיית Find Enriched Words

נניח שנשאר בעולם הפשוט בו חלבון נקשר רק לאתר שלו – שאלה חישוביות שקשורה לכך היא Find Enriched Words. בהינתן אוסף מילים בטקסט שברוב הפסקאות מופיעה מילה אחת, לא ידוע איפה, ואנחנו

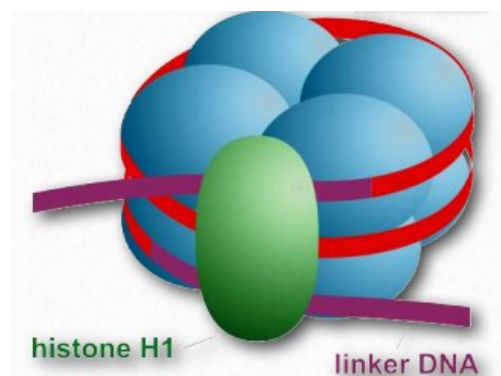
³ Immuno – תלוי במערכת החיסון, Precipitation – השקעה.

נדרשים למצוא אותה. כדי שלא נמצא מילים שכיחות אנחנו נקבל גם את יתר הגנום 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 כאלו דומה לבעיה בטקסט מעל א"ב, רק שהא"ב יותר מנוון מאשר באנגלית, וגם המילים אינן שמורות (מדובר בענן של מילים דומות עם יכולת החלפה, האות השנייה למשל חייבת להיות נכונה אבל את השלישית אפשר להחליף וכו'). אפשר לקחת את התוצאות של ניסוי כזה ולזהות אתרי קשירה שמסביר למה זיהינו אתרים מסוימים ולא אחרים עד כדי טעויות מדידה וכדומה.

אפשר לבחון את הניסוי גם על אותה חיה בתנאי גידול שונים כדי לבדוק האם הפקטור משנה את אתרי הקישור שלו בתגובה למצב.

חיידקים הם יצורים שעברו אופטימיזציות כדי למזער את מספר הרכיבים הפועלים בהם כדי לקבל מערכת שעובדת היטב ללא חלקים מיותרים. בקפיצה האבולוציונית מחיידקים ליצורים עם גרעין קרה תהליך שאפשר ליצורים האלו לבצע פונקציות יותר מורכבות, ועמו הגיע גם בקרה. בצורה מופשטת, כל חתיכה של DNA בחיידקים יכולה לעבור שעתוק. ביצורים אוקריוטים (בעלי גרעין) ה-DNA נראה מעט שונה מזה של חיידקים, ועטופים בחלבונים בצורה יותר אדוקה מאשר אצל חיידקים.

מעבר לשאלה האם החלבון נקשר או לא נקשר, נוסיף ממד שמתייחס לקישור ברירת מחדל של חלבונים שנקראים היסטונים שיוצרים קומפלקס (מצבור) של חלבונים שסביבו מלופף DNA דמוי דיסקית. אפשר לבצע 2 ליפופים של ה-DNA סביב הדיסקית. הדיסקיות מאוד דביקות ל-DNA (נדרשת הפעלת אנרגיה כדי להפריד אותו). לקומפלקס כולו (היסטונים ו-DNA) קוראים נוקלאוזום.



(התמונה היא תוספת שלי רק להמחשת המבנה, לא הזכרנו מהו linker DNA)

התכונה הבסיסית של יצורים אוקריוטים זה שיש "מכונות" שכאשר הן נתקלות ב-DNA פנוי הן עוטפות אותו לכדי מבנה של נוקלאוזום. בררת המחדל עברה מבקטריה שבה צריך בצורה מתוכננת להפריע כדי לשעתק, לברירת מחדל שבה כל DNA פנוי ייתפס.

השימוש של TF הרבה פעמים הוא רק להיקשר ל-DNA ולהפריע לקישור נוקלאוזומים – בגלל שהם נקשרים שם אי אפשר להרכיב שם נוקלאוזום.

לפעמים פקטורי השעתוק מטרם לסמן אזורים שאנחנו לא רוצים שיהיו עטופים בנוקלאוזום – נוצר משחק של להחביא את הגנום או לא. הרבה פעמים נדבר על האזורים שלא עטופים בנוקלאוזומים בתור החלק הנגיש של הגנום (כל מה שנקשר ל-DNA יכול לעשות זאת). זה שימושי כי אם נוכל לבקר על אילו אזורים חשופים, על אותו רצף גנומי נוכל להציג קומבינציות שונות (כמה גנומים שונים בתוך רצף אחד).

כשיצורים הפכו לרב-תאיים זה עבר עוד שלב של התמחות (בתא כבד אנחנו רוצים תכונות ספציפיות שלא רלוונטיות לנוירון למשל). הדרך לעשות זאת היא להסתיר את רוב הגנום חוץ מהחלקים שרלוונטיים לפעילות התא.

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

[מה אפשר לבדוק?](#)

נוכל לבחון לאן TF נקשרים בתאים שונים. דבר נוסף הוא לבחון איפה אין קישורים (ניסוי פופולארי לאחרונה הוא ATAC-Seq שבגדול מפעיל אנזים שחותך DNA ומסמן אותו – כשיש שני אירועי חיתוך קרובים נצליח לרצף, והאזורים האלה יקרו בדו"כ באזור פתוח). כך מקבלים תמונה של חוסר נוכחות.

אנחנו רק צריכים שהתחלת השעתוק תהיה נגישה – המכונה שעושה שעתוק באוקריוטים יודעת לעבור דרך אזור שארוז על ידי נוקלאוזום.

בגנום אוקריוטי הגנים בדרך כלל רחוקים אחד מהשני (לא קיים לחץ אבולוציוני לצמצם את אורך הגנום לעומת חד-תאים למשל). הגנים למרות שמקודדים לחלבון קצר יכולים להיות מאוד ארוכים. גם בתוך הגן וגם מחוצה לו יכול להיות ארוז, והאזורים הנגישים צריכים להיות אזורים ליד תחילת השעתוק (TSS – Transcription start sites או promotor) ולפעמים רחוקים יותר (נקראים regulator regions או enhancers) ובדרך זו משפיעים על הפעילות.

[בעיה חישובית](#)

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

האם אפשר להסביר אילו גנים יתבטאו בתא? אפשר לתאר את האזורים הנגישים ולהגדיר גן א' יבטא, גן ב' לא יבטא.. אם נבין את הבעיה הזו נהיה צעד אחד קדימה לקראת להסביר מדוע גנים מסוימים (למשל מחלות)

מתבטאות בשלב מסוים בהתפתחות וכדומה. לכאורה יש לנו את כל אבני הבניין לענות על השאלה, אבל ההתקדמות בתחום היא חלקית.

תחנות עגינה

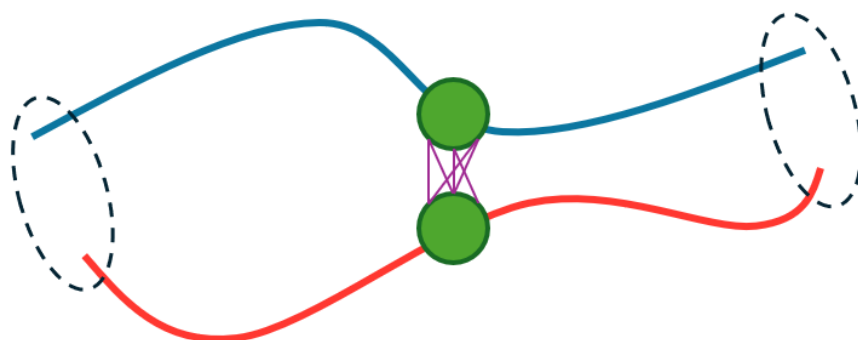
כשביוולוג מסמן אזור כ-promotor יש 3 תצפיות שיכולות לגרום לכך:

1. בעזרת מסווג כלשהו (שיקולים שלא נכנס אליהם)
2. נמצא התחלת שעתוק של RNA לכן מתחיל שם שעתוק
3. לקחנו חתיכה של ה-DNA ושמנו אותה לפני גן מדווח (reporter) שאין לו פרומוטור ופתאום הופיע ביטוי לגן.

בעזרת גנטיקה (שינו באופן טבעי/מלאכותי את ה-DNA) אנשים גילו שבעת שינוי מספר בסיסים משפיע על גן שממוקם רחוק מאוד. כשהוציאו את הרצף הזה גילו שהגן אכן לא התבטא. מה שאנחנו כן יודעים היום זה שיש הרבה אתרים שהם enhancers. מכל מיני ניסויים (צ'יפ, סימון חלבון בעזרת מיקרוסקופיה) אנחנו יודעים ש-enhancers עובדים על ידי זה שהם מתקרבים לפרומוטור שהם פועלים עליו. זה מעלה שאלות – איך הוא יודע למי להתחבר? מה גורם לו לזהות אותו?

צריך שחלבונים מסוימים יקשרו ל-enhancers ואחרים יקשרו לאזורי TSS Promotor. זה לא עונה על השאלה איך ה-enhancer יודע לאיזה גן להיקשר. התשובה הקצרה – יש שילוב של חלבונים שמזהים את האתרים ועם נטייה בסיסית להתחבר, והגנום ארוז בצורה תלת-ממדית מעניינת, יש אתרי עגינה ספציפיים שמפוזרים בגנום, אבל בחלקם הם צפופים, ויש חלבונים שקושרים בין אתרי עגינה שונים. הלולאות האלו גורמות למצב שאתרי עגינה שונים רחוקים בגנום אבל קרובים פיזית, וכן יש מכונות שתופסות DNA ומנסות להעביר אותו דרך מעין טבעת. הטבעת תתקע בנקודות העגינה שתיארנו ואז או שהטבעת תישאר (ועוד אחת תגיע, והטבעות יערמו – נכוץ את כל החוט ביחד, נפוץ בתהליך חלוקת תא) או שהיא פשוט תיפול.

משפחת שיטות ניסוייות (ש-Hi-C אחת מהן) לוקחות את הגנום ומפעילות חומר משמר (פורמלדהיד) בתור דבק שגורם לחלבונים להידבק אחד לשני. זה יוצר קשרים בין חלבונים – תהליך שנקרא קיבוע, ואז שוברים את הגנום. כך נקבל קטעים שהיו בעבר חתיכת DNA אחת, רק שעכשיו ייצמדו אליהן חתיכות גן שהיו בעבר רחוקות. הקסם הוא לבצע ריאקציה שמחברת בין הקצוות (באזור המקווקו):



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



ואתו אפשר לרצף. זו עדות לכך ששני הרצפים האלו היו קרובים. עכשיו אפשר להתחיל למלא מפה של קואורדינטות בגנום מול קואורדינטות בגנום ולספור כמה פעמים ראינו דברים שמתחילים במקום אחד ונגמרים באחר. מטבע הדברים רוב מה שנקבל יהיה על האלכסון של הטבלה, אבל ליד האלכסון נקבל אזורים צפופים יותר שמעידים שיש הפרדות פיזיות כמו שתיארנו.

בעיה אחת שעולה היא שכדי לאכלס מטריצה כזאת נדרשים למלא את מספר המקומות בגנום **בריבוע** – צריך לרצף המון כדי לוודא שאירועים לא קרו באקראי.

מסקנות שידועות היום

1. הרבה מנקודות העגינה שמורות בכל התאים והרבה מהאזורים של הלולאות חוזרים על עצמם בהרבה תאים עם הבדלים קטנים בין תאים שונים.
 2. יש enhancers שהתפקיד שלהם הוא לא להפעיל את הגן ליד אלא לגייס מכונות שפותחות את הלולאה או אורזות אותו וקוברות אותו. כך אפשר לכבות חתיכות שלמות בגנום.
- אותו enhancer יכול לפעול על כמה גנים, אותו גן יש כמה enhancers שיכולים לפעול עליו. לאבולוציה קל לעשות דברים כשהמערכת לא מבוקרת (פתאום גנים חדשים יכולים לקבל ביטוי).

סימונים כימיים על החלבון

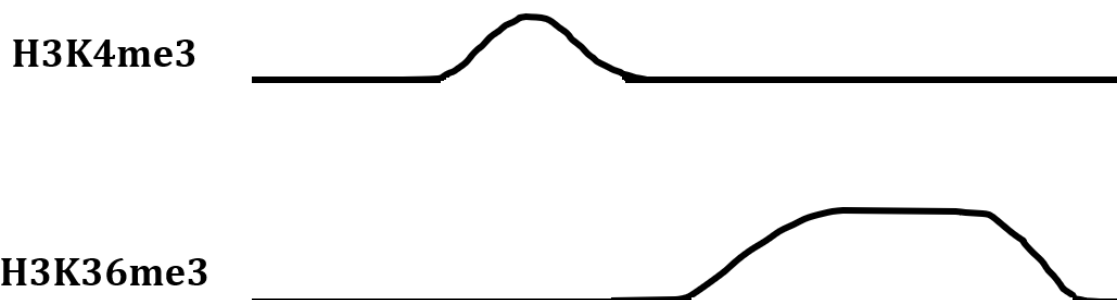
לפני 150 שנה גילו שעל החלבונים שסביבם ה-DNA ארוז אפשר לסמן סימונים כימיים. הסימונים מאפשרים לסמן חלבונים עם מגוון סימונים כימיים (פחות קריטי לדיון כרגע). מבחינה כימית מדובר בשינויים אחרי התרגום בצורה של פוספורילציה, מתילציה וכו' שמתקבלת בעמדות ספציפיות של החלבון.

כשגילו את זה, גילו גם אנזימים שמסמנים את החלבונים בצורה הזו. יש חלבון שהתאים מייצרים שבסיטואציות מסוימות מבצעים את ההוספה הזו. בהמשך גילו שיש חלבונים שנקשרים ל-DNA רק אם הוא עטוף בצורה מסוימת. יש דרך לכתוב ויש דרך לקרוא - אז מדובר בזיכרון זמני שמסמן את ה-DNA. זה הוביל למאמרי דעה חשובים שיש קוד נוסף ב-DNA - "קוד גנטי נוסף" מעבר לקוד ה-DNA (אפי-גנטי).

מה האינפורמציה השמורה? הדרך שאנשים ניגשו לבעיה היא בעזרת מיפוי איפה רואים את הסימנים השונים בגנום בשלבים שונים. איך ממפים? לתפוס את המקומות בגנום שמחזיקים רק את ההיסטון הרלוונטי בעזרת הנוגדן המתאים.

ביצוע מיפוי

היום אנחנו יודעים לבצע מיפויים כאלה בעזרת נוגדנים נגד הסימונים. בגדול, אנשים התחילו לשים לב שכשהם עושים ניסויים כאלה למשל בעזרת H3K4me3 וקיבל peak במקום מסוים, ואז עם H3K6me3 קיבל אובזרוציה אחרת.



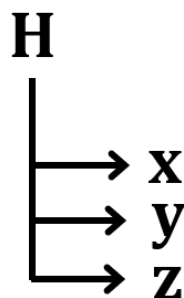
אם יש גן שאנחנו יודעים שהוא פעיל אז בדרך כלל נראה peak אחרי החץ שמסמן את האזור שבו מתחיל השעתוק, כלומר אפשר לזהות איפה יש גנים פעילים בחלבון רק על ידי מיפוי הפיקים.

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

מספר שנים אחרי זה יצא פרויקט שמטרתו למצוא את כל המקומות שמעניינים בגנום. אחת הטכנולוגיות הייתה בעזרת צ'יפ, וסימון גנים פעילים ולא פעילים. ביצעו סימונים במספר סוגי תאים וכך חתכו את הגנום במקומות המעניינים. האובזרוציות של peak מסוג אחד ואחריו peak מסוג אחר מובילים לגן. האם אפשר לבצע מיפוי בצורה יותר אוטומטית?

ChromHMM

מדובר ב-HMM עם כרומטין - עשו דברים מעט שונים ממה שדיברנו עליו בכיתה. ראינו שמצב חבוי H מוביל לתצפית x וכך הלאה. התצפית כאן במקרה שלנו היא רשימה של פרמטרים. באלגוריתם שהם פיתחו הם דיברו על מספר תצפיות:



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

ציר הזמן ובינאריזציה

מה ציר הזמן כאן? לגנום אין בדיוק חתיכות לפיהן הוא נשבר. הם החליטו לשבור את הגנום בצורה פשוטה של יצירת חתך באורך 200 בסיסים בכל פעם. אם בתוך ה-200 יש peak מספיק גבוה האחת התצפיות אבל בשני לא, היה צריך לבצע בינאריזציה (להחליט איפה 0 ואיפה 1) - כלומר רצף מאוד עשיר לרצף מעט קצת יותר עם 0 ו-1, בערך 15 מיליון מקומות.

בשלב זה היה צריך ללמוד את המצבים החבויים והמעבר ביניהם. אפשר לקחת את כל הנתונים ולהתעלם מאיפה הם נמצאים ולהסתכל רק על הקואורדינטות - האם בשניהם נמצא 1 או 0, זה ייתן לנו ניחוש התחלתי על המצבים החבויים. בהמשך השתמשו בזה בתור input לאלגוריתם EM שלמדנו בכיתה.

Hidden States והשימוש למחקר בביולוגיה

במודל היה משחק כמה hidden states להכניס וכדומה, אבל בגדול למדו שכש-HMM רץ על הגנום הוא עושה אנוטציה ואומר באיזה מצב אנחנו בכל מקום בגנום. השאלה היא אם אפשר לחבר את זה לביולוגיה? או שמדובר ברעש?

אפשר להסתכל על המקומות שבהם אנחנו חושבים שאנחנו יודעים מה קורה, למשל התחלות של הגנים שאנחנו בן מכירים. באופן מפתיע (או לא) ה-*hidden state* שהם למדו היה התחלה של גן פעיל והתחלה של גן פעיל נמוך או מוכן לפעילות, וגם התחלה של גן מושתק - כלומר מלמעלה מ-10 מצבים חבויים הגיעו לכ-4. טבלת המעברים הייתה שאם נלך משמאל לימין אם היינו בהתחלת גן סביר שנכנס לגוף הגן (או מעבר הפוך). הבעיה הייתה שכאשר הוא היה באזור התחלה של גן המודל לא זכר האם הוא כבר ראה גן (מאיזה כיוון הגיע). כשהולכים לסוגי תאים אחרים ומריצים את אותו מודל, מקבלים אנוטציות אחרות (כי דברים שונים מתבטאים בתאים שונים). לוגיקה של מתי גנים מתבטאים שונה בין תא לתא, אבל שיטת הסימון זהה בין תאים שונים. המטרה הייתה לתת ארכיטקטורה של HMM בעזרת שימוש אמיתי בעולם. היום השימוש של ChromHMM משמש גם אזורי בקרה בגנום ואילו אזורי בקרה פעילים. יש סימונים אחרים שביצעו למשל ל-*enhancers*, לאזורים שעוברים השתקה, אזורי חיבור ל-TF וכדומה. הכל נעשה בעזרת HMM.

למה התכונה המרקובית מסייעת כאן? ללא HMM המסווג היה מצליח לא רע, אבל האינפורמציה שהוא מקבל מ-HMM זה הסתכלות רחבה יותר על הדאטה וזיהוי שחלק מהמקומות יכולים להיות רעש, אבל בהינתן אינפורמציה נוספת מהסביבה אנחנו יכולים להסיק שהמקום עליו אנחנו מסתכלים כנראה לגיטימי.

איפה זה נופל?

בעיה אחת היא שלפעמים יש תכונות ממקטעים רחוקים יותר, ל-HMM קשה עם זה כי אם נרשה המון *hidden states* כדי להתייחס למצבים רחוקים יותר, אנחנו עלולים להגיע למצב של *overfit*. אם צריך לקלוט מספר אירועים מהעבר, המצב מתחיל להסתבך (הקבלה לחישוביות - האוטומט לא יכול להיות סופי).