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: פרופ' ניר פרידמן

תוכן עניינים

שבוע 1 – הרצאה 1 4

הקדמה – ביולוגיה מולקולרית 4

הדוגמה המרכזית של הביולוגיה המולקולרית – Central Dogma of Molecular Biology 5

קידוד לחומצות אמינו 6

פיתוח אלגוריתם לבעיה ביולוגית 7

שבוע 1 – הרצאה 2 8

עימוד רצפים – Sequence Alignment 8

תכנון דינמי – Dynamic Programming 9

עימוד במקום לינארי – Linear Space Alignment 11

עימוד מקומי – Local Alignment 13

שבוע 2 – הרצאה 3 14

Probabilistic Models and Decisions 14

Two Hypotheses 14

למת ניימן-פירסון (1933) 17

One Hypothesis 18

שבוע 2 – הרצאה 4 20

למידה על הסתברות מנתונים 20

דרך שיערוך בייסיאנית 21

שימוש בהסתברות בעימוד רצפים 22

שבוע 3 – הרצאה 5 (הרצאת אורח) 24

הקדמה 24

מידול אינטגרטיבי של מערכות ביולוגיות דינמיות 25

שלבים ביצירת מודל 26

שימוש במודל 27

שבוע 3 – הרצאה 6 28

היוריסטיקות לעימוד רצפים 28

שרשראות מרקוב 29

שבוע 4 – הרצאה 7 32

מודלים מרקוביים חבויים – Hidden Markov Model 32

שבוע 4 – הרצאה 8 37

בעיית השחזור – Maximum Probability Reconstruction 37

שבוע 5 – הרצאה 9 42

אלגוריתם Max-Max (MM) 42

אלגוריתם Exp-Max (EM) 43

שבוע 5 – הרצאה 10 47
הגינום וכרומטידים 47
שבוע 6 – הרצאה 11 54
סימונים כימיים על החלבון 54
ChromHMM 55
שבוע 7 – הרצאה 12 57
Phylogenetic Trees 57
אלגוריתם UPGMA 62
אלגוריתם Neighbour Joining 65
שבוע 7 – הרצאה 13 67
חישובי מרחקים 69
שבוע 8 – הרצאה 14 (הרצאת אורח) 74
הקדמה והגדרת הגישה 74
De-Bruijn Graph 74
Metagenomic Assembly 78
שבוע 8 – הרצאה 15 (הרצאת אורח) 81
הקדמה 81
נקודת פתיחה 81
הגדרת האנרגיה 82
בעיות לא קמורות 84
חזרה למערכת הביולוגית 85
סיכום 87
שבוע 9 – הרצאה 16 89
מודלים של זמן בעצים פילוגנטיים 89
Kimura Model 89
Stationary Probability 92
Multiple Sequence Alignment - MSA 94
שבוע 9 – הרצאה 17 (הרצאת אורח) 97
מבנים מולקולריים 97
Protein Structure Prediction 99
שבוע 10 – הרצאה 18 113
סרטן וגנטיקה 113
זיהוי ודיאגנוזה 116
שבוע 10 – הרצאה 19 (הרצאת אורח) 118
אבסטרקציה וספציפיקציה 118

סוגים שונים של שיטות single-cell 120
טיפול ברקמות והשהיית תאים 120
Droplet based scRNAseq 121
Cell Filtering 123
Doublets / Multiplets 124
נורמליזציה והורדת ממד 125
tSNE\UMAP 126
מידול הבדלים ב-Gene Expression 127
שבוע 11 – הרצאה 20 (הרצאת אורח)..... 128
ספקטרומטריית מסות 128
בעיות במדידה 130
מה אפשר למדוד רק ב-Mass-spect? 132

שבוע 1 – הרצאה 1

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	$10^{-5}\ m$	Cells	
	$10^{-6}\ m$	Organelles	Mitochondria
	$10^{-7}\ m$	Macro-molecules	DNA/Protein
	$10^{-9}\ m$	Small molecules	Amino Acid
	$10^{-10}\ m$	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:

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

נוקלאוטידים של DNA ושל RNA מעט שונים מבחינה כימית.

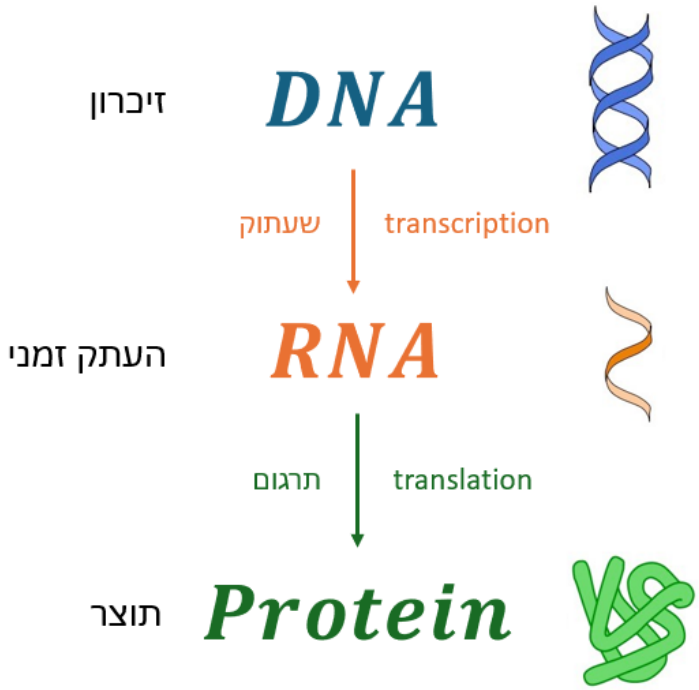
DNA

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:

$$\begin{array}{ccccccccccc} 5' & & & & & & & & 3' & & \\ A & \rightarrow & C & \rightarrow & C & \rightarrow & T & \rightarrow & A & \rightarrow & \cdots \\ \Updownarrow & & \Updownarrow & & \Updownarrow & & \Updownarrow & & \Updownarrow & & \\ T & \leftarrow & G & \leftarrow & G & \leftarrow & A & \leftarrow & T & \leftarrow & \cdots \end{array}$$

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: $A \Leftrightarrow \boldsymbol{U}, C \Leftrightarrow G$. 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:

64 different options

AAA	Lys
$\vdots$	$\vdots$
TTT	$\vdots$

20 amino acids

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

$$\cdots \text{ CA } \underbrace{\text{ATG}}_{start} \text{ TAA CTG GTT } \underbrace{\text{UGA}}_{stop} \text{ CC } \cdots \underbrace{\text{ATG}}_{start} \cdots$$

הרצף ATG יכול להופיע גם באמצע רצף ולקודד למתיונין, 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-3×10^9 אותיות.

פיתוח אלגוריתם לבעיה ביולוגית

לא נתמקד רק באלגוריתם, אלא גם ב**למה** פיתחנו אותו, ו**על 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 מישהו 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.

שבוע 1 – הרצאה 2

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: $s = \text{AACT} \quad t = \text{AGT}$

<u>עימוד</u> – 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 ונזיז אותם כך שיהיו דומים? לשם כך נגדיר את **חוקי המשחק:**

1) מותר לבצע הזזות 2) מותר להכניס רווחים –

נגדיר א"ב מורחב: $s^*, t^* = \{\text{A,T,C,G,} -\}$ ונמצא שני רצפים כך שמתקיים $|s^*| = |t^*|$ וכך שמתקיים $remove - (t^*) = t,\ remove - (s^*) = s$ כאשר הפונקציה $remove -$ 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 indel[1] כי מופיעה אות מול רווח.
בכתום – מופיעה אי-התאמה/mismatch כי אין שוויון באותיות.

<u>הערה</u>: 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 **in**sert ושל **del**etion, מאחר שעבור הרצף העליון עלינו לבצע $delete(\text{T})$ ועבור הרצף השני עלינו לבצע $insert(\text{T})$.

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:

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

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.

<u>הערה</u>: 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.

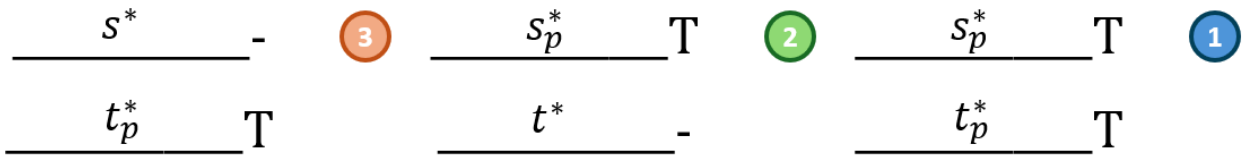
תכנון דינמי – Dynamic Programming

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:

$$\underbrace{\quad * \quad}_{t_{prefix}} \mathrm{T} \qquad \underbrace{\quad * \quad}_{s_{prefix}} \mathrm{T}$$

יש למעשה שלוש אפשרויות למקם את הרצפים זה מול זה:



ולכן:

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

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

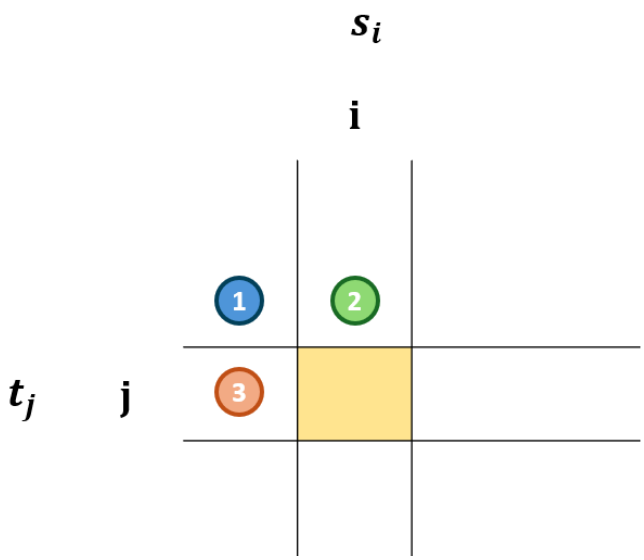
<u>מקרי בסיס</u>: $V[0,0] = 0 \quad V[*,-1] = \infty \quad V[-1,*] = \infty$

<u>הגדרת הרקורסיה</u>:

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

<u>מילוי הטבלה</u>:

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:

s		t			
			A	G	T
		0	1	2	3
	0	0	-2	-4	-6
A	1	-2	1	-1	-3
A	2	-4	-1	0	-2
C	3	-6	-3	-2	-1
T	4	-8	-5	-4	$\boxed{-1}$

<u>סיבוכיות מקום</u>: $O(n \cdot m)$ כמספר התאים בטבלה.

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

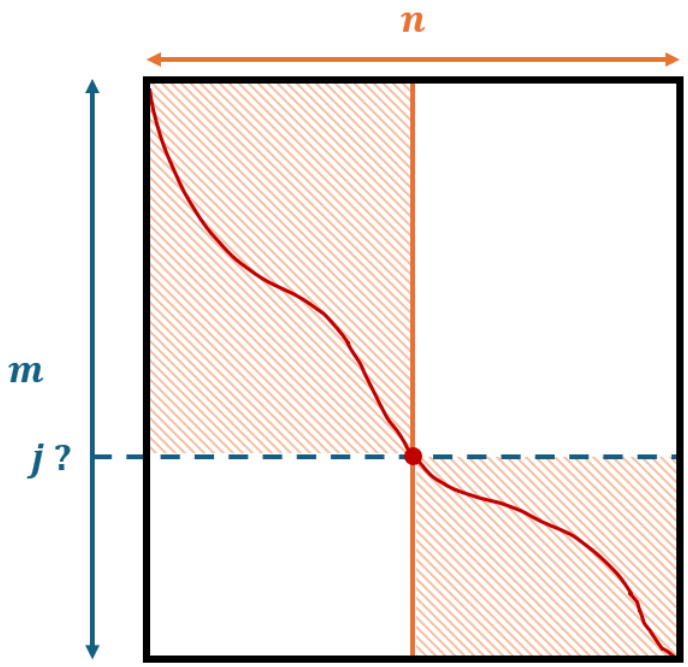
הזמן שקיבלנו הוא פולינומי, אבל אם מבצעים עימוד גנים של אדם מול גנים של עכבר, עדיין מדובר בסדרי גודל עצומים ומתחילה גם בעיית מקום ומוגבלות hardware. כיצד נתמודד? כדי לחשב את הציון בלבד, אין צורך במטריצה כולה, אבל עבור העימוד הסופי עלינו לשמור יותר מידע.

עימוד במקום לינארי – Linear Space Alignment

<u>פתרון נאיבי</u>: 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 הרישות בתוספת העימוד הכי טוב של הסיפות ייתן את העימוד הכולל הטוב ביותר.

<u>רישות</u>: $V[i:j] = \max(s[1:i], t[1:j])$

<u>סיפות</u>: $U[i:j] = \max(s[i+1:n], t[j+1:m])$

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 $m \cdot \frac{n}{4}$, ואז על $m \cdot \frac{n}{8}$ 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:

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

<u>מילוי טבלה</u>:

	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).

שבוע 2 – הרצאה 3

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).

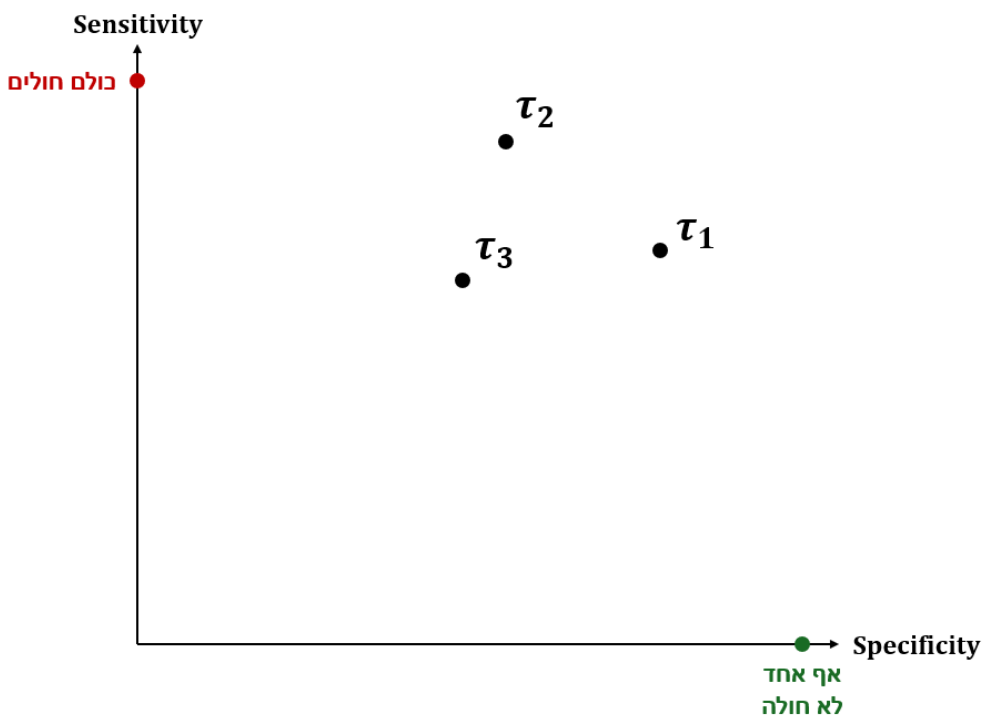
<u>העלינו כמה יחסים מעניינים</u>:

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

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

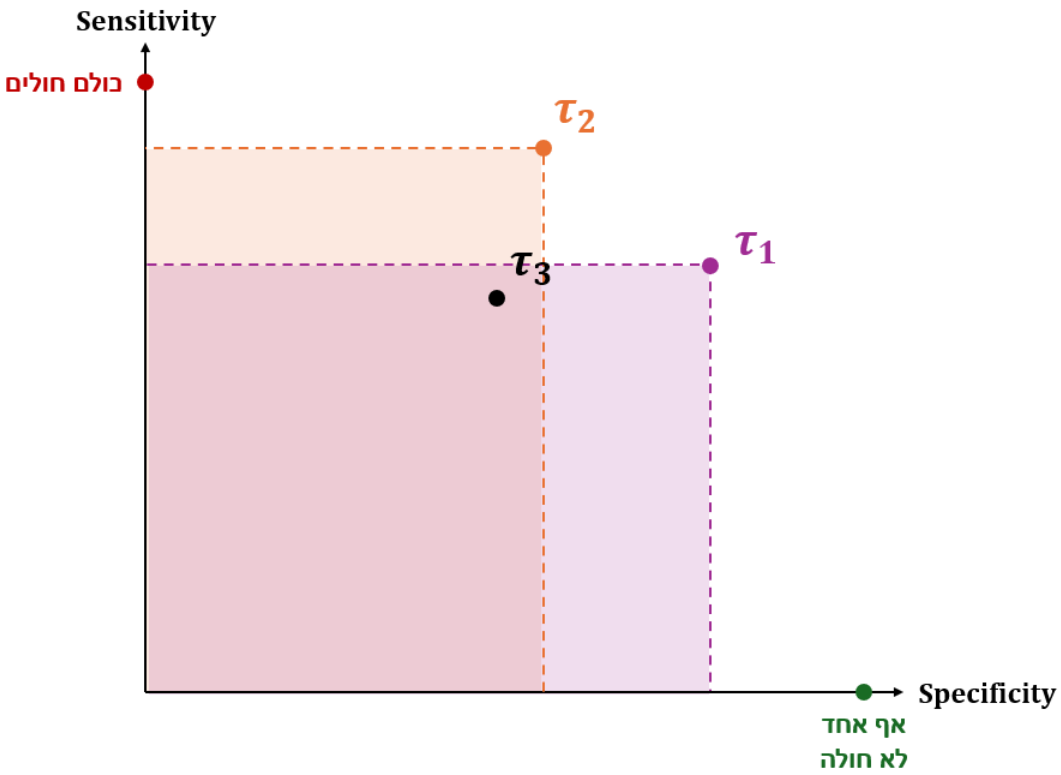
לפירוט המלא: *Sensitivity and specificity - Wikipedia*

אפשר להמחיש גם בגרף:

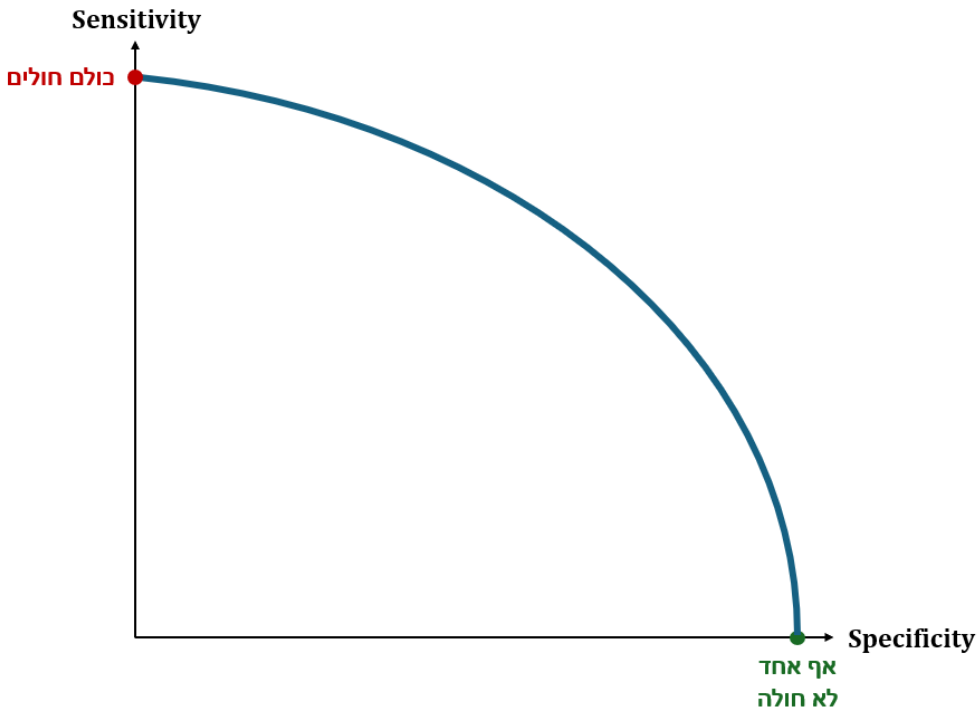


כל חוק החלטה הוא נקודה במרחב הזה (עם ביצועים בשני הממדים). האם נעדיף את τ_1 או את τ_2? תלוי במה שנדרש מאיתנו. לעומת זאת, לעולם לא נעדיף את τ_3 על פני אחד מהם, כי הוא גרוע מהם בשני הפרמטרים. בשלב הזה, זה שיקול של המשתמש מה חשוב יותר.

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:



למת ניימן-פירסון (1933)

אם נסתכל על חוק החלטה כלשהו:

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

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

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

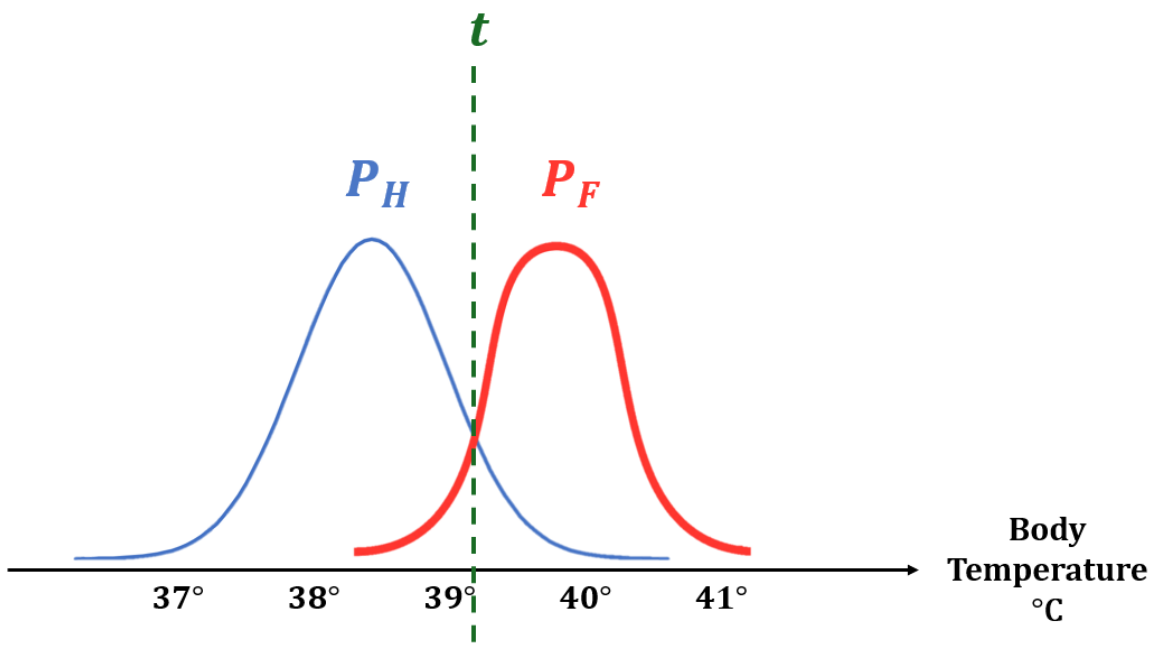
כלומר, החוק מקסימלי אם"ם קיים סף t שעבורו זה מתנהג כמו יחס הנראות. במילים אחרות, אם מציבים את חוקי המשחק (2 הסתברויות) כל החוקים שצריך לשקול בעולם יהיו מהצורה הזו (אם לא, נהיה בתוך האזור שמתחת לקו הכחול שראינו למעלה).

גישה בייזיאנית אומרת שבהינתן תצפית אפשר להכריע איך פרט מתנהג:

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

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

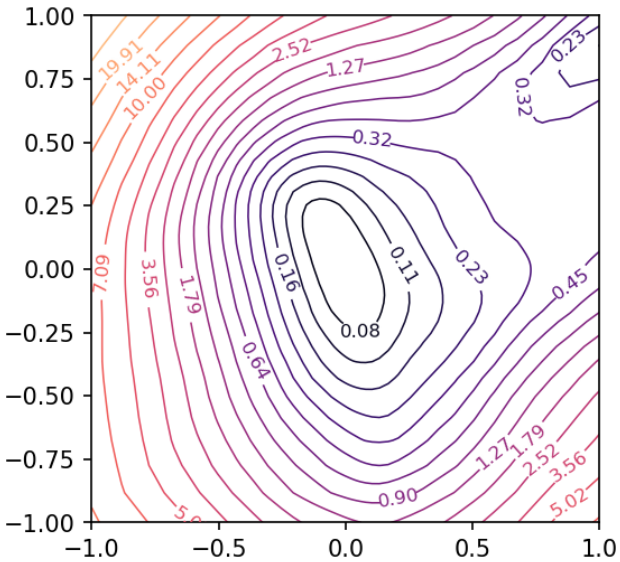
$$\mathbb{P}(x|+) = P_F(x) \quad \mathbb{P}(x|-) = P_H(x) \quad \mathbb{P}(+), \mathbb{P}(-) = \text{Prio (ידע מקדים)}$$



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 $P_H(temp > 38.5°) = \overbrace{0.0001}^{P\ value}$.
נוח לדבר במונחים של הסתברות המשלים. העניין מסתבך כאשר הבדיקה שלנו היא ביותר 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 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:

- $P_{H_0} \equiv P_0$ וגם $P_{H_1} \equiv P_1$
- p_0, p_1 תהיה דגימה של אות
- s_i ו-t_i האות ה-i ברצף.

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 כן 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:

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

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

ולכן נגדיר:

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

שבוע 2 – הרצאה 4

למידה על הסתברות מנתונים

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

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

המימד ה-n-י 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: $x[i] \in \{H, T\}$ $\quad x[1], \dots, x[N]$.

נרצה לשערך את הסתברות המאורע H (זהה לשערוך 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 למקסימום את הפונקציה, כלומר: $\hat{\theta} = \arg\max_\theta L(\theta) = \arg\max_\theta \ell(\theta)$

נסמן $MLE := Maximum\ Likelihood\ Estimation$. 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 <u>סטטיסט</u> – 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 שיערוך 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 השיערוך הזאת תיתן את אותו הערך גם לדאטה מהצורה $\{T, T, \dots, T\}$ כך 100 פעמים – ברור לנו ששתי קבוצות הדאטה שונות זו מזו.

<u>השיטה הבייסיאנית</u>

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

שימוש בהסתברות בעימוד רצפים

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)

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

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

חקר הפונקציות מסתבך – כי מדובר עכשיו בפונקציה רב-ממדית. צריך שהנגזרת החלקית בכל כיוון תהיה 0:

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

עלינו לפתור את הבעיה $\max_\theta f\left(\vec{\theta}\right) \quad s.t\ g\left(\vec{\theta}\right)=0$ והדרך לזו היא באמצעות כופלי לגראנז':

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

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

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

כשעושים 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:

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

בהינתן 20 רצפים וההנחה שכל אות נדגמה באופן בלתי תלוי. לשערך נראות מקסימלית משמעותה לספור כמה פעמים כל אות הופיעה. אילו רצפים נכניס 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 רווה, 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.

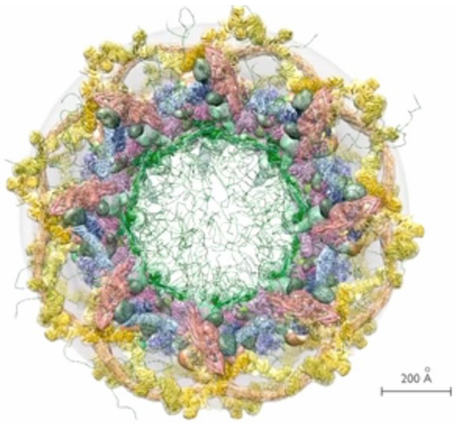
דיון – מהם חיים?

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.



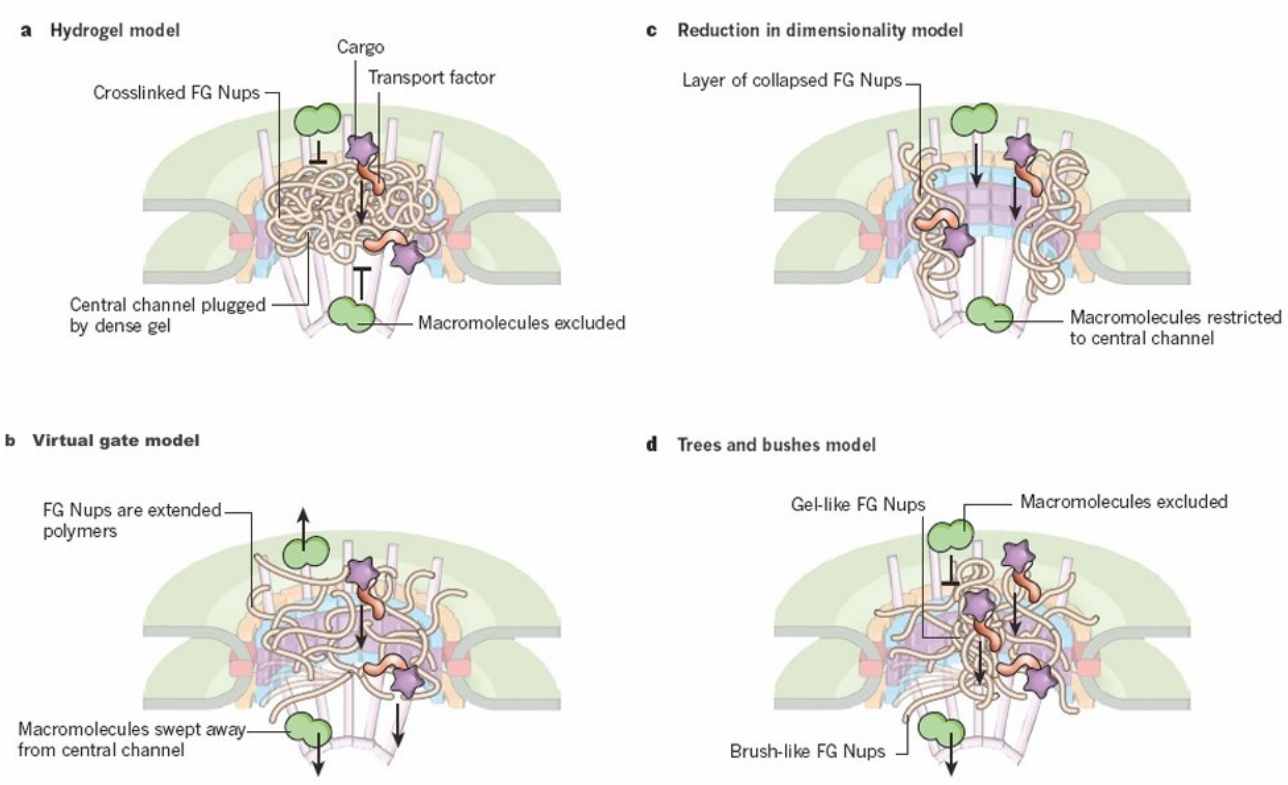
למה צריך מודלים והדמיות? לחזות להסביר תכונות, להכריע בין היפותזות, לשלוט בתכונות המערכת, לזהות פערי מידע ולאתר מידע חסר.

מידול אינטגרטיבי של מערכות ביולוגיות דינמיות

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

נעבוד לאורך ההרצאה עם דוגמה – מודל של תנועה מולקולרית דרך חריריות הגרעין. החריריות 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.

שלבים ביצירת מודל

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 שייוצגו כבדור 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.

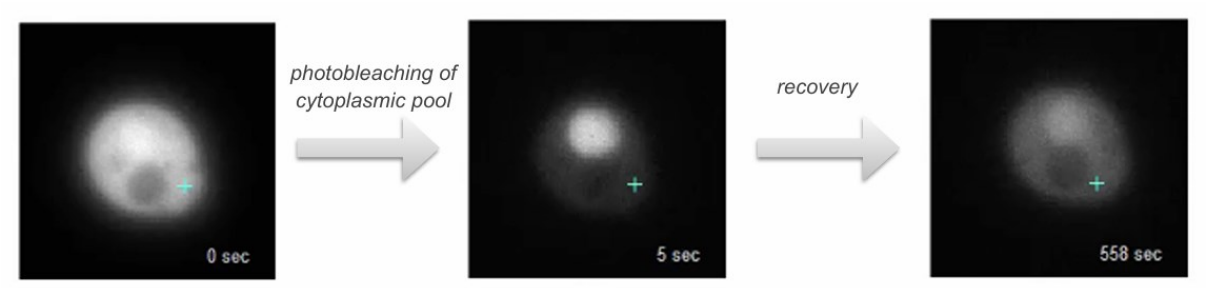
4. **אישוש נכונות ואיכות המודל והפקת תובנות חדשות על המערכת –**

באמצעות מידע ב"ת תבוצע ואלידציה של התוצאות. בשלב זה ננסה להבין אילו אירועים לא טריוויאלים 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.

שימוש במודל

מה המודל טוען לגבי הסתננות של מולקולות לפי גודלן?

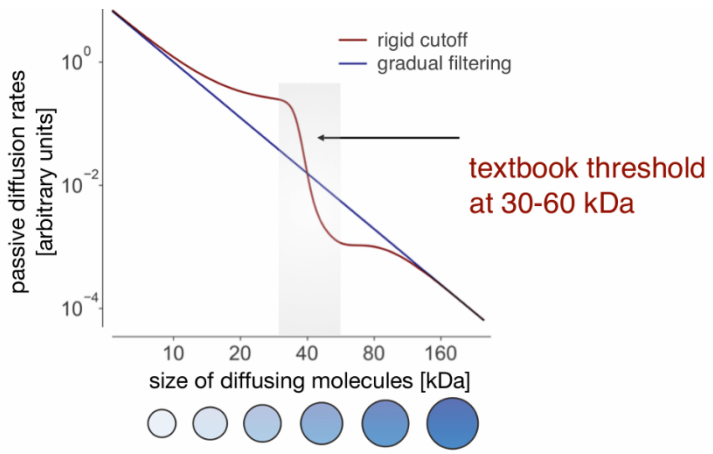
כל מולקולות ה-GFP בציטופלסמה עברו Photo Bleaching ולכן תוך 5 שניות כל הציטופלסמה הפסיקה לזהור:



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 ואלידציה.

שבוע 3 – הרצאה 6

היוריסטיקות לעימוד רצפים

לעתים הפתרון המוכח הוא יקר, ראינו שאפשר להגיע לזמנים קוואדרטיים 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**.

<u>שרשרת מרקוב</u> – בדרך כלל מוגדרת על ידי תכונת מרקוב:

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

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

<u>שרשרת מרקוב אחידה</u> – מכילה את התכונה הבאה:

$$\forall \boldsymbol{i}, \boldsymbol{j} \quad \mathbb{P}(\boldsymbol{x_{i+1}} = \boldsymbol{a} \mid \boldsymbol{x_i} = \boldsymbol{b}) = \mathbb{P}(\boldsymbol{x_{j+1}} \mid \boldsymbol{x_j} = \boldsymbol{b}) \overset{\star}{=} \boldsymbol{\tau}[\boldsymbol{a}, \boldsymbol{b}]$$

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:

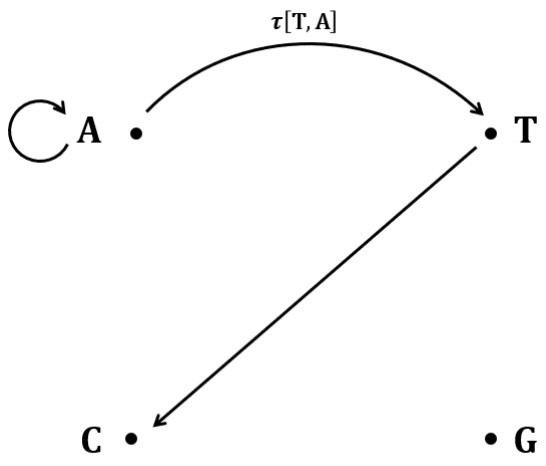
$$\mathbb{P}(x_1, \dots, x_n) \overset{\text{שרשרת}}{=} \mathbb{P}(x_1) \cdot \mathbb{P}(x_2 \mid x_1) \cdot \dots \cdot \mathbb{P}(x_n \mid x_1, \dots, x_{n-1}) \overset{\text{Markov}}{=}$$

$$= \mathbb{P}(x_1) \cdot \prod_{i=1}^{n-1} \mathbb{P}(x_{i+1} \mid x_i) \overset{\text{התפלגות אחידה}}{\overset{\star\ \text{שינוי נוטציות}}{=}} \mathbb{P}_0(x_1) \cdot \prod_{i=1} \tau[x_{i+1}, x_i]$$

מטריצת זוגות אותיות שמגדירה את הסיכוי לאות אחת בהינתן אחרת.

State Diagram

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**.

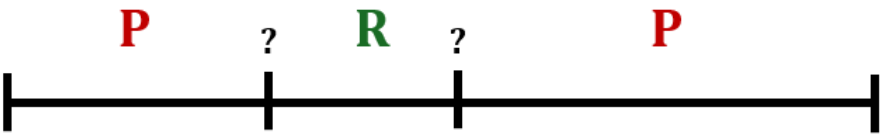
איך היינו בונים מטריצת מעבר לפי האם האזור עשיר/עני ב-CpG? נגדיר שתי מטריצות שיבדלו ביניהן בתא המסומן:

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\R ואת המעברים ביניהם אפשר לשייך לאורך אופייני של Strech של אזור עשיר או עני.

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

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

עדיין יש סימני שאלה – איך משתמשים במודל כזה?

שבוע 4 – הרצאה 7

מודלים מרקוביים חבויים – Hidden Markov Model

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

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

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:

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

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

שימושים במודל:

1. בהינתן תצפיות, מה היה המצב בזמן i? $\mathbb{P}(H_i \mid x_1, \dots, x_n)$
2. מהו $\arg\max_{h_1, \dots, h_n} \mathbb{P}(H_1 = h_1, \dots, H_n = h_n \mid x_1, \dots, x_n)$? 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-$\star$, ולהתחיל להפעיל חוקים אלגבריים. אנחנו נציג דרך אחרת באמצעות תכנון דינאמי. לשם כך, נכתוב את הבעיה ללא הסכום האקספוננציאלי.

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

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

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

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:

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

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:

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

<u>ניקח דוגמה מספרית</u>:

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	$1/6$	$1/2$

		From	
	τ	F	L
To	***F***	0.99	0.05
	L	0.01	0.95

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 ♣:

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

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 $\star+\spadesuit=0.01397$.

מה הסיבוכיות? n כפול גודל ה-א"ב 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 נזין את המספרים במחשב, נקבל underflow.

<u>איך נטפל בבעיה</u>? נשים לב שבנוסחה יש $b=a\cdot\sum c_i$, ובמקרים כאלה נפוץ לשמור log representation כלומר $\tilde{x}=\log b$, ואז $\log b=\log\left(e^{\tilde{a}}\sum_i e^{\tilde{c}_i}\right)$. נבחין כי $\tilde{b}=\tilde{a}+\log\sum_i e^{c_i}$. 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 שיסכם ל-1.

אבל, לפעמים נרצה לדעת מה קרה קודם לכן:

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

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

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

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

<u>לסיכום</u>: $\mathbb{P}(H_i = s \mid x_1, \dots, x_n) = F_i[s] \cdot B_i[s]$.

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

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

נבחין שאת כל החלקים המסומנים באדום אפשר להוריד בזכות התכונות של שרשרת מרקוב. נשארנו עם:

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

ולכן אפשר לענות על $\boldsymbol{F}orward$ ב-$O(n)$ וגם על $\boldsymbol{B}ackward$ 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.

שבוע 4 – הרצאה 8

תזכורת וסיכום השיעור הקודם

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

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

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

בעיית השחזור – Maximum Probability Reconstruction

בעיה של מציאת ה-h-ים שגורמים לתצפיות שלנו להיות הכי סבירות:

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

<u>הערה</u>: אם **נתנה** בתצפיות, כלומר נמקסם את $\mathbb{P}(H_1=h_1,\dots,H_n=h_n\mid x_1,\dots,x_n)$ 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:

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

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

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

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**.

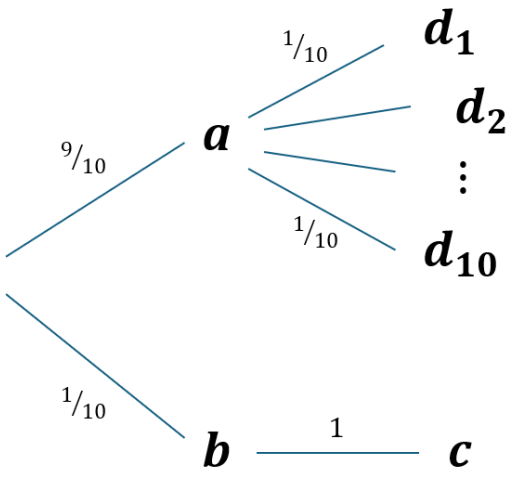
שאלת הבנה

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

מה אפשר להגיד על הערך $\mathbb{P}\left(H_i = \hat{h}_i \mid x_1, \dots, x_n\right)$ 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 $\hat{h}_i \overset{?}{=} \tilde{h}_i$? אינטואיטיבית היינו רוצים שזה יהיה המצב, אך זה לא כך. אחת הדרכים היא לייצר דוגמה מנוונת לחלוטין שעושה את זה. נבנה שרשרת מרקוב שבה השרשרת הכי סבירה אינה הערך הכי סביר:



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

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 $a \to c$ 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.

הבעיה הקלה – Full Data

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:

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

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

$$\vdots$$

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

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 חלבונים (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:

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

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

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

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

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

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

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?

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

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).

עכשיו 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:

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

הבעיה **הכתומה** זהה ללמידת קובייה:

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

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:

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

באופן דומה הבעיה **הסגולה**:

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

<u>סיכום</u>:

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

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

הבעיה הקשה

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

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

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

$$\vdots$$

גם כאן נתחיל מהנראות:

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

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.

שבוע 5 – הרצאה 9

תזכורת

אנחנו נמצאים במצב שבו הדאטה שלנו נראה בצורה הזו:

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

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 סטטיסטים ($N_{0,s}, N_{s,t}, N_{s,x}$), וגם ההפך. נתחיל עם ניחוש התחלתי של פרמטרים. הניחוש ההתחלתי משפיע על התוצאה הסופית, אבל נשאיר את הדיון על כך להמשך.

סקיצה של האלגוריתם:

```
Θ^1 ← Guess
For j = 1,...
      N^j =???
      Θ^(j+1) ← MLE [N^j]
```

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

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

אלגוריתם Max-Max (MM)

מה יהיה ה-Max Step באלגוריתם?

נגדיר: $\vec{H}^j \leftarrow \arg\max_H \mathbb{P}_{\Theta^j}\left(\vec{H}^j \,\middle|\, \text{Data}\right)$

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

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 Exp-Max (EM)

אנחנו נתמקד באלגוריתם הזה. נגדיר: $\vec{N}^j = \mathbb{E}\left[\vec{N} \,\middle|\, \vec{X}, \Theta^j\right]$. 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: $\sum_{\vec{H}} \mathbb{P}\left(H \,\middle|\, \vec{X}, \Theta^j\right) \cdot \vec{N}(H)$, 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 $N_{0,x}$

$$\mathbb{E}\left[N_{0,x} \,\middle|\, \vec{X}, \vec{\Theta}\right] \overset{\text{ליניאריות}}{\underset{\text{התוחלת}}{=}} \sum_m \mathbb{E}\left[1\{H_1[n] = s\} \,\middle|\, \vec{X}, \vec{\Theta}\right] \overset{\text{תוחלת של}}{\underset{\text{אינדיקטור בינארי}}{=}} \sum_m \mathbb{P}(H_1[m]$$
$$= s \,\middle|\, \vec{X}, \vec{\Theta}\Big) \overset{i.i.d}{=} \sum_m \mathbb{P}\left(H_1[m] = s \,\middle|\, \vec{X}[M], \vec{\Theta}\right) =$$

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:

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

חישוב עבור $N_{s,x}$

$$\mathbb{E}\left[N_{s,x} \,\middle|\, \vec{X}, \vec{\Theta}\right] \overset{\text{ליניאריות}}{\underset{\text{התוחלת}}{=}} \sum_m \sum_i \mathbb{E}\left[\{x_i[m] = x, H_i[m] = s\} \,\middle|\, \vec{X}, \vec{\Theta}\right]$$
$$= \sum_m \sum_i \mathbb{P}(x_i[m] = x, H_i[m] = s)$$

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

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

חישוב עבור $N_{s,t}$

$$\mathbb{E}\left[N_{s,t}\ \middle|\ \vec{X},\vec{\Theta}\right]\underset{\text{התוחלת}}{\overset{\text{ליניאריות}}{=}}\sum_{m}\sum_{i}^{n_m-1}\mathbb{E}\left(1\{H_i[m]=s,H_{i+1}[m]=t\}\ \middle|\ \vec{X},\vec{\Theta}\right)$$

$$=\sum_{m}\sum_{i}^{n_m-1}\mathbb{P}\left(H_i[m]=s,H_{i+1}[m]=t\ \middle|\ \vec{X}[m],\vec{\Theta}\right)$$

חזרה לסטטיסטים

כשדיברנו על סטטיסטים הייתה שאלה מה הערך של:

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

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

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

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

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

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

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

הביטויים מופיעים לפי הסדר.

פירוט האיטרציה

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

```
Θ^1 ← Guess
For j = 1,...
      Run Forward/Backward for all m
      Collect E[N] with F/B
      Θ^{j+1} ← MLE(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 הסטטיסטים במעבר). גם מבחינת מקום, נקבל סאב-לינארי בגודל האינפוט.

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

משפט

1. $\mathbb{P}\left(\vec{X}, \Theta^{j+1}\right) \geq \mathbb{P}\left(\vec{X}, \Theta^j\right)$
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 של הפונקציה $L\left(\vec{\Theta}\right) = \mathbb{P}\left(\vec{X}, \vec{\Theta}\right)$.

תנאי העצירה

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

$$\ell\left(\Theta^j\right) \xrightarrow{j \to \infty} \ell\left(\Theta^*\right)$$

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

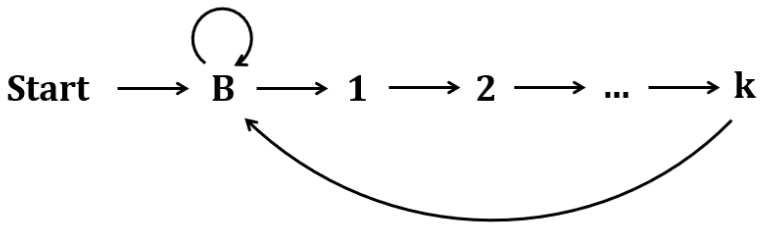
מבט מלמעלה

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 דפוסי 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

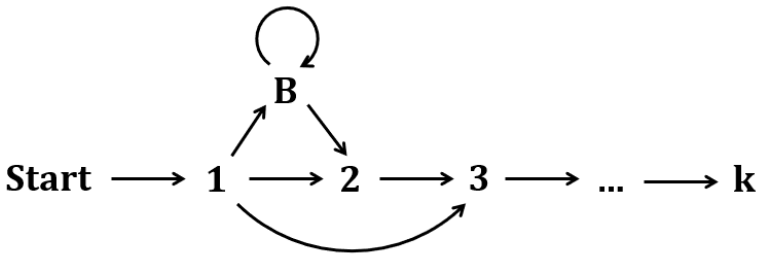
1. **מילה באורך k**



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 חלבונים כולל שינויים**



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

שבוע 5 – הרצאה 10

הגינום וכרומטידים

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

גינום – 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 אזורים 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.

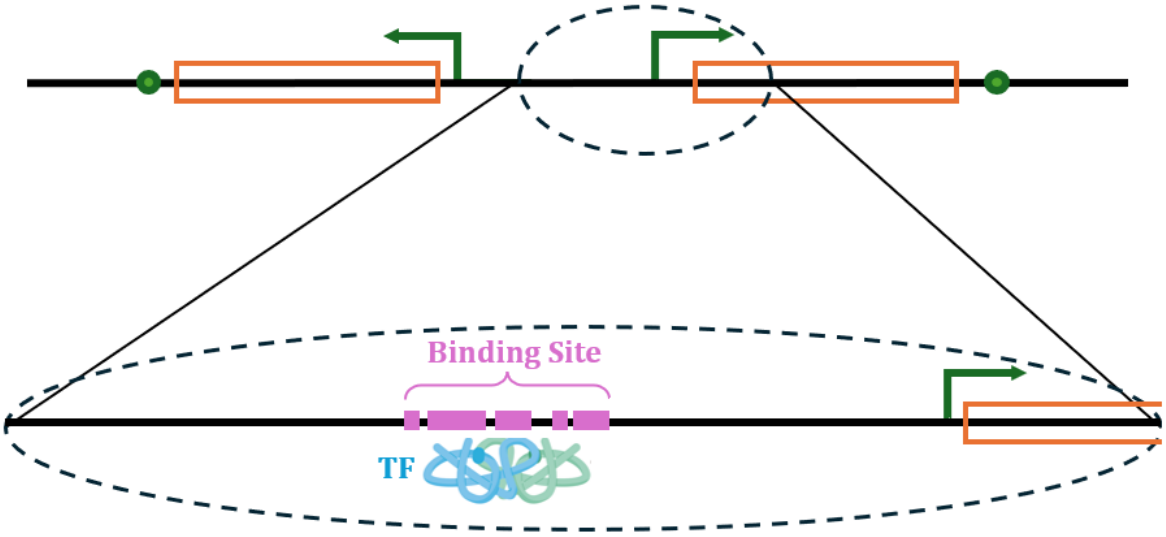
יש עוד אינפורמציה במהלך הרצף שמכוונת את התאים להשתמש ב-DNA בצורה ה"נכונה". השאלה היא מהי האינפורמציה הזו ואיך ניגשים אליה. אם נתחיל בשאלה שניצבת בפני יצור פשוט כמו חיידק, רוב ה-DNA שלו הם גנים והוא מצליח להשתמש בהם נכון בעזרת שימוש בעובדה שבאזור שלפני החץ הירוק יש מילה או כמה מילים שמורות שמופיעות בנקודה 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 Factors[2]. 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 שיודע 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 Precipitation[3]. 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 בקונבולוציה 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 חישובית שקשורה לכך היא Find Enriched Words. בהינתן אוסף מילים בטקסט שברוב הפסקאות מופיעה מילה אחת, לא ידוע איפה, ואנחנו

---

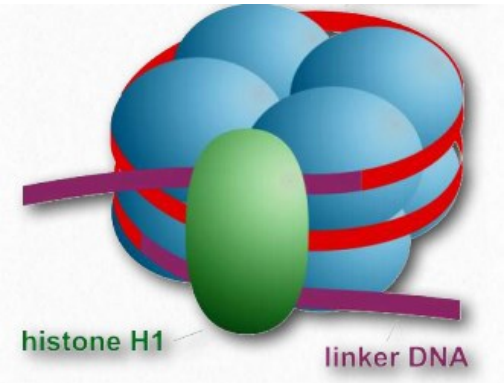
[3] Immuno – תלוי במערכת החיסון, Precipitation – השקעה.

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 היסוטנים 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. ברירת 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 מטרתם 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 בדר"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 הגינום 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' יבוטא, גן ב' לא יבוטא.. 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

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

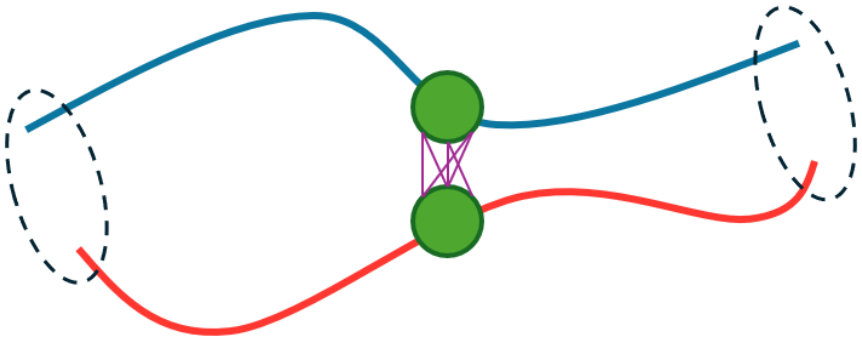
1. בעזרת מסווג כלשהו (שיקולים שלא נכנס אליהם)
2. נמצא התחלת שעתוק של RNA לכן מתחיל שם שעתוק
3. לקחנו חתיכה של ה-DNA ושמנו אותה לפני גן מדווח (reporter) שאין לו פרומוטור ופתאום הופיע ביטוי לגן.

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 – נכווץ את כל החוט ביחד, נפוץ בתהליך חלוקת תא) או שהיא פשוט תיפול.

Hi-C

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):

ואותו 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 בגינום **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).

שבוע 6 – הרצאה 11

סימונים כימיים על החלבון

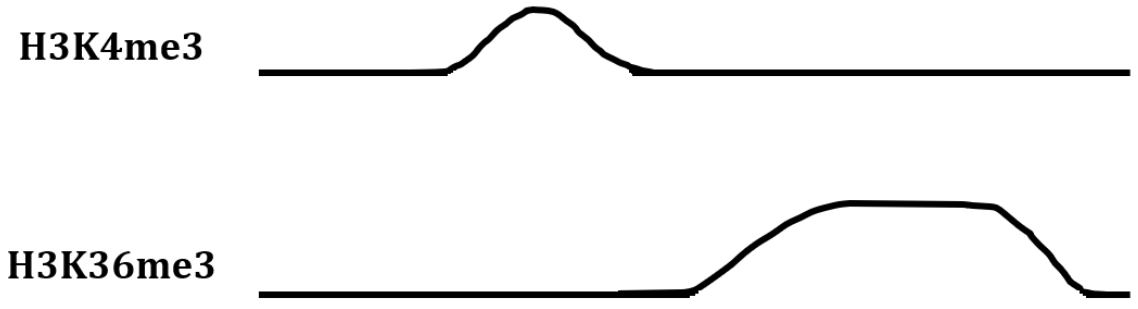
לפני 150 שנה גילו שעל החלבונים שסביבם ה-DNA ארוז אפשר לסמן סימונים כימיים. הסימונים מאפשרים לסמן חלבונים עם מגוון סימונים כימיים (פחות קריטי לדיון כרגע). מבחינה כימית מדובר בשינויים אחרי התרגום בצורה של פוספורילציה, מתילציה וכו' שמתקבלת בעמדות ספציפיות של החלבון.

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 H3K6me4 קיבל אובזרוציה אחרת.



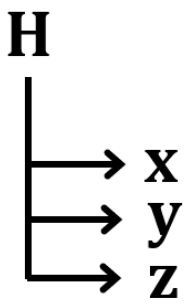
אם יש גן שאנחנו יודעים שהוא פעיל אז בדרך כלל נראה peak אחרי החץ שמסמן את האזור שבו מתחיל השעתוק, כלומר אפשר לזהות איפה יש גנים פעילים בחלבון רק על ידי מיפוי הפיקים.

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 באחת 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 בגינום. 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 כן 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?

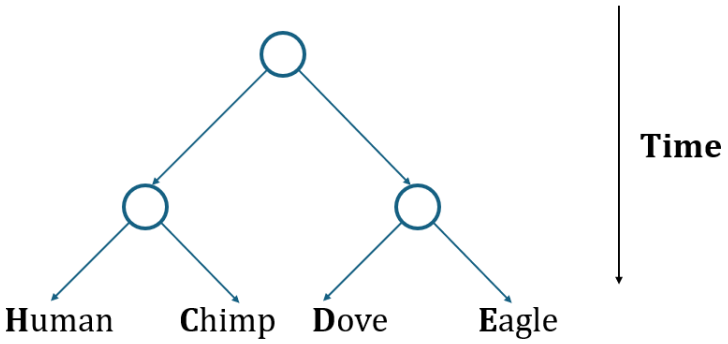
בעיה אחת היא שלפעמים יש תכונות ממקטעים רחוקים יותר, ל-HMM קשה עם זה כי אם נרשה המון hidden states כדי להתייחס למצבים רחוקים יותר, אנחנו עלולים להגיע למצב של overfit. אם צריך לקלוט מספר אירועים מהעבר, המצב מתחיל להסתבך (הקבלה לחישוביות - האוטומט לא יכול להיות סופי).

שבוע 7 – הרצאה 12

Phylogenetic Trees

להסתכל על העצים האבולוציונים שאנחנו יכולים לסדר מינים שונים בהם. כפי שכבר דיברנו בנושא עימודים, אנחנו מניחים הנחה שיש להרבה דברים בביולוגיה מקור אבולוציוני משותף. ידוע לנו שכל היצורים החיים משתמשים כמעט באותו קוד גנטי - סימן למקור אבולוציוני משותף. יש הרבה ויכוחים על תיאוריית האבולוציה - אנחנו לא נכנס לכל השיקולים האלו.

בגדול אנחנו מסתכלים על אבולוציה כתהליך שמתפצל - עץ מכוון על ציר הזמן:



מינים

ההגדרה של מין לצורך העניין בקורס זה הוא שלכל מין יש רצף גנטי משלו, וגם בתוך מין מסוים יש הבדלים, אבל הדמיון גדול. כשמדברים על חיות וצמחים מין זה אוסף היצורים שיכולים להתרבות זה עם זה לאורך הרבה דורות (סוס וחמור יכולים להביא פרד, אבל פרד לא יכול להתרבות לכן סוס וחמור אינו מאותו המין). כשעוברים לחיות פשוטות יותר כמו בקטריות ההגדרה לא מספיקה - לכן לא נתעסק בקורס בהגדרה המדויקת של מין. באיזשהו שלב באבולוציה הגנטיקה בין אדם לשימפנזה השתנתה. יש תאוריות על מעבר של קבוצה מסוימת ממין את הנהר, התרבו אלו עם אלו בנפרד מקבוצה אחרת, ולאורך זמן שינויים גנטיים גרמו להיווצרות של מין חדש, בהתאם לתנאים של הסביבה הגיאוגרפית בה הם נמצאים.

היווצרות של אגמים גדולים באפריקה (כמו ויקטוריה) כתוצאה מתהליך גאולוגי שאנחנו יודעים לתארך, גרמה לדגים שהיו בנהר שהפכו לאגם גדול לעבור אבולוציה יחסית מהירה שהתבטאה במראה שונה.

כעת נציג מספר שאלות:

האם יש שעון?

ההנחה היא שהיה פיצול ובעקבות זה היו שינויים. כשאנחנו מדברים על עץ כזה שבו הזמן הולך מלמעלה למטה. לפעמים נעשה הנחות: בין כל ענף לכל ענף יש מרחק, אם הענפים מודדים מרחק אז אנחנו מצפים

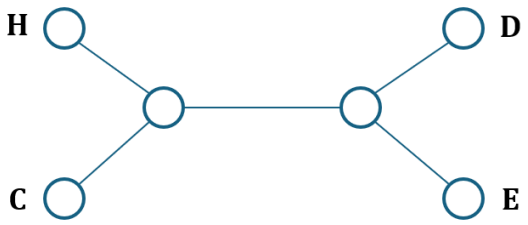
שהענפים יסתדרו כרונולוגית: כל העלים יהיו נקודת הזמן הנוכחית, וכל הקודקודים אחורנית הם נקודות קודמות בזמן - עלינו לשמור על טווח זמן. לתכונה זו קוראים לעתים שעון אוניברסאלי או יחידת זמן. כשמודדים שינויים במרחקים גנטיים במקום זמן אולי זה לא כל כך טריוויאלי להשתמש בהנחה הזו. יש ויכוח האם שינויים גנטיים מצטברים לאורך זמן תלויה רק בזמן או גם בדברים אחרים.

שאלה - האם העץ חייב להיות בינארי?

מתמטית אם נרשה ענפים קצרים אפשר לפצל כמה פיצולים לסדרת פיצולים בינארית ולכן לשם פשטות נניח שהעץ בינארי.

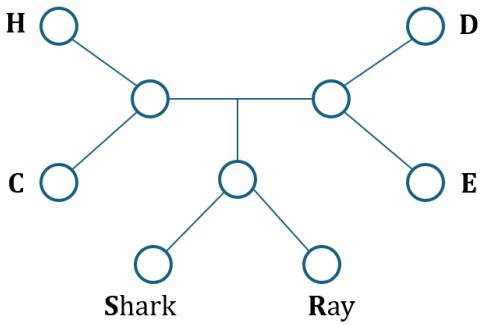
האם העץ בהכרח מכוון?

לפעמים נדבר על עצים לא מכוונים. אפשר לצייר גם כעץ לא מכוון:



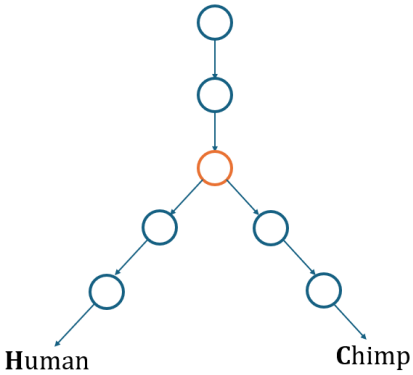
וחסר לו את הקודקוד העליון. עדיין, human ו-chimp יותר קרובים מאשר Eagle ו-Dove. נרצה עץ לא מכוון כשאנחנו לא באמת יודעים איפה השורש. בהתאם למיקום השורש נקבל עצים שונים. כאשר אין סיבה טובה להחליט מה בא לפני מה נדבר על עצים לא מכוונים כדי לא להתחייב לשורש בצורה מלאכותית.

הגרסה של עץ בינארי בעצים לא מכוונים מקיימת שדרגת קודקוד פנימי הוא 3. הגירסה הלא מכוונת נוחה יותר כשלא מניחים שיש שעון, לפעמים ננסה להסיק איפה נמצא האב הקדמון האחרון על ידי כך שנכניס עוד אלמנטים בעץ ("כל החיות באזור מסויים בעץ הן חולייתניות"). אפשר להכניס בעץ שתי חיות דומות לחולייתנים אבל הן בעלות סחוס:



לפעמים הקבוצה הזו נקראת outgroup כי הענף ממנו יוצאים המינים האלה יכול להיות השורש.

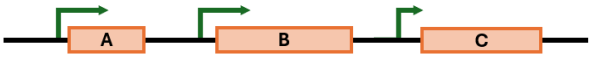
המין האחרון שהוא עדיין אב קדמון של שני מינים נקרא LCA - last common ancestor (מסומן באיור בכתום). המינים האחרים באמצע כל עוד אין פיצול שהיה בין לבין אין הרבה מה להגיד עליהם.



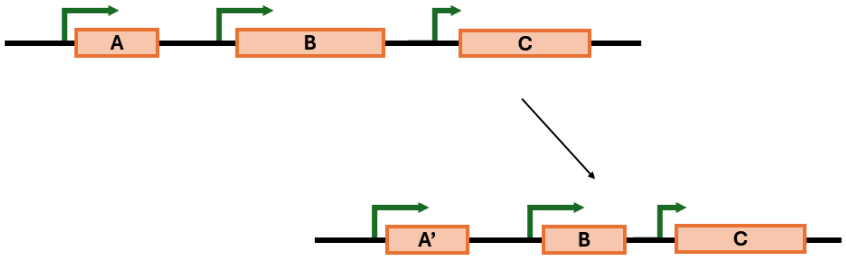
הרבה פעמים לא נדע לשים את האצבע על ה-LCA של שני מינים כמו למשל של ציפור ושל בני אדם.

עד עכשיו הסתכלנו על עצים כשהאובייקט בקצה הוא מין. לפעמים נרצה לקודד על העלה כגן, למשל גן שמקודד להמוגלובין. מה אנחנו יודעים על הגנים האלו?

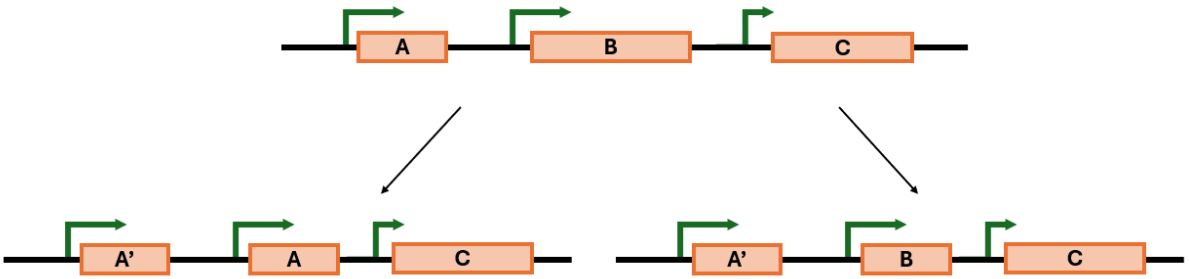
לכאורה, אפשר לשכן את האבולוציה של הגן בתוך האבולוציה של המינים - אם למין מסוים היה גן להמוגלובין הוא עבר הלאה והתפרס בעץ. בחיים זה מעניין יותר - כשאנחנו מדברים על גנים ועל גנומים באופן כללי, יכולים לקרות אירועים מעניינים ברמת הגן:



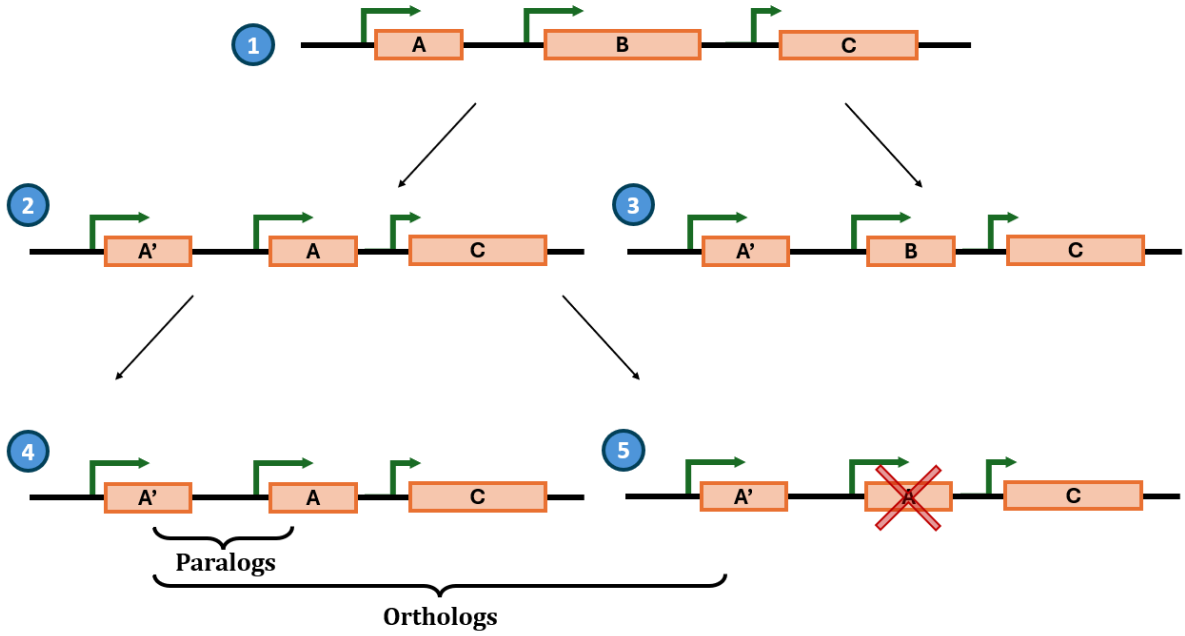
היינו מניחים שבאבולוציה כשנסתכל מה קרה נקבל משהו שנראה דומה, כשהרצף השתנה קצת:



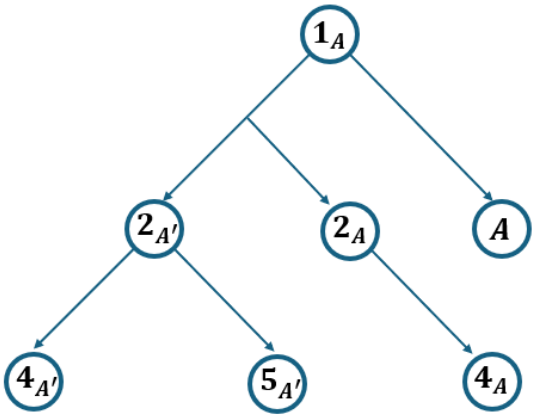
אבל לפעמים יש שינויים יותר גדולים: הגן של A השתכפל ועכשיו מופיע פעם נוספת בגנום, ו-B נמחק לגמרי:



בסיטואציה כזו הפרספקטיבה של הגנים שאנחנו רואים שונה בין המינים למרות שיש להם אב קדמון משותף. מה יקרה בשלב הבא של האבולוציה? יש גנים "מועדים לפורענות" ולכן באבולוציה נמצא אותם יותר. מה לגבי הגן B? ה-B המקורי לא יחזור יותר.



זה בא לומר שכאשר מסתכלים על אבולוציה של גנים ומסתכלים על אובייקט כמו A' אם נצייר את עץ הגנים קיבלנו:



וכאן אנו רואים שהענפים כבר לא מסודרים לפי עץ המינים.

גנים קרובים נקראים paralogs וגנים רחוקים יותר ortholog.

Ortholog – הוא גן או רצף גנים בשני מינים שונים שמקורם באותו גן קדום משותף, על ידי התפצלות מינים.

Paralog – מתייחס לשני גנים או יותר שהתפצלו מאותו גן קדום באמצעות אירוע של שכפול גנים.

בדרך כלל אחרי הכפלה של גנים - או ששני ההעתקים ביססו לעצמם תפקיד חשוב ואז יהיה קשה להיפטר מאחד מהם, או שהם מכפים אחד על השני - שניהם עושים את אותה עבודה ואז היעלמות אחד לא תשפיע באותה מידה כי התפקוד נשמר. יש מקרי קצה - פונקציה שמעניינת רק בתנאים מסוימים. דוגמה לכך היא למשל אנזים שעובר הכפלה, ואז עובד הרבה יותר יעיל - יש מקרים בהם יהיה יתרון לפרק סוכר לאט, ויש מקרים בהם יהיה יתרון לפרק מהר. יש סיטואציות בהם קורים מקרים דרמטיים בהם **כל הגנום מוכפל**. כשמכפילים גן אחד או שניים יוצרים חוסר תאימות בין ההעתקים שלו להעתקים של אחרים. בצמחים לא תמיד יש שני העתקים של הגנום (שני זוגות כרומוזומים). כשמשתכפל כל הגנום קורים דברים מעניינים - אם קודם לא היה אפשר לשנות חלבון כי שינוי גרם לבעיות, עכשיו יש גיבוי.

שימושים בעץ גנים

בחיים האמיתיים אנחנו לא רואים את האיבר ההיסטורי, רק את היצורים בני זמננו. הבעיה שאנחנו רוצים להתמודד איתה היא בעיית שחזור על סמך תצפיות בנות ימינו. השחזור הוא שימושי לצורך מתן מענה על שאלות "האם הגנים האלה זהים או התפצלו מזמן?"

לפעמים נשים יצורים על העץ האבולוציוני (כמו למשל וירוס HIV שנמצא שמקורו בקופים, חיסונים לפוליו הכילו זן מומת של פוליו שבהמשך התברר שהזן יכול להפוך להיות אקטיבי - היום כבר לא נמצא בחיסון).

שחזור אבולוציוני עוזר גם להבין פונקציה של דברים - מה נשמר ומה לא, על מה הייתה הסלקציה?

אנחנו מקבלים אוסף של אובייקטים ורוצים לשחזר את העץ הכי סביר שיוביל למה שאנחנו רואים היום תוך הנחת מספר שיקולים לא טריוויאליים:

כשציירנו עץ יש מרחק בעץ בין קודקודים (המסלול שהולכים בין שני קודקודים) - הנחת עבודה היא ששינויים מצטברים ככל שעובר יותר מרחק אבולוציוני. אם עברנו יותר מרחק מכיוון שבכל ענף מתרחשים אירועים באופן ב"ת אז יש פחות שינויים בין קודקודים עם מרחק קטן יותר.

הגדרת הבעיה

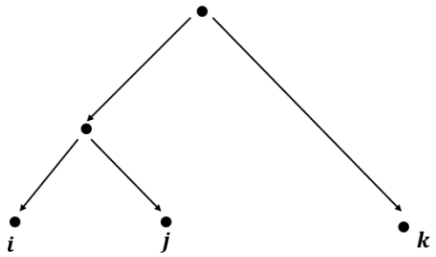
מקבלים סט של אובייקטים: gene seq, genomes,..

נציג שני אלגוריתמים.

הפשוט מביניהם היא לקחת את האובייקטים ולהפוך אותם למטריצה D שהיא distance matrix.

מטריצת מרחק סימטרית וחיובית שמתארת את המרחק בין כל שני זוגות. מהמטריצה נרצה ללמוד עץ - מגדיר מטריקת מרחק, בעץ יש רק מסלול אחד קצר ביותר בין שני אובייקטים. עץ טוב זה עץ שמתאר מטריצת מרחקים D_T שדומה ל-D שהגדרנו. רוב האלגוריתמים שנדבר עליהם יניחו שיש עץ כזה.

נניח שהאבולוציה "מתנהגת יפה" ויש שעון אבולוציוני - כל הקודקודים נמצאים בנקודת הזמן העכשווית. אם נסתכל על שלישייה יש זוג שהתפצל יותר מאוחר בזמן ומישהו שהתפצל יותר מוקדם.



מטריצה נקראת Ultra Metric אם לכל שלישייה i,j,k אפשר לסדר אותם כך שמתקיים $d_{i,j} < d_{i,k}$ או $d_{i,j} < d_{j,k}$ או דרישה קשה יותר שיתקיים שלכל שלישייה יש מישהו שאפשר להוציא החוצה כמו למשל ה-k ולקבל $d_{i,j} < d_{j,k} = d_{i,k}$.

אלגוריתם UPGMA

מה נעשה עם עצים אולטרה-מטריים? נציג אלגוריתם פשוט: UPGMA.

אלגוריתם חמדני - ניקח את השניים שכי קרובים אחד לשני: ה-input הוא מטריצת מרחקים בין העלים, ה-output הוא עץ. בכל שלב נבצע:

select i,j s.t $\forall\, l,k\ d_{i,j} \le d_{l,k}$

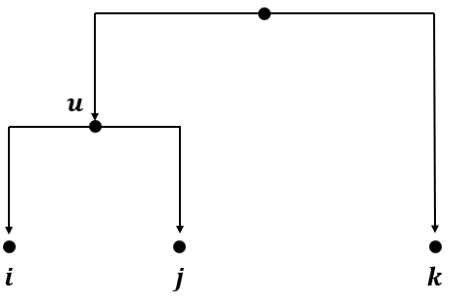
את הזוג הזה נוציא ונכניס במקומם קודקוד חדש u:

create node u

$$d_{u,k} \leftarrow \frac{1}{2}\left(d_{i,k} + d_{j,k}\right)$$

remove i,j

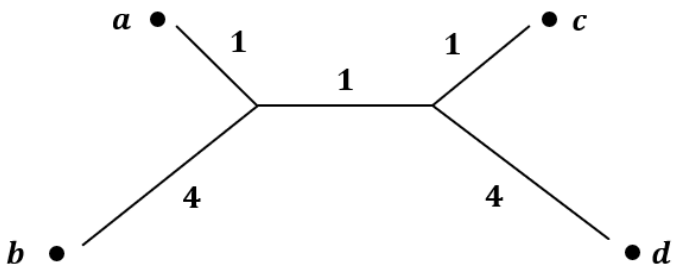
זה מתכנס ב-n-1 איטרציות. נכתוב את העץ שייראה בתור:



גובה הקודקוד הוא אמצע הדרך בין i ל-j.

דוגמה

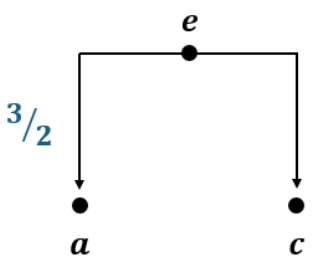
בדוגמה הזו נשתמש בעץ לא אולטרה-מטרי כדי שנראה את הבעיות שעולות בדרך. בדוגמה a, b קרובים בעץ אבל המרחקים ביניהם לא מסתדרים.



	a	b	c	d
a	0	5	3	6
b		0	6	9
c			0	5
d				0

המטריצה באמור סימטרית. הזוג הקרוב ביותר הוא a ו-c אז יוצרים מהם את העץ הבא עם הקודקוד e, ונעדכן את המטריצה בהתאם:

	a	b	c	d	e
a	0	5	3	6	1.5
b		0	6	9	5.5
c			0	5	1.5
d				0	5.5
e					0

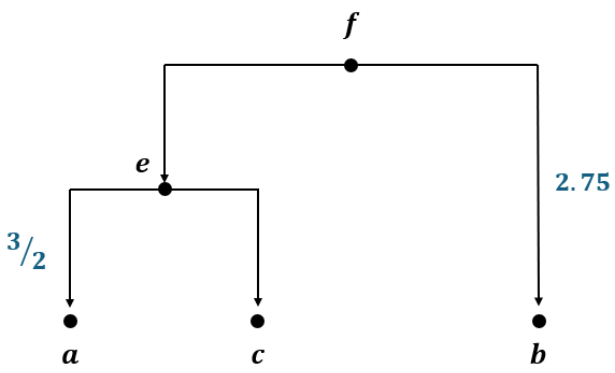


דוגמה לאופן החישוב:

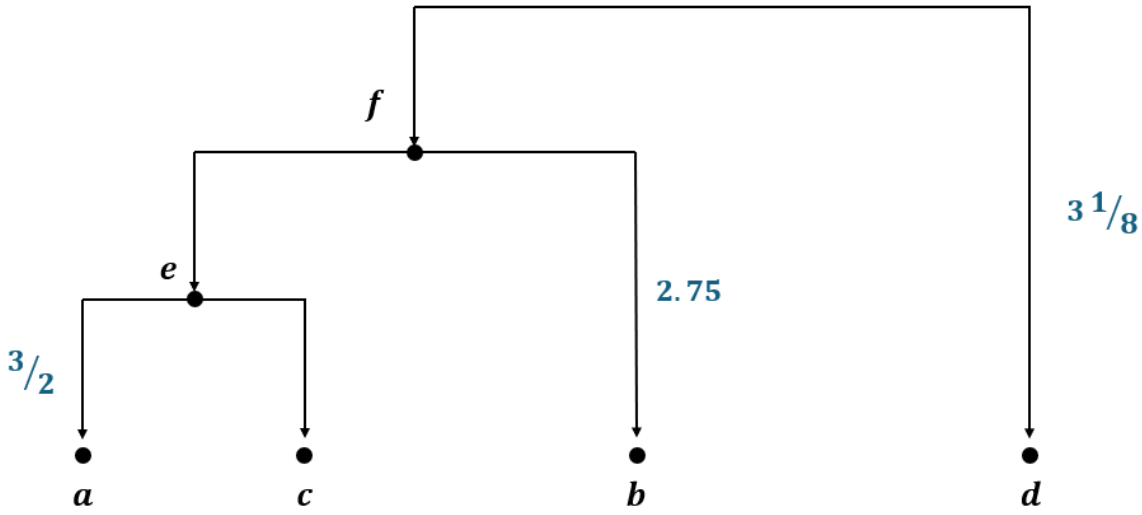
המרחק של b ל-e הוא הממוצע של המרחק של a ל-b (5) ו-c ל-b (6) לכן 5.5. נשים לב שאחרי שאיחדנו את a ואת c השורות שלהם לא רלוונטיות יותר.

עכשיו המרחק הקטן ביותר הוא למשל b ל-e. נעדכן עם הנקודה f:

	a	b	c	d	e	f
a	0	5	3	6	1.5	
b		0	6	9	5.5	2.75
c			0	5	1.5	
d				0	5.5	7.25
e					0	2.75
f						0

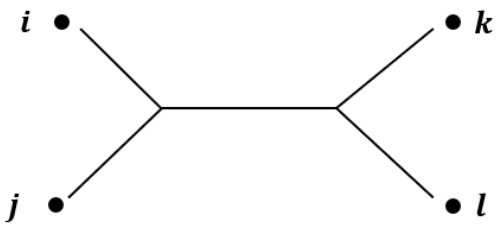


ונסיים עם העץ הסופי ($7.25 \div 2 = 3\,{}^{1}/_{8}$):



קל לבנות את העץ, אבל הוא לא משקף את העץ המקורי כי a ו-c נבחרו ראשונים בגלל המרחק ביניהם. העץ גם יוצא לא סימטרי. אלגוריתם UPGMA לא משחזר את העץ הנכון אפילו אם מטריצת המרחקים אולטרה-מטרית.

מה אפשר לעשות לעצים לא אולטרה-מטריים? נסתכל בתכונה של העץ הבא:



$$\star \quad \forall\, i,j,k,l \quad d_{i,j} + d_{k,l} \le d_{i,k} + d_{j,l} = d_{i,l} + d_{j,k}$$

התכונה הזו לא טריוויאלית, לכל רביעייה נוכל להפריד אותה לשניים כך שתתקיים התכונה הזו.

מה ההפרש ביניהם? הצלע באמצע.

יש אלגוריתם שמוכח שהוא משחזר את העץ הנכון, UPGMA לא עושה את זה גם כשהעץ אולטרה-מטרי.

אלגוריתם Neighbour Joining

נגדיר מטריצה חדשה:

$$Q_{i,j} = R_i + R_j - d_{i,j}$$

כאשר:

$$R_i = \frac{1}{(n-2)} \sum_j d_{i,j}$$

ננסה לשכנע את עצמנו בתכונה של ה-Q שהגדרנו. נניח שהעץ האמיתי הוא כזה כך ש-i,j אין ביניהם מישהו נוסף לפני k אז:

$$Q_{i,j} - Q_{i,k} = R_i + R_j - d_{i,j} - \left(R_i + R_k - d_{i,k}\right) = R_j - R_k + d_{i,k} - d_{i,j}$$

$$= \frac{1}{n-2} \cdot \left[\sum_{l \ne j} d_{j,l} - \sum_{l \ne k} d_{k,l} + (n-2) \cdot d_{i,k} - (n-2) \cdot d_{i,j} \right]$$

בשלב הזה החלפנו את הביטויים לפי הצבעים, והוצאנו $\frac{1}{n-2}$ מ-$d_{i,k}$ ומ-$d_{i,j}$ אז היינו צריכים להחזיר אותו.

$$= \frac{1}{n-2} \left[\sum_{l \ne i,j,k} \left(d_{j,l} - d_{k,l} + d_{i,k} - d_{i,j}\right) \right] + d_{j,k} + d_{i,j} - \left(d_{j,k} + d_{i,k}\right) + d_{i,k} - d_{i,j}$$

כל מה שמחוץ לסכום צריך להצטמצם (בתקווה). מה זה מזכיר לנו? את התנאי שסימנו ב-⋆. כלומר אם נעביר אגפים במשוואה שקיבלנו עכשיו, נקבל:

$$Q_{i,j} - Q_{i,k} \geq 0$$

כלומר הוכחנו ש-$Q_{i,j}$ גדול יותר מכל אחד אחר, וזה עובד בזכות העובדה ש-i,j שכנים. לכן הראינו שאם i,j הוא זוג אז הערך של $Q_{i,x}$ גדול יותר מכל x שהוא לא j, ובסימטריה גם j גדול יותר עם i מאשר עם כל אחד אחר (כלומר בהצלבת השורה ה-i עם העמודה ה-j). זה לא מוכיח שהמקסימום יהיה כזה – כלומר שהנקודה הכי גדולה ב-Q חייבת להיות זוג כזה, בשביל זה צריך לעבוד טיפה יותר קשה – אך לא נוכיח במסגרת הקורס.

בשיעור הבא נדון איך מעדכנים את המרחקים ואיך משתמשים בהם.

שבוע 7 – הרצאה 13

חזרה

בשיעור הקודם התחלנו מאוסף של רצפים מהם אנחנו מעוניינים לחשב מטריצה D, ויש אלגוריתמים שמהמטריצה מנסים לשחזר מה העץ שקיים כך שהמרחקים במטריצה מתאימים למרחקים בעץ.

אלגוריתם UPGMA מניח שהמרחקים הם אולטרה-מטריים, וכשההנחה לא מתקיימת העץ שמתקבל שגוי.

התחלנו לדבר על אלגוריתם Neighbor Joining שמניח additive distance, ועבדנו עם דוגמה רצה שהיום נריץ עליה את Neighbor Joining.

הגדרנו גם: $Q_{i,j} = R_i + R_j - (n-2) \cdot d_{i,j}$ ו-$R_i = \sum_k d_{i,k}$ (זה שקול להגדרה מההרצאה הקודמת).

הוכחנו Lemma שאם i, j שכנים (כלומר האב הקדמון המשותף שלהם מיד מעליהם בעץ) אז לכל k מתקיים $Q_{i,j} \geq Q_{i,k}$. המשפט שלא נוכיח הוא שאם $Q_{i,j} \geq Q_{k,m}$ לכל k, m אז i, j שכנים (כלומר ה-Q הכי גדול חייב להתקבל מזוג שכנים).

האלגוריתם

$$(i^*, j^*) = \arg\max_{i,j} Q_{i,j}$$

$$\text{introduce } u \ (from\ i\ and\ j)$$

אנחנו צריכים להגדיר את המרחקים $d_{u,k}$ ואת $d_{i,u} = d_{j,u}$.

$$d_{i,u} = \frac{1}{n-2} \sum_{k \neq i,j} (d_{i,k} - d_{u,k})$$

$$d_{j,u} = d_{i,j} - d_{i,u} \quad \text{(or same as above)}$$

$$d_{u,k} = \frac{1}{2}(d_{i,k} + d_{j,k} - d_{i,j})$$

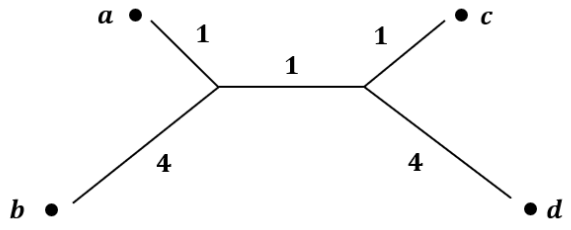
$$\text{remove} \quad i, j$$

$$\text{add} \quad u$$

$$n \leftarrow n - 1$$

דוגמה

סט העלים שנתחיל איתו הוא: $Leafs = (a, b, c, d)$



נחשב את R סכום המרחקים מכל קודקוד לאחרים:

$Leaf$	R
a	14
b	20
c	14
d	20

ואת המטריצה Q:

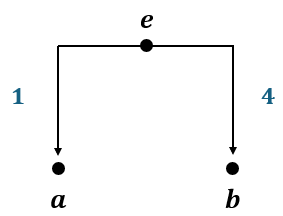
Q	a	b	c	d
a	0	24	22	22
b		0	22	22
c			0	24
d				0

דוגמה לחישוב התא ac: הסכום של $R(a) + R(c) = 28$ ואז חיסור $2 \cdot d_{a,c} = 6$.

באופן שרירותי נבחר a, b. נוסיף קודקוד שנקרא e. ונחשב:

$$d_{e,c} = \frac{1}{2}\left(d_{a,c} + d_{b,c} - d_{a,b}\right) = \frac{1}{2}(3 + 6 - 5) = 2 \qquad d_{e,d} = 5$$

הענפים הם:



- סוף שלב ראשון -

$Leaf$	R	
a	14	
b	20	
c	14	7
d	20	10
e		7

Q	c	d	e
c	0	12	12
d		0	12
e			0

קיבלנו תכונה שבעץ בגודל 3 כל זוג שנבחר הם שכנים – נשאר רק לשחזר את המרחקים.

סיכום

האלגוריתם יחסית מוצלח, מוכח להיות נכון כשהמרחקים באים מעץ, והוא מאוד robust גם כשהמטריצה לא מדויקת, ולכן שימושי יותר מ-UPGMA. החולשה החזקה היא שהוא עובד רק עם מרחקים – בזה נטפל בהמשך הקורס.

חישובי מרחקים

שאלה שהיינו רוצים לשאול – האם מספר ה-mismatches מתנהגים כמו שהייתי רוצה שיתנהגו מרחקים?

נניח שיש לנו חתיכה מהעץ עם 3 קודקודים, שהמרחק ביניהם הוא 1. אדיטיביות אומרת שאנחנו מצפים שהמרחק בין a ל-c יהיה כמובן 2. השאלה האם התנהגות כמו מספר השינויים מתנהגת בצורה דומה. לצורך הדוגמה, נניח שהרצף הוא בינארי (כדי להתעלם מהעובדה של "מה היה השינוי").

נגדיר: $\boldsymbol{t} = \%$**base changes** (אם $t = 1\%$ אז 1 מכל 100 רצפים השתנו). יהיו מספר רצפים:

$$a_1, \dots, a_{1000} \quad ; \quad b_1, \dots b_{1000} \quad ; \quad c_1, \dots, c_{1000}$$

אז $\mathbb{P}(b_i = a_i) = 0.99$. באותו אופן $\mathbb{P}(b_i = c_i) = 0.99$, $\mathbb{P}(b_i = \bar{a}_i) = 0.01$.

$$\mathbb{P}(a_i = c_i) = ?$$

האינטואיציה אומרת שלפי אדיטיביות התשובה היא 0.98. נחשב:

$$\mathbb{P}(a_i = c_i) = \mathbb{P}(a_i = b_i = c_i) + \mathbb{P}(a_i \neq b_i, b_i \neq c_i) = 0.99 \times 0.99 + 0.01 \times 0.01$$
$$= 0.9802$$

ואם היינו לוקחים $t = 10\%$ הפער היה אפילו יותר גבוה.

אנחנו רוצים לדבר על האבולוציה כתהליך הסתברותי שמקבל פרמטר של יחידות זמן ומדבר על השינוי, אבל אנחנו רוצים שתהיה התנהגות אדיטיבית. נעבור להתנהגות של שרשרת מרקוב, בכל יחידת זמן נקפוץ יחידה, ולמען האמת כבר במשוואה למעלה הנחנו מרקוביות על פני ציר זמן.

נרצה להגדיר הסתברות מעבר τ. מה הסתברות המעבר כשהולכים k צעדים בשרשרת? נצטרך לסכום על כל מה שקורה עד אותה נקודה (סימון – הסתברות המעבר בזמן $k+1$ מהאות x לאות y):

$$\tau_{x,y}^{(k+1)} = \sum_{z} \tau_{x,z}^{(1)} \cdot \tau_{z,y}^{(k)} \stackrel{?}{=} \sum_{z} \tau_{x,z}^{(k)} \cdot \tau_{z,y}^{(1)}$$

ה-? נמצא מעל השווה כי זה משהו שהיינו רוצים לקבל.

בעצם – יש כאן מכפלת מטריצות: $\tau^{(k+1)} = \tau^{(k)} \cdot \tau^{(1)} = \left(\tau^{(1)}\right)^{k+1}$. למדנו מזה שאם אנחנו רוצים להסתכל על זמן בתור משהו אדיטיבי, אני צריך לקחת את השינויים שאני לוקח ביחידת זמן, ולהעלות אותם במספר יחידות הזמן. הנחה מובלעת שהאבולוציה אותו דבר כל יחידת זמן אבל בלתי תלויה בעבר.

למשל עבור המטריצה ו-$k = 10$:

τ	A	C	G	T
A	0.99	0.003	0.003	0.003
C	0.003	0.99	0.003	0.003
G	0.003	0.003	0.99	0.003
T	0.003	0.003	0.003	0.99

To (columns), *From* (rows)

נחשב את החזקה ונגלה איפה נהיה. מה הבעיה? כמו בהילוך שיכור, כשממשיכים יותר מדי זמן לא זוכר את נקודת ההתחלה – כלומר נגיע להסתברות שלא תלויה בנקודת ההתחלה.

היינו רוצים לדבר על מרחקים אבולוציוניים לא ביחידות שלמות – יותר טבעי לחשוב על אבולוציה על כמשהו שמתנהג בצורה רציפה (ולא בדידה).

הגדרת פונקציה ונגזרתה

אנחנו רוצים פונקציה: $\mathbb{P}: \quad \mathbb{R}^+ \to [prob]$ (מספר ב-$\mathbb{R}^+$ אל מטריצת הסתברויות), כלומר $\mathbb{P}(t)$ ייתן את המטריצה של לחכות זמן t, ואנחנו רוצים שהיא תהיה רציפה. פתרון שאפשר להציע זה לדבר על לחלק את יחידת הזמן t ל-ε במרווחים יותר ויותר קצרים, כמו: $\mathbb{P}(t) = \mathbb{P}\left(\frac{t}{n}\right)^n$.

נפתור את הבעיה בעזרת נגזרות:

$$\mathbb{P}'(t) = \lim_{\varepsilon \to 0} \frac{\mathbb{P}(t+\varepsilon) - \mathbb{P}(t)}{\varepsilon}$$

$$\mathbb{P}(t+\varepsilon) = \mathbb{P}(t) \cdot \mathbb{P}(\varepsilon)$$

כמו לשבור את הזמן ל-2: לחכות $t+\varepsilon$ זה כמו לחכות t ואז לחכות ε. לכן:

$$\mathbb{P}(t+\varepsilon) - \mathbb{P}(t) = \mathbb{P}(t)[\mathbb{P}(\varepsilon) - I]$$

נקבל:

$$\mathbb{P}'(t) = \mathbb{P}(t) \cdot \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}[\mathbb{P}(\varepsilon) - I]$$

ונגדיר בעזרת פיתוח טיילור מסדר 1:

$$\mathbb{P}(\varepsilon) = \mathbb{P}(0) + \varepsilon \cdot \mathbb{P}'(0) = I + \varepsilon \cdot \mathbb{P}'(0)$$

בזמן 0 שום דבר לא משתנה, לכן $\mathbb{P}(0) = I$. לכן אם נציב חזרה בנוסחת הנגזרת נקבל:

$$\boxed{\mathbb{P}'(t) = \mathbb{P}(t) \cdot \mathbb{P}'(0)}$$

קצב השינוי

הגדרה – Rate Matrix שתגדיר את קצב השינוי: $R = \mathbb{P}'(0)$.

תכונות:

$$R_{a,a} \leq 0$$

כי מדובר בשינוי (הנגזרת) שיהיה לערך "1", הוא חייב לרדת, ולכן הסימן שלילי. ככל שעובר זמן הסיכוי שאין שינוי קטן.

$$R_{a,b} \geq 0 \quad (a \neq b)$$

מאותה סיבה. נרצה שיתקיים גם:

$$\sum_b R_{a,b} = 0$$

כיוון שגם אחרי ε זמן, אנחנו עדיין מסתכלים על הסתברויות ולכן הסכימה צריכה להתאים ל-1.

$\boldsymbol{R}$	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

$\boldsymbol{\alpha \neq 0}$

שינוי יוניפורמי כאשר α קבוע שאומר כמה מהר השינוי קורה (אם נדחס את הזמן נזיז גם את α). למטריצה הזו יש שם – המודל של **Jukes Cantor**, שני האנשים שהמציאו אותו כמודל יוניפורמי לאבולוציה.

$$\boxed{\mathbb{P}'(t) = \mathbb{P}(t) \cdot R}$$

אנחנו רוצים לשחזר את $\mathbb{P}(t)$ ממשוואת הנגזרת. מדובר במשוואה דיפרנציאלית. הפתרון:

$$\mathbb{P}(t) = \mathbb{P}(0) \cdot e^{tR} = I \cdot e^{tR}$$

$$\boxed{\mathbb{P}(t) = e^{tR}}$$

אבל מה זו e בחזקת מטריצה? מטריצה:

$$e^x = e^0 \cdot \sum_n \frac{1}{n!} (e^0)^{(n)} x^n = \sum_n \frac{1}{n!} x^n$$

$$e^A = \sum_n \frac{1}{n!} A^n$$

האם זה לא פשוט המטריצה כשכל קומפוננטה שלה בחזקה? **כן, כשהמטריצה אלכסונית.**

$$D = \begin{bmatrix} d_{1,1} & & 0 \\ & \ddots & \\ 0 & & d_{n,n} \end{bmatrix} \Longrightarrow e^D = \begin{bmatrix} e^{d_{1,1}} & & 0 \\ & \ddots & \\ 0 & & e^{d_{n,n}} \end{bmatrix}$$

במקרה של פירוק אורתוגונלי מתקיים:

$$A = U^T D U \Longrightarrow A^n = U^T D^n U$$

ואז:

$$e^A = U^T \cdot \left[\sum_n \frac{1}{n!} D^n\right] U = U^T e^D U$$

אז אלו שתי דרכים לחשב את הפתרון של Jukes Cantor. יש את הדרך הפשוטה יותר עבור JC:

$$\mathbb{P}(t) \overset{\text{סימטריה}}{=} \begin{bmatrix} r & & s \\ & \ddots & \\ s & & r \end{bmatrix}$$

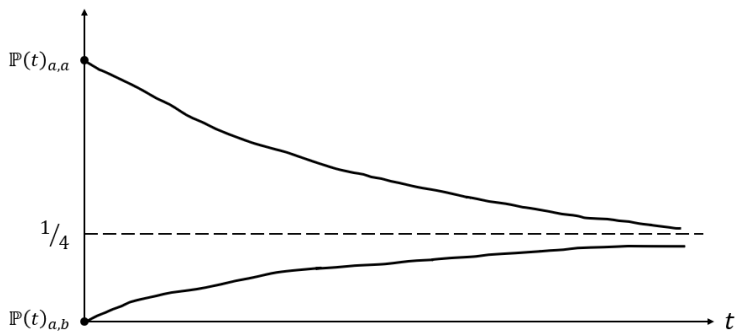
כלומר בזמן t יהיו רק 2 סוגים של הסתברויות. מאחר שחייבת להיות סכימה ל-1: $r = 1 - 3s$

ניזכר $\mathbb{P}'(t) = \mathbb{P}(t) \cdot R$ ונמשיך לפתח את המטריצה:

$$\begin{bmatrix} -3\alpha r + 3\alpha s & & \alpha r - \alpha s \\ & \ddots & \\ \alpha r - \alpha s & & -3\alpha r + 3\alpha s \end{bmatrix} = \cdots \Longrightarrow$$

$$\boxed{s = \frac{1}{4}(1 - e^{4\alpha t}) \quad ; \quad r = \frac{1}{4}(1 + 3e^{-4\alpha t})}$$

מתקיים עדיין שבזמן 0 המטריצה R היא היחידה. כש-$t \to \infty$ אנחנו שואפים ל-$1/4$:



לסיכום, למדנו על מטריצות של זמן רציף. למדנו מטריצת Rate שנקראת Jukes Cantor, וראינו תכונה של JC כאשר t מאוד גדול. בהמשך נראה שתכונת ההתכנסות לא מקרית.

שבוע 8 – הרצאה 14 (הרצאת אורח)

הרצאת אורח: מורן יסעור, Assembly Sequencing

הקדמה והגדרת הגישה

פעם היו מקבלים מאות רצפים או רצף אחד מאוד ארוך, הפכנו את הצירים להמוני רצפים קצרים (כ-20 מיליון רצפים, כל אחד באורך של כ-150 נוקלאוטידים). איך מזה נוכל לרצף גנום של חיידק? Assembly.

Read – מילה על פני א"ב של ACGT.

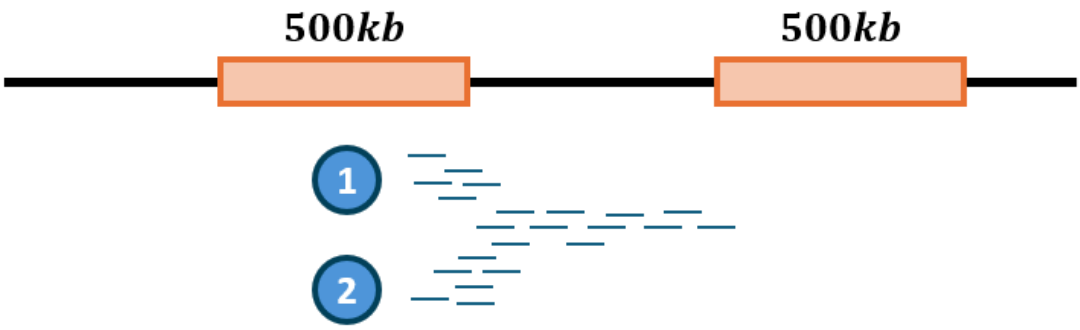
יש כמה גישות:

(1) reference based – היצמדות לידע קודם על הרצף.

(2) reference free – לא יודעים מה התמונה הכללית.

מתי reference free יעבוד? כשיש הרבה חפיפה. נניח שיש מספיק reads ביחס לאורך שרוצים להרכיב.

נניח שיש קטע באורך 500 נוקלאוטידים שחוזר אחרי כ-2000kb, יש שתי דרכים להשלים את ה"דומינו" באזור שהוא התפר בין האזור החוזר לאזור הרגיל:

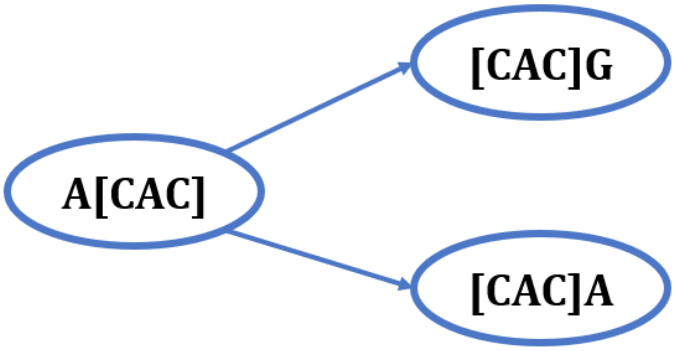


k-mer – כמו k-יה, מעל הא"ב {A,C,G,T}.

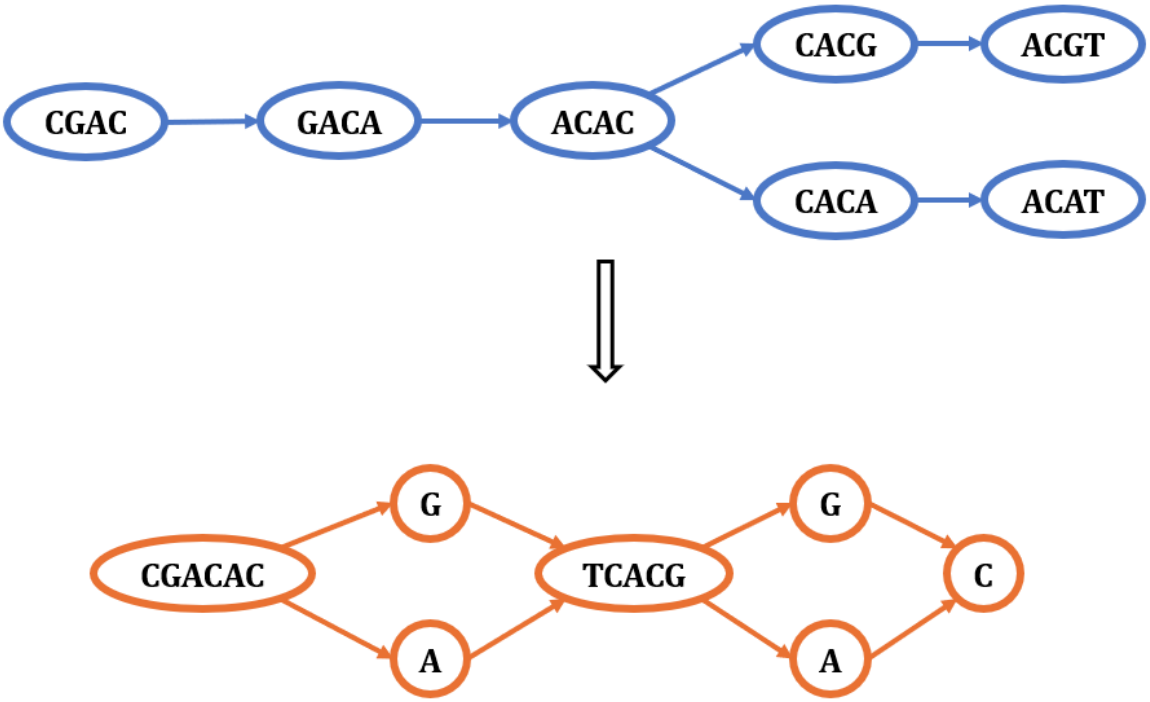
De-Bruijn Graph

נפרק כל רצף באורך 150 לכל ה-k-יות שנמצאות בתוכו. לצורך הציור נניח שיש mer-4 (בפועל כנראה נסתכל על mer-25). תהיה צלע בין שני k-mer אם:

1. הם חופפים את ה-k-1-mer (סיפה מתאימה לרישה, מסומן בתרשים למטה ב-[ ]).
2. יש מילים בדאטה המקורי שיכילו את ה-k+1-mer (כלומר אם היה ACAC ו-CACG צריך שהמילה ACACG תהיה בדאטה).

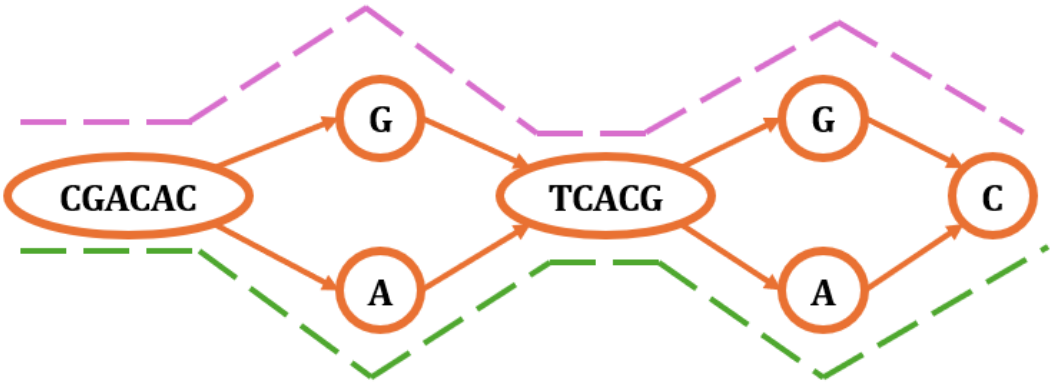


אפשר גם למשקל את הצלעות. האם כל מסלול בגרף הוא read? לא - יכול להיות מסלול ארוך יותר מ-read. ה-de-bruijn גרף הזה מאוד ארוך, אז נהפוך אותו לקומפקטי יותר בעזרת assembly graph. נרצה לעבור לגרף שבו נוכל להפעיל אלגוריתמים של גרפים, אז נעבור ללינאריזציה בלי פיצולים:



איך נשמור את המידע על משקל הצלעות? אפשר לשמור משקל ממוצע, חציוני, משקל לכל אות..

האם כל read מיוצג על ידי מסלול בגרף הזה? כן. כי שמרנו את כל המידע שהיה לנו במקור. האם כל מסלול מייצג read? לא - המסלול יכול להיות ארוך מעבר לאורך של read. יכול להיות גם פיצול נוסף בהמשך.



כאן יכול להיות מסלול שלא היה במקור, כלומר לא הייתה מילה שיוצרה קודם שהייתה תומכת בהם.

אנחנו רוצים שהגרף ייצג את המקור כמה שיותר במדויק. כל מסלול בו צריך להיות נאמן למקור של ה-20M reads שריצפנו. איך נפתור המקומות בהם יש צומת ורק חלק מהמסלולים נכונים? אם זה קצר, נוכל לפתור על ידי reads בודדים, אחרת לא נוכל לפתור את המבנה לרצפים ארוכים.

איך דוגמה עם הרצף החוזר תראה ב-assembly graph? החלק החוזר rep1 יהיה עם מעגל, ובגנום האנושי יש הרבה חזרות כאלו. באלגוריתמים של גרפים, ברוב המקרים הנחנו שאין מעגלים (DAG - Directed Acyclic Graph).

מתוך הגרף נוציא את all probable paths מתוך הנחה שהמסלול היה יכול להופיע בגנום המקורי. איך נעשה את זה? אפשר לחשוב על כיסוי שהיינו מנסים לשמור אחיד, אבל הכי פשוט לקחת את כל הקודקודים שאין להם node שנכנס אליהם, נחבר אותם לקודקוד התחלה, ואת כל מי שלא יוצא ל-node אחר לקודקוד בור, נוריד את המסלול עם הזרימה הטובה ביותר, נמצא את האחד הבא וכך הלאה.

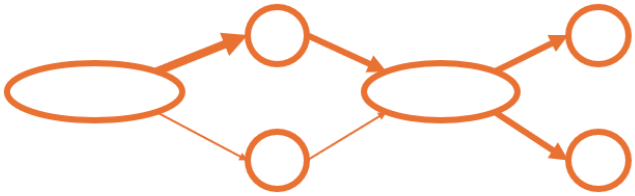
RNA Assembly

מה קורה אם במקום לעשות DNA Assembly נעשה RNA Assembly?

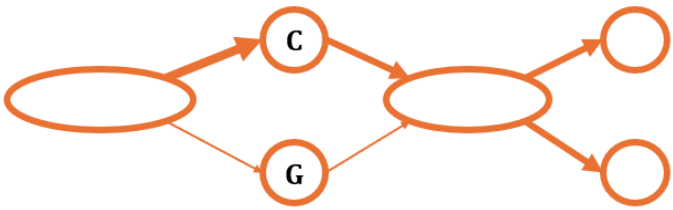
ה-DNA דו-גדילי, ב-RNA נייצר רצף אחד חד גדילי. נייצר המון רצפים מהגן שמתבטא הרבה, ומעט רצפים מהגן שמתבטא מעט. נעשה assembly באותו האופן. נקבל גרף שנראה דומה, ועכשיו ננסה לשחזר מהם את ה-RNA, את כל ה-transcriptome כלומר "all mRNA molecules".

השלב הראשון הוא להבין שנקבל הרבה יותר מילים מגן שמבוטא הרבה. אנחנו רוצים מתוך ה-assembly לדעת כמה פעמים גן הופיע.

אם הגנים מספיק שונים אחד מהשני, נצפה לקבל רכיב קשירות לכל transcript. יהיו רכיבי קשירות עם coverage גבוה, וכאלה עם coverage נמוך (לפי כמה reads תמכו בצלע). כשה-coverage גבוה נוכל לדעת שהיו יותר reads שתומכים בו, כלומר זה כנראה גן שהתבטא הרבה.



מתי יכול להיות פיצול? כשהאללים שונים - שוני קטן בין שני הכרומוזומים והגן התבטא משניהם.



אם יש הרבה החלפות כאלה נקבל הרבה "בועות" בגרף, ונוכל להבדיל בין כמות הטרנסקריפטים מהגן האימהי לגן האבהי. כך נוכל לפתור את מבנה X והם מספיק קרובים כך שהחלק שחוזר לאותו צומת נוכל "ליישר" את ה-X ובסוף ליצור שני רצפים במקום את כל הקומבינציות האפשריות.

בכל שיש יותר שונות, והנקודות של השונות מספיק תכופות אפשר לבצע phasing - מימוש הפתרון לכל ה-X-ים בדרך.

מתי נרצה להשתמש ב-reference free? במקרים שבהם אין לנו את הגנום של היצור באיכות טובה, אפשר להיות אגנוסטיים לעובדה האם יש או אין את הגנום המלא. מה קורה לגנום בסרטן? בלגאן שלם. באחד הכרומוזומים יכולה להיות פתאום תוספת, ולהצמד ל-reference יכול להוות הגבלה.

אורך המילים ה-reads גם מהווה פקטור. ריצופים שנותנים reads ארוכים יותר יובילו לרצפים באורך 1,000 עד 10,000, זה יכול לפתור בעיות בגרף שלא יכולנו לפתור קודם. hybrid assembly משתמש ברצפים קצרים וארוכים יחד.

Metagenomic Assembly

במקום לגדל חיידק בודד נוציא דוגמית שמכילה את כל החיידקים באזור מסוים - ויש עכשיו המון גנומים שנרצה להרכיב. נוכל להשתמש בשיטות reference free כמו assembly כדי להרכיב אותם. כל האזורים החזרתיים בגנום מוגברים עכשיו פי כמה, כי מדובר בגנים שמופיעים בכל אחד מהחיידקים שדגמנו.

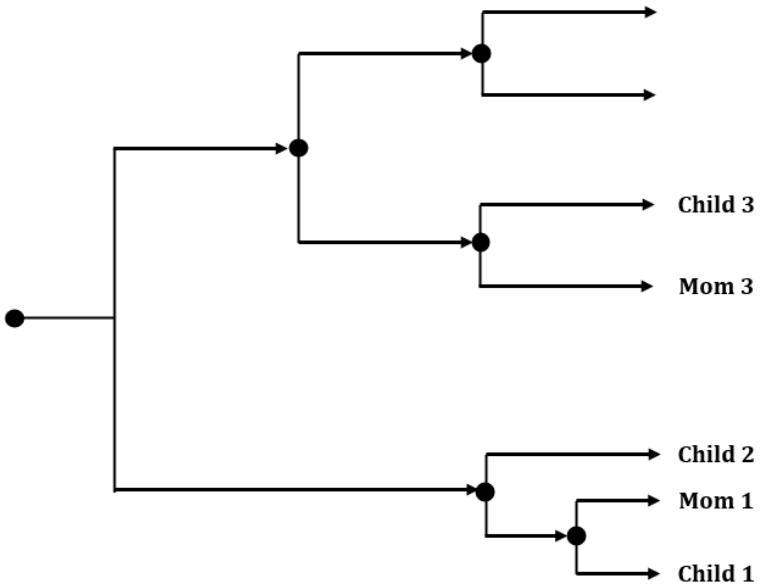
בעולם קל, כל רכיב קשירות היה חיידק אחד, אבל בגלל העובדה שיש להם המון גנים דומים, נקבל את החזרתיות. כאן כבר ממש יש צורך גדול לפתור את הלופים (10 פאזלים שהחתיכות שלהן מעורבבות, וצריך להרכיב אותם בלי תמונה ל-reference).

עצים פילוגנטיים

עץ שמתאר התפצלות של מינים לאורך זמן. איך בונים עץ כזה? צריך גנומים של יצורים בני זמננו, ומחשבים מטריצת מרחק שממנה בונים את העץ.

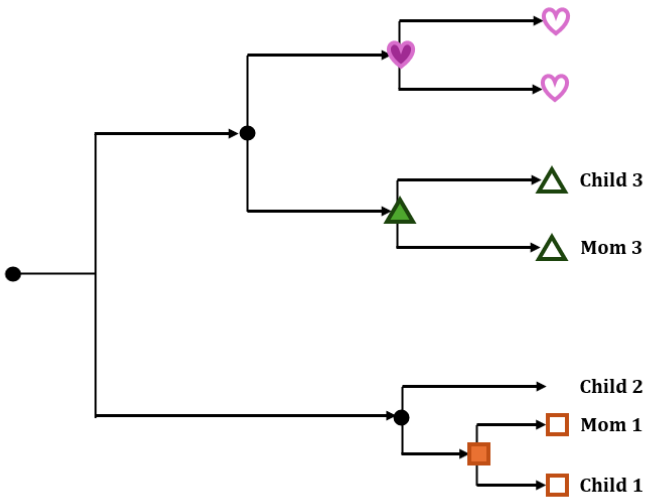
אם נבנה gene tree לגן מסוים, נקבל קונספט דומה. העלים יהיו רצף גנים בני זמננו, והקודקודים הפנימיים יהיו גרסאות קודמות של אותו הגן. יכולה להיות הכפלה של גנים (או הגנום כולו) בעץ המינים, שתבוטא בהכפלה שלמה בעץ הגנים, ואז נקבל שני מקטעים בעץ שהם למעשה זהים.

כשאנחנו מסתכלים על עצים פילוגנטים בעולם מיקרוביום, אפשר לזהות את החיידקים בכל אחת מהנישות - עור, מעי וכו. שאלה שאפשר לשאול היא האם החיידק שאנחנו רואים אצל האמא הוא החיידק שאנחנו רואים אצל הילד, למשל ecoli. האם לכולנו יש את אותו ecoli? סביר שלא. נעבור מ-species tree (עץ מינים) ל-starins tree (עץ זנים) - למשל כדי להשוות בין זנים של ecoli.

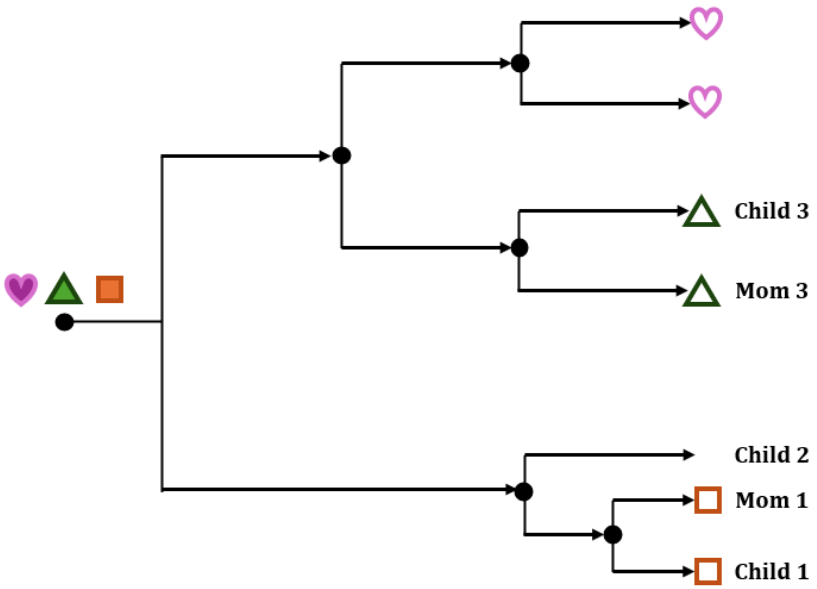


מדובר בעץ אולטרה-מטרי, וכל עלה מייצג דוגמה. גם כאן צריך מטריצת מרחק, באמצעותה נשווה בין דוגמאות. לכל דוגמה (ילד 1, אמא 1) נרצה לדעת מה הזן ecoli הכי בולט.

אפשר לייצר עץ שכל קודקוד בו הוא זן בדוגמה שיש לנו, ואפשר לשאול אילו מהזנים יותר קרובים אחד לשני. יהיו חיידקים עבורם כשיש את הדוגמה גם אצל הילד וגם אצל האם מדובר במשפחה. נוכל להניח שיש קשר בין המצאות של זן אצל האם והילד. כל תת עץ מייצג משפחה.



הזנים משותפים בתוך המשפחות. מנגד יהיו עצים שכל האבות הקדמונים המשותפים יהיו רחוקים יותר. יהיו מינים שבהם העץ יפה ומסודר והאבות הקדמונים המשותפים קרובים, ויהיו מינים שבהם רחוק יותר בעץ.



במצב הזה, זה יכול להיות כי הילד קיבל חיידק ממקום אחר מהאם, כלומר חיידק חדש נכנס והשתלט. מינים שבהם יש תחלופה חזקה מאוד בין המינים ייראה כך.

איך בונים עץ שמכיל מין אחד ומבדיל בין הזנים השונים שלו?

יש אוכלוסיית חיידקים A,B,C,D. יש לנו את הגנום של כל אחד מהם. נניח שנרצה להסתכל רק על E.coli. מה נוכל לעשות? נניח שיש 20M reads מכל דוגמה אבל רק 5M מהם שייכת לאיקולי. איך נבנה מטריצת מרחקים שמשווה בין ה-4 זנים של ה-e.coli?

ננסה למפות את כל ה-20M, אזורים שבהם אנחנו רואים חפיפה, יצביעו לנו על אזורים ייחודיים לאיקולי שיכולים להעיד על הימצאותו. נקרא להם marker genes. אם יש לנו כ-100 כאלה, נייצר מהם אחד ארוך שהוא חיבור של כל ה-marker genes, ועכשיו אפשר לחשוב על זה כמו העצים שפגשנו קודם - כל שונות על פני ה-marker genes היא שונות שנבעה מהזנים השונים של החיידק.

הערה: הסתמכות על ה-marker gene מבוססת על זה שאנחנו יודעים שהוא לא מזהה אף חיידק אחר. גם אם הוא משויך גם לחיידק אחר, נזכור שעדיין אנחנו מסתכלים על כ-100 גנים אז גם אם אחד מהם לא מדויק רק עבור איקולי - it evens out. כמו כן, תמיד נסתכל גם על reference based.

כמעט כל שיטה שנשתמש בה יש לה הנחה על ה-reference, תמיד נרצה להבין האם אנחנו בעולם reference based או reference free. ה-free דורש יותר מאמץ, אבל פותח אותנו ליותר אפשרויות.

שבוע 8 – הרצאה 15 (הרצאת אורח)

הרצאת אורח: פרופ' מור ניצן, תהליכי היווצרות מבנים רב-תאיים

הקדמה

בקבוצת המחקר של פרופ' ניצן מפתחים שיטות חישוביות שעוזרות לחלץ מידע מתוך דאטה ביולוגי גדול רועש ורב-ממדי. למשל, מידע שמגיע מהרבה גנים במקביל בצורה מעורבבת או בקונטקסט מרחבי. פיתוח שיטות המבוססות על למידת מכונה, אופטימיזציה, מערכות דינאמיות וכו' – על מנת לשחזר שכבות של אינפורמציה שאי אפשר לקודד באופן ישיר. בין השאר, אחד מהדברים שהתחילו מהם ביחד עם פרופ' ניר פרידמן היה שחזור מבנה מרחבי של רקמות מתוך מידע על תאים בודדים שלא נמדדים בקונטקסט המרחבי שלהם.

היום נחשוב על השאלה הזו מזווית אחרת – איך תאים הגיעו לקונפיגורציה הזו מראש? הרבה פעמים נדבר על פיתוח של שיטות חישוביות/ אלגוריתמים כדי להבין משהו על המבנה, הפונקציה, דפוסים בתוך דאטה ביולוגי. אנחנו נתמקד בגישה שחוקרת איך המערכת הביולוגית הגיע למצב שלה מנקודת מבטה, ולא רק על ידי מדידת המצב הנוכחי מבחוץ. "האם התאים עצמם יודעים איפה הם נמצאים במרחב?" - לאורך השיעור נצלול לעומק של דוגמה ספציפית.

הגדרת השאלות

איך תבניות ומבנים פיזיים נוצרים? נתמקד בתהליכים self-organized (כמו בתהליך התפתחות) כלומר קבוצת תאים שמתארגנת לכדי מבנה תלת-ממדי.

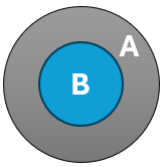
1) איך מבנים רב-תאיים נוצרים בדרך רובסטית?
2) עיצוב עקרונות לבניית מבנים רב-תאיים רובסטיים.

המחקר נעשה על ידי: Guy Pelc, Wesley Mckeithen, Michael Brenner, Wendell Lim.

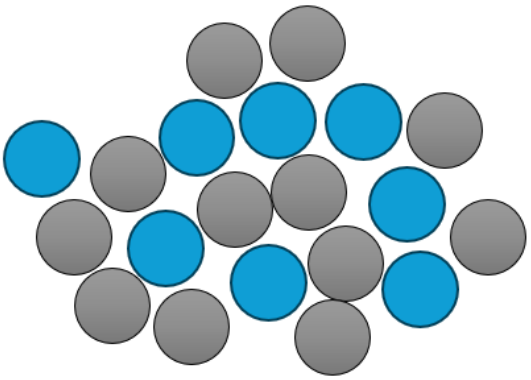
אם נרצה לדבר על תהליכי היווצרות של ארגונים רב-תאיים בתהליך כלשהו, נתחיל מדוגמה פשוטה שמהווה אבן בניין.

נקודת פתיחה

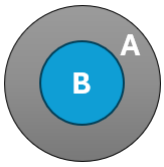
נקודת הפתיחה היא basic multicellular building block. נתייחס אליו כמבנה שבנוי מליבה (מסוג אחד של דברים, למשל סוג תא **B**), ומעטפת (למשל מסוג תא **A**).



נניח שבנקודת התחלה יש לנו cluster של תאים, חלקם סוג A וחלקם סוג B. השאלה היא איזה תהליך או מנגנון יכול לגרום לנו להתכנס למבנה שתואר לעיל:



בדרך כלל במחקר, נרצה קודם למצוא הסבר בצורה הפשוטה ביותר. הפתרון במקרה הזה הוא אכן כך – לכל התאים יכולת להסתובב במרחב, אבל לתאים מסוג B יש נטיית "הידבקות" אלו לאלו, יותר מאשר B ל-A או A ל-A. הרעיון נקרא Differential Adhesive (והוא הוצע על ידי שטיינברג בשנת 64). זה יוצר סוג של מיון המביא לתוצאה של מבנה כזה:



התאים שנדבקו ממשיכים לנוע ביחידה אחת, אבל פחות מתא בודד – נדרש יותר מקום כדי לנוע וכדומה. אפשר היה לחשוב על מנגנונים אחרים (בנוסף להדבקות) שעשויים לספק הסבר להיווצרות המבנה הזה.

הגדרת האנרגיה

(אנרגיה = במובן מסוים כמו Loss של המערכת). נגדיר:

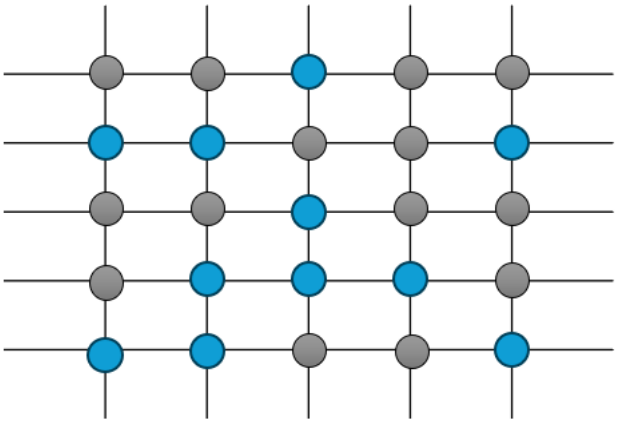
$$H = \sum_{i,j \in NN} J\big(\sigma(i), \sigma(j)\big)$$

עוברים על כל הזוגות i, j שהם Nearest Neighbors בקונפיגורציה שלנו, ולכל אחד מהם שואלים מה ה-state שלו $\sigma(i) \in \{A, B\}$. כתלות בזה, שואלים מה רמת הדביקות ביניהם. המערכת שואפת להגיע למצב שבו התאים כמה שיותר דבוקים. J מספק לנו ערך לכל זוג שכנים, במקרה הזה מטריצת 2×2 סימטרית, כשאנחנו יודעים $J_{BB} < J_{BA}, J_{AA}$.

הקונפיגורציה שממזערת את האנרגיה תהיה הקונפיגורציה הארוזה שהצגנו. בביולוגיה אנחנו לא יכולים מראש לסדר את המבנה האופטימלי ולשחרר, אלא תהליך של ארגון פנימי. נחזור לקונפיגורציה הראשונית של תאי A ו-B מעורבבים ונותנים להם לנוע. נשאל – לאן המערכת תתכנס?

Cellular Potts Model

כמו הרבה מקרים בביולוגיה, כשרוצים להבין התנהגות מסוימת ליצור אבסטרקציה למערכת הביולוגית המורכבת על ידי מודל פשוט שתופס מורכבות ותכונות בצורה פשוטה יותר. נחשוב על זה כ-grid שבכל מקום שלו יש תאים מסוג A או B:

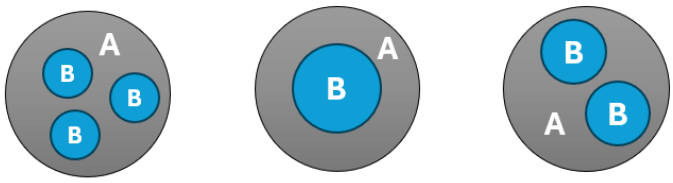


עכשיו מתחיל "משחק החלפות" – כל תא יכול להחליף פוטנציאלית את מיקומו עם אחד משכניו. סדרת צעדים, בכל תא מגרילים רנדומלית תא מהקונפיגורציה, מגרילים רנדומלית את אחד משכניו, ושואלים אותם האם הם רוצים להתחלף. התשובה שלהם תהיה שההסתברות של אותו זוג שנבחר להחליף מיקום (המצב של תא i יתחלף עם של תא j) באופן הבא:

$$\mathbb{P}\big(\sigma(i) \leftrightarrow \sigma(j)\big) = \begin{cases} e^{-\frac{\Delta H}{T}} & \Delta H \geq 0 \\ 1 & \Delta H < 0 \end{cases}$$

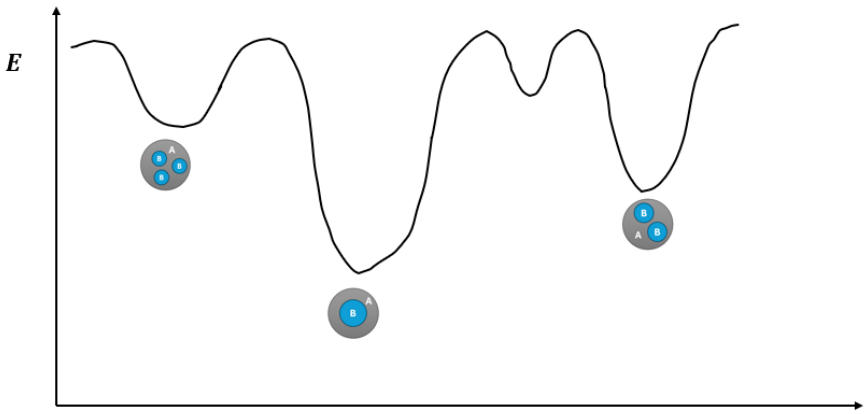
אם ההחלפה מורידה את האנרגיה הכללית של המערכת (נשאל לוקאלית) – מבצעים את ההחלפה. אחרת, החלפה תתבצע בהסתברות $e^{-\frac{\Delta H}{T}}$ כלומר e בחזקת מינוס ההפרש באנרגיה (מהמצב ההתחלתי לסופי) חלקי T (טמפרטורה/רעש, סוג של סקאלת נרמול).

אנחנו יכולים להניח שה-grid יתכנס למצב כמו המבנה שראינו (המסודר) – אבל זה לא תמיד המצב:



וקונפיגורציות נוספות..

בשפה של משטחי אנרגיה/פונקציית Loss מדובר במינימום **לוקאלי**:



לכאורה אם נריץ את הסימולציות, לא מובטח לנו שנגיע לאנרגיה האופטימלית הכי נמוכה במערכת. התהליך הזה לא יהיה reproducable – נתכנס בכל פעם לקונפיגורציה שונה. בסופו של דבר בטבע אנחנו ברוב המוחלט של המקרים מתכנסים לקונפיגורציה הנכונה, אבל על ידי הפשוט של adhesion זה לא באמת קורה – נתקע במינימות לוקאליות, ההבנה הנוכחית שלנו של המערכת הביולוגית לא מסתדרת עם תכונה בסיסית שלה.

כל ההערות שנתנו לאורך המידול רלוונטיות גם למודלים מורכבים יותר ממה שהצגנו, כלומר המורכבות לא נובעת מכך שהמודל "פשוט מדי".

בעיות לא קמורות

רוב בעיות הלמידה המעניינות הן לא קמורות, גם הביולוגיה צריכה להתמודד עם זה – וגם האלגוריתמים שאנחנו יוצרים. נישאר בראש עם התמונה של מערכת ביולוגית שפונקציית ה-Loss שלה לא קמורה, יש גם מינימום לוקאליים. לנסות לחשוב איך המערכת הביולוגית פותרת את זה יכול להיות מורכב אז נעבור לכובע האלגוריתמי – איך אנחנו פותרים בעיית אופטימיזציה לא קמורה?

כדי לצאת מהמינימום הלוקאלי – "נשקשק את המערכת" נזרוק אותה למקום אחר. במערכת ביולוגית זה מעט קשה ליישם דבר כזה (בניגוד למערכות פיזיקליות כמו קופסה עם כדורים).

Simulated Annealing (SA)

טכניקה הסתברותית לשערוך של אופטימום גלובלי בהינתן פונקציה. זו דרך אחרת להימנע מהתכנסות למינימום לוקאלי. נחשוב על חלקיק שמטייל לאורך משטח האנרגיה. מיקום החלקיק מייצג קונפיגורציה של המערכת. נחשוב על אלגוריתם פשוט של Gradient Descent (GD) שבכל פעם לוקח צעד בכיוון מיטבי (לוקאלית) לכן ייתקע במינימום הלוקאלי הקרוב אליו.

הרעיון של SA הוא להוסיף רעש למערכת. בתחילת התהליך נחשוב שהחלקיק מאוד רועש (למשל חלקיק עם טמפרטורה גבוהה), ההבדלים באנרגיה עוברים scaling על ידי הטמפרטורה מהאופן בו הגדרנו למעלה את $\mathbb{P}\big(\sigma(i) \leftrightarrow \sigma(j)\big)$.

מדי פעם החלקיק "מורעש" מספיק כך שהוא מצליח לצאת מהמינימה הלוקאלית ויכול להגיע למינימה הגלובלית. תוך כדי התהליך מידת הרעש (הטמפרטורה) של החלקיק הולכת וקטנה. נשים לב שזה אומר שההסתברות לקבל או לא לקבל מעבר – תלויה בזמן ($e^{-\frac{\Delta H}{T}}$).

ככל שנהיה יותר מוקדמים בתהליך הדינמי, נטה יותר לקבל מעברים גם אם הם פחות טובים אנרגטית.

נבחין כי זה שקול להגיד שפונקציית ה-Loss/משטח האנרגיה משתנה בצורה הדרגתית לאורך זמן. אפקטיבית אפשר לחשוב על משטח אנרגיה שטוח שבאופן הדרגתי על פני ציר הזמן הוא משנה בצורה לינארית את צורתו – והוא בעצם מתעוות באופן רציף ולינארי בזמן.

חזרה למערכת הביולוגית

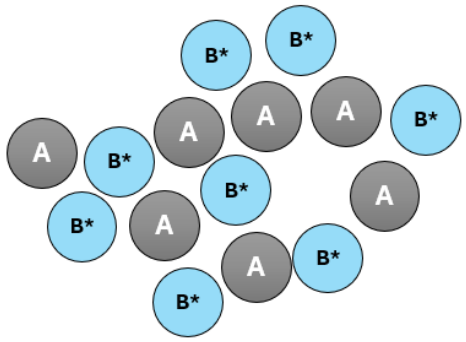
אולי בעצם קיים פה מנגנון שנכון למערכות פיזיקליות או אבסטרקציה – ונכון גם למערכות ביולוגיות? אם נוכל למצוא את המנגנון הזה האם נוכל למנוע מהמערכת להתכנס למינימות לוקאליות?

מה הבעיה? המנגנון של SA כמו שהוא מאוד מורכב למימוש ביולוגי, או ביולוגי בצורה סינטטית. נדרשה זווית אחרת: יצירת מסגרת שהיא הכללה של SA שבה התמונה המחשבתית זהה, אבל מורידים את האילוץ הלינארי, כלומר לא נחייב שכל חלק במשטח האנרגיה ישתנה באותו קצב כמו האחרים.

זה אומר שאנחנו אולי נוכל לפתור את הבעיה אם המערכת הביולוגית עצמה תוכל להשתנות תוך כדי הארגון הספונטני לצורה התלת-ממדית, וזה נשמע כמו משהו שקורה כל הזמן. תהליך דיפרנציאציה משמעותו אפשר להתחיל עם סט מסוים של תאים, ובמהלך ארגונם במרחב הם ישנו את התכונות שלהם.

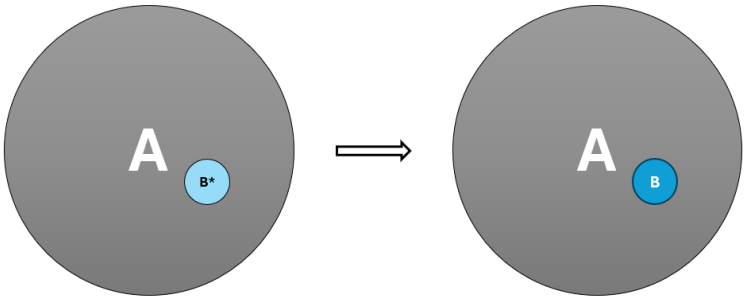
נדבר על אלטרנטיבה שבה אנו מתחילים רק עם תאים מסוג A. אין הבדל בין התכונות, לכן השאלה של מצב האנרגיה אחרי חילוף פוטנציאלי חסרת משמעות כרגע – כי איננו מבדילים בין תאים. אפשר לחשוב על המשטח במצב זה כשטוח (כל קונפיגורציה תהיה שקולה אנרגטית).

יש קאץ' – באלטרנטיבה הזאת יש גם תאי B* עם תכונות בדיוק כמו A. תאי A ותאי B* נמשכים אחד לשני, ו-A יכול לשנות את תכונות B* כך שיהפוך להיות תא מסוג B.

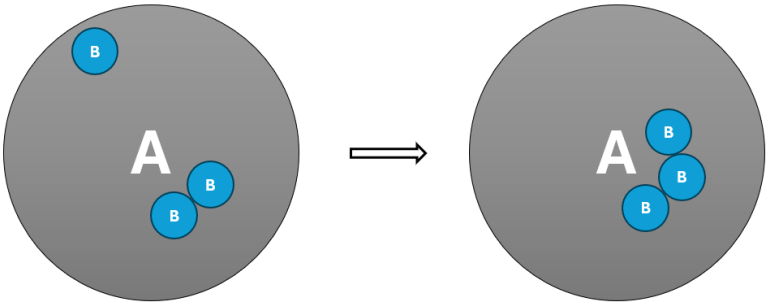


כש-A פוגש תא *B יש לו הסתברות מסוימת להפוך אותו לתא B עם תכונת הדביקות לתאי B אחרים.

למה זה עוזר לנו? זה מסייע לנו למדל את התהליך של שינוי מרחב האנרגיה בצורה לא לינארית. אנחנו יודעים שבאופן הדרגתי ורציף אנחנו מעוותים את המשטח כתלות בשינויים מתאי *B לתאי B. איך המידול הזה מעלה את הסיכוי להתכנס למינימום גלובאלי? ניקח מקרה קיצון שבו קצב האקטיבציה של תאי B כל כך נמוך שאנחנו מגיעים למצב הבא:



כלומר cluster ענק של תאי A, ובתוכו תא *B אחד שעומד לעבור אקטיבציה. תא B כזה יטייל בין תאי ה-A ולא יבחין בשום דבר חריג. אחרי כמה זמן עוד אחד עובר אקטיבציה – והם בסוף מוצאים זה את זה. אחרי כמה זמן שוב..



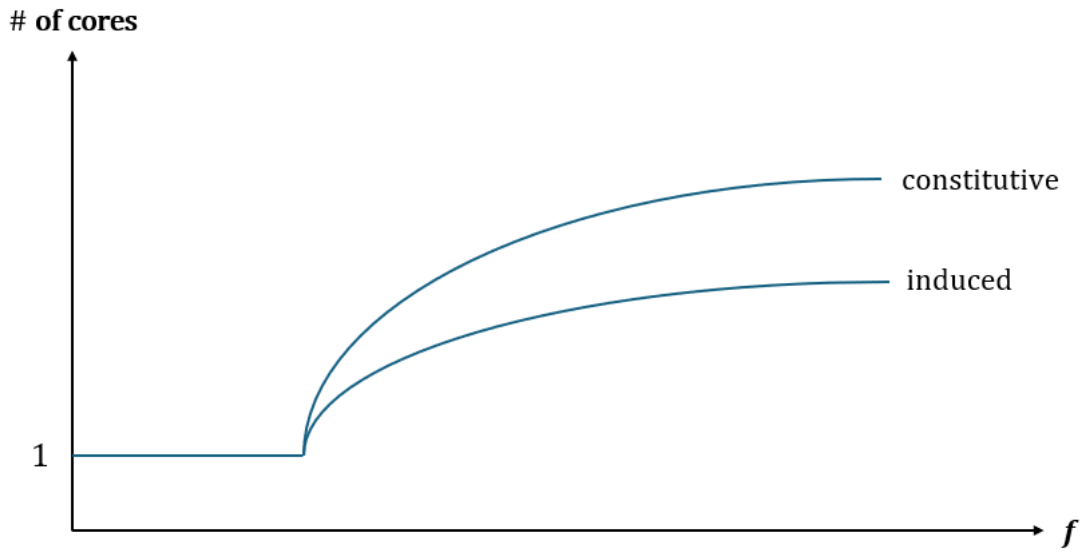
נשים לב שלצמד התאים יותר קשה לזוז במרחב, אז סביר שמי שימצא אותם יהיה תא ה-B הבודד. כך אנחנו מקבלים תהליך הדרגתי (מזכיר תהליך של יצירת גבישים). ההנחות האלה נכונות גם כשהקצב יותר מהיר (בלי הנחת מקרה הקיצון של קצב אקטיבציה איטי מאוד).

באופן יותר כללי, אפשר לראות גם ממודלים ריאליסטיים וגם מהילוכים מקריים (שלא נכנס אליהם), שמהירות (S) תנועה של B cluster בקונפיגורציה הכוללת עם שטח מסוים בדו-ממד (עבור תלת-ממד נתאר נפח), ניתנת לתיאור באופן הבא:

$$S \sim \left(\frac{1}{A}\right)^f$$

כאשר A הוא השטח (ניתן לחשוב עליו כמספר התאים בתוך cluster), באקספוננט כלשהו f.

אם נשרטט את מספר ה-cores של תאי B במערכת (באשר האופטימום הוא ב-1), נקבל תמונה כזו:



כאשר המערכת ה-induced היא זו שהצגנו מאוחר יותר בהרצאה (עם תאי *B) והמערכת ה-constitutive היא זו שהצגנו בתחילת ההרצאה. ככל ש-f גדל (כלומר ההשפעה של גודל ה-cluster על המהירות גדלה) רואים הבדל שהולך ונעשה יותר ויותר משמעותי בין מספר ה-cores במקרה המקורי למקרה ה-induced שבו משנים את המערכת תוך כדי ההתכנסות. אינטואיטיבית, זה ש-f נהיה יותר גדול אומר של-clusters גדולים נהיה יותר קשה לנוע במרחב.

בין וריאנטים שונים של מידול המערכת הפרמטר שמשתנה הוא f, שנקבע מחוקים פיזיקליים בסיסיים של המערכת. עבור f נתון נוכל לסמלץ את המערכת הרבה פעמים – בוורסייה המקורית לעומת המערכת שמתחילה כהומוגנית ומשתנה, ולשאול בכל פעם כמה clusters יש.

בסוף, אפשר לספר לנו שאחרי מחקר המודל ה-induced עוזר לנו להתכנס למינימום נמוך יותר, להתכנס למינימום גלובלי, קונפיגורציה יציבה יותר של ה- cores וכן הלאה.

סיכום

בהרבה מקרים זו התוצאה שאנחנו רואים: המנגנון של ה-induction (שאכן קורה בביולוגיה, תהליך התמיינות) – עוזר למערכת למצוא קונפיגורציה יציבה יותר (מינימום נמוך יותר).

האם זה קורה במערכות ביולוגיות? בדיקה של מנגנונים שונים בצורה מבוקרת זה תהליך מורכב בקונטקסט של מערכת ביולוגית, אבל כן אפשר לבצע בצורה סינטטית. אפשר להסתכל על דינמיקה של תאים אמיתיים והתכנסות למבנה תלת ממדי במעבדה, כשמה שמפריד ביו הגישות (constitutive ו-induced) זה הינדוס של תכונות של התאים, שמאפשר להם לעבור induction או לא.

בגישה ה-constitutive

שני תאים A ו-B שאם נותנים להם להסתדר רנדומית במרחב הם נדבקים אחד לשני ב-cluster לא מוגדר של תאים, אבל אם אחד מהם מבטא חלבונים שמכסים את התא מקבלים מודל שדומה לגישה הזו, ו-B נעשה דביק יותר כמו שהגדרנו. כדי לעקוב אחרי תאים אלו צבעו אותם בצבעים פלורסנטיים שונים ובעזרת שיטה מיקרוסקופית שמאפשרת לקבל תמונה תלת-ממדית עקבו אחריהם לאורך 21 שעות ברווחים של שעה.

זה אפשר לחוקרים לבחון את הארגון הדינמי של התאים ברזולוציה של תא בודד וברזולוציה גבוה בזמן, ולבחון את ההתארגנות – תוך ידיעה מי היה תא A ומי היה תא B.

בחנו ב-40 ניסויים ורסיות שונות (ratios בין התאים וכדומה). התוצאות היו התכנסויות שונות כתלות בפרמטר הנבדק, כאשר כ-**10% בלבד** היו ל-core יחיד.

בגישה ה-induced

כאן כמו שתיארנו יש תא A ותא *B שהופך לתא B אחרי אינטראקציה עם A. התאפשר בעזרת הנדוס התאים כך שבעת התחברות בין תא A לתא *B, תא *B מייצר ביטוי למה שגורם לו להפוך לתא B דביק. באותן שיטות מעקב עם ניסוי דומה רואים שב-**75%** מהניסויים התכנסו למבנה של core יחיד.

מסקנה עיקרית

במערכת הזו, בדומה לאינטואיציה במערכות מתמטיות, עלייה הדרגתית במורכבות המערכת תוך כדי התהליך מובילה להתכנסות רובסטית יותר.

היום מנסים לחפש את התהליך הזה במקומות שונים, מעבר לגבולות מה שתיארנו בהרצאה.

שבוע 9 – הרצאה 16

מודלים של זמן בעצים פילוגנטיים

תזכורת:

$$\mathbb{P}_{a,b}^{(t)} = \mathbb{P}(a \xrightarrow{t} b)$$

$$\mathbb{P}^{(t)} = e^{tR}$$

$$\mathbb{P}^{(0)} = I$$

$$\mathbb{P}^{(t)} \xrightarrow{t\to\infty} \begin{bmatrix} 1/4 & \cdots & 1/4 \\ & \vdots & \\ & & \end{bmatrix}$$

דיברנו על מודל Jukes-Cantor שבו ה-rate היה המטריצה:

$\boldsymbol{R}$	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

$\boldsymbol{\alpha \neq 0}$

כמו כן הגדרנו:

$$\mathbb{P}_{a,a}^{(t)} = ¼(1 + 4e^{-4\alpha t})$$

$$\mathbb{P}_{a,b}^{(t)} = ¼\,(1 - e^{-4\alpha t})$$

Kimura Model

יש מוטציות מסוג transition ומסוג transversion:

מוטציית transition: כאשר אנחנו משנים $A \leftrightarrow G$ או $C \leftrightarrow T$.

מוטציית transversion: זה ההחלפה בין $A \leftrightarrow C$ או $G \leftrightarrow T$.

עכשיו נוסיף עוד פרמטר, α ל-transition ו-β ל-transversion. נקבל עדכון למטריצה:

$\boldsymbol{R}$	A	C	G	T
A	$-\alpha-2\beta$	β	α	β
C	β	$-\alpha-2\beta$	β	α
G	α	β	$-\alpha-2\beta$	β
T	β	α	β	$-\alpha-2\beta$

מה שיש על האלכסון נובע מהעובדה שיש דרך אחת לעשות transition ושתי דרכים לעשות transversion.

אם $\alpha=\beta$ חזרנו ל-JC, וההנחה היא שבדרך כלל $\alpha>\beta$ – כלומר transitions מעט סבירים יותר מ-transversions.

$$\mathbb{P}^{(t)}_{\text{transition}}=\frac{1}{4}\cdot\left(1+e^{-4\beta t}-2e^{-2(\alpha+\beta)t}\right)$$

$$\mathbb{P}^{(t)}_{\text{transversion}}=\frac{1}{4}\cdot\left(1-e^{-4\beta t}\right)$$

$$\mathbb{P}^{(t)}_{\text{no change}}=1-\frac{1}{2}\left(1-e^{-4\beta t}\right)-\frac{1}{4}\left(1+e^{-4\beta t}-2e^{-2(\alpha+\beta)t}\right)$$

מספר השינויים לו אנחנו מצפים הוא לא לינארי לאורך זמן, מספר השינויים גדל מהר יותר. (תזכורת - גרף בהרצאה הקודמת בנושא). עם הזמן יכול להיות שנחזור לאותה האות שאנחנו כבר נמצאים בה. חוסר השינויים יורד עם הזמן, אבל לא יירד מתחת ל-¼. נצטרך להסתכל על מודל שמתרגם שינויים ליחידות של זמן.

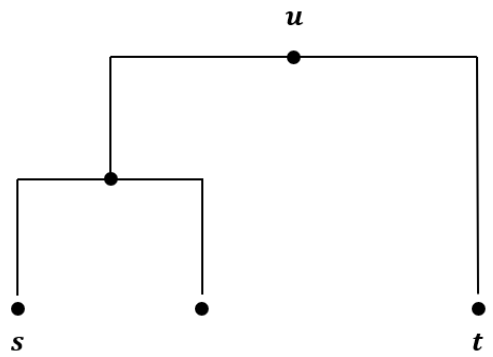
איך נשתמש בזה?

דרך אחת פשוטה להשתמש בזה היא בהינתן שני רצפים לדעת כמה זמן עבר ביניהם:

s	A	C	A	T	T	A	G
t	A	G	T	T	A	A	G

במקרה של JC אנחנו יודעים שמה שבתוך המכפלה אפשר להציב את $\mathbb{P}_{a,a}^{(t)}$ ואת $\mathbb{P}_{a,b}^{(t)}$ ולקבל איך שינוי ב-t משנה את $L(t)$. באינסוף t מתכנס ל-¼ בחזקת אורך הרצפים. אפשר לשאול איזה t נותן maximum likelihood לשני הרצפים שלנו.

כך אפשר לחשב אורכי ענפים, כמה זמן עבר בין זוג עלים. אילו בעיות יש במודל הזה? לא משנה אם מדברים על עץ מכוון או לא מכוון. היינו מצפים שהמרחק ביניהם הוא המרחק לאב הקדמון המשותף אל s ועוד המרחק לאב הקדמון המשותף אל t: $d_{s,u} + d_{u,t}$



האם זה שווה למרחק שאם נכניס את העץ בגרסה הזו?

מה קורה כש-t שואף לאינסוף? הזמן מתנהג כמו שרשרת מרקוב עם צעדי אפסילון. כמעט בכל שרשרת מרקובית נגיע למצב שבו נשכח מאיפה הגענו. אפשר לבנות שרשרת מרקובית שהולכת במעגלים, אבל לא במקרה הזה נתמקד.

Stationary Probability

אחרי שנחכה יותר מ-0 זמן יש סיכוי כלשהו לעבור מ-a ל-b, הוא יכול להיות קטן אבל קיים, אז הגבול של t באינסוף הוא כך:

$$\text{if} \quad \forall t > 0 \; \forall a, b \quad \mathbb{P}_{a,b}^{(t)} > 0$$

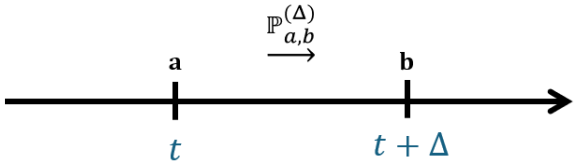
$$\lim_{t\to\infty} \mathbb{P}^{(t)} = \begin{bmatrix} - & \pi & - \\ - & \pi & - \\ & \vdots & \end{bmatrix}$$

$$\pi(a) = \lim_{t\to\infty} \mathbb{P}_{b,a}^{(t)}$$

יש פה הנחה לחוסר תנועה אבל זה לא שהערכים לא משתנים אלא שההסתברות שלהם לא משתנה. מה שזה אומר זה שאם נחכה מספיק זמן ונדגום את המצב של הדברים נראה את a בהסתברות $\pi(a)$ ללא קשר לאיפה היינו קודם, זה לא אומר גם שאם נחכה עוד קצת זמן נראה שוב את a כי ראינו אותו קודם.

תכונות של Stationary Probability

נסתכל על ציר הזמן, באיזושהי נקודה t נשאל איפה אנחנו נמצאים ובנקודה אחרת $t + \Delta$ נשאל שוב:



אפשר לכתוב את זה גם בצורה הבאה:

$$\mathbb{P}(x_{t+\Delta} = b) = \mathbb{P}(x_t = a) \cdot \mathbb{P}_{a,b}^{(\Delta)}$$

אפשר לכתוב גם במונחים מטריציוניים:

$$\pi(b) = \pi(a) \cdot \mathbb{P}_{a,b}^{(\Delta)}$$

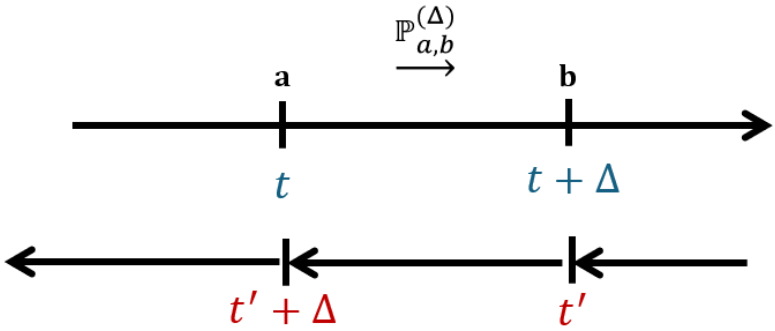
$$\vec{\pi} = \vec{\pi} \cdot \mathbb{P}^{(t)} \quad ; \quad \vec{\pi} = \left[\mathbb{P}^{(t)}\right]^T \cdot \vec{\pi}$$

כדי להיות נייח אתה צריך להיות וקטור עצמי, ואפשר להוכיח שיש רק אחד כזה - והוא מתאר את ההתפלגות.

רברסיביליות בזמן

אם ניקח סרט של משהו קורה, נצטרך לתאר האם הוא מתקיים עם כיוון הזמן או נגד כיוון הזמן. למשל כדור לא נוכל לדעת כשהוא קופץ באיזה כיוון הסרט מופעל - ביצה שנופלת לרצפה לעומת זאת, ברור לנו באיזה כיוון הסרט רץ.

תהליך הוא רברסיבילי בזמן אם היה אפשר לתאר את ציר הזמן ממקודם רק לכיוון ההפוך:



ואז מקבלים משוואה:

$$\mathbb{P}^{(t)} \text{ is reversible if:}$$

$$\forall a,b,\Delta \quad \pi(a)\cdot \mathbb{P}^{(\Delta)}_{a,b} = \pi(b)\cdot \mathbb{P}^{(\Delta)}_{b,a}$$

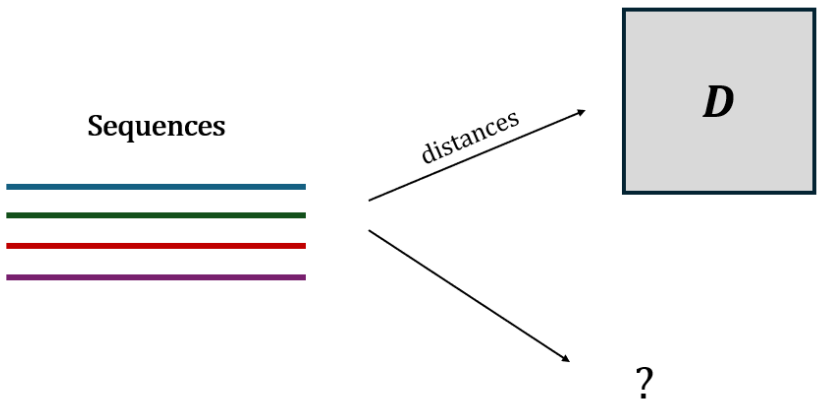
כלומר אי אפשר להבדיל מה כיוון הסרט, שני האירועים בשני הכיוונים היא עם אותה ההסתברות. זה מוביל לכך שהנראות מ-s ל-t או מ-t ל-s היא אותו הדבר.

בדרך כלל נניח שההסתברות היא reversible. מצד אחד נוח מתמטית, מצד שני יש את אלו שטוענים שזו אינה הנחה ריאלית.

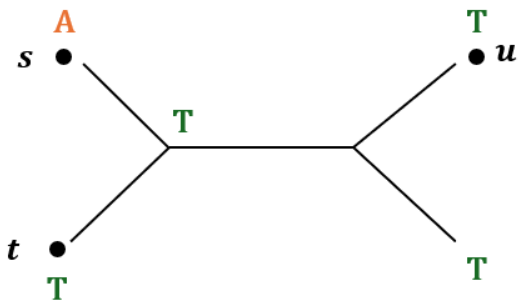
<u>חזרה לנראות</u>:

$$L(t) = \prod_i \pi(s_i)\cdot \mathbb{P}^{t}_{s_i,t_i}$$

כשהתחלנו לדבר על עצים, אמרנו שגישה אחת זו לשחזר את כל הרצפים ומשם לשחזר עץ - ושיש גישה נוספת. היום נראה את הגישה הנוספת, יתרונותיה וחסרונותיה.



נניח שהעץ לא מכוון (כי הזמן רברסיבילי):

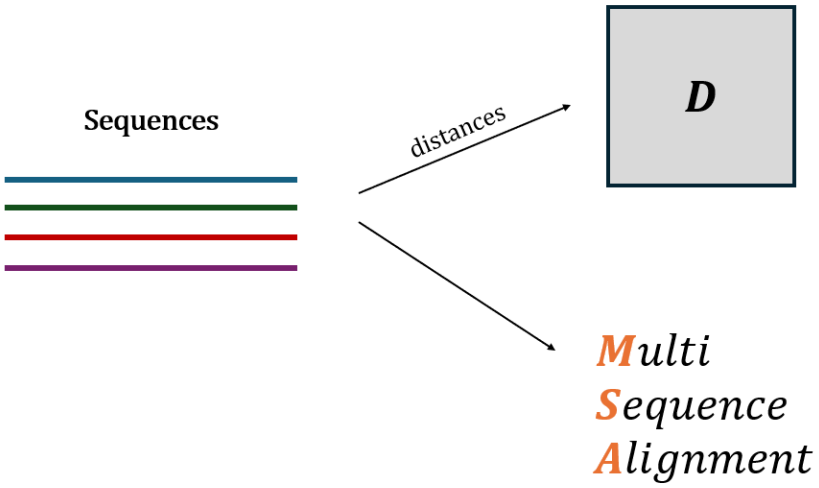


כשאנחנו מחשבים את המרחק מ-s ל-u אנחנו מניחים שקרו דברים באבולוציה על המסלול המסומן. אפשר להגדיר את המרחק בין שתי נקודות בעזרת הסתברות שלמה לסכום על כל הנקודות שביניהם. כשמחשבים את המרחק בין t ל-u אנחנו לא לוקחים עדיין בחשבון את העובדה שלשניהם יש אב קדמון משותף שהם חייבים לעבור דרכו. אם בענף אחד היה T ובשני A אנחנו מחזקים את הסבירות שהשינוי קרה ל-s אחרי הפיצול. לעומת זאת אם אנחנו רואים שינויים בין A ל-T לאורך כל העץ, יהיה יותר קשה לתת הנחה כזו.

האינטואיציה היא שאם נדע את העץ ואת האותיות שמתאימות לכל עמדה, היינו יכולים לתת את הנראות בצורה יותר מעניינת.

Multiple Sequence Alignment - MSA

כאן נתמקד בחלק השני שעוד לא ראינו:

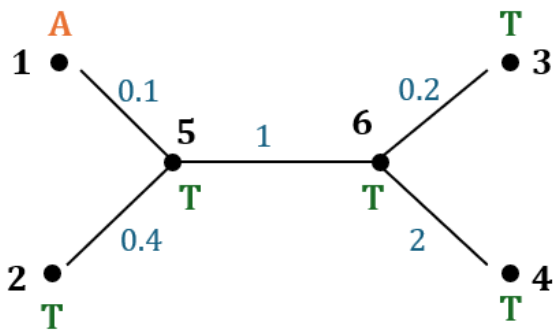


נתבונן בדוגמה הבאה:

s_1	A	-	T	G
s_2	A	T	T	G
s_3	T	T	T	-

אם יש לנו אלגוריתם למציאת MSA, נוכל לחשב נראות של עצים. בהינתן MSA נניח שכל האותיות באותה עמדה הן בעצם עלים בעץ. בהינתן MSA נוכל לחשב $L(T)$ שזה אובייקט שמתאר עץ. נוכל לחשוב על תהליך אופטימיזציה שמחפש עץ עם הנראות הכי גבוהה.

נניח שכל העמדות בלתי תלויות, ונשאל מה הנראות של עמודה אחת. מניחים ש-T מתאר טופולוגיה וענפים. דוגמה:



אנחנו רוצים לחשב את הנראות של העץ הזה. נניח שנתנו לנו את הערכים הבאים:

$$x_1 = \mathrm{A} \quad ; \quad x_2 \dots x_6 = \mathrm{T}$$

אז נוכל לחשב את ההסתברות לכך בהינתן למשל שהשורש הוא 1:

$$\mathbb{P}(x_1 = \mathrm{A}, x_2 = \mathrm{T}, \dots, x_6 = \mathrm{T}) = \pi(x_1) \prod_{i \neq 1} \mathbb{P}_{x_i, x_{a(i)}}^{(t_i, a(i))}$$

אם ההסתברויות רברסיביליות אפשר בעזרת טיולים להזיז את השורש לכל מקום, לכן במצב הזה מיקום השורש לא ישנה.

הבעיה היא שבמודל הזה אנחנו מניחים שאנחנו יודעים מה קורה בקודקודים הפנימיים, מה שלא ריאלי למציאות. לכן צריך לעדכן את זה לפי המידע שכן בידינו (בדומה למה שעשינו עם מצבים חבויים ב-HMM) ולקבל:

$$\mathbb{P}(x_1, \dots, x_4) = \sum_{x_5} \sum_{x_6} \mathbb{P}(x_1, \dots, x_6)$$

בעצים גדולים זה כמובן יתחיל להסתבך. גם כאן נשתמש בתכנון דינאמי, ובמקום Forward Backward נשתמש ב-Up and Down. הרעיון: אם יש קודקוד, מתחתיו עלים ומעליו המשך, אז יש עלים שעוברים דרך קודקוד i בדרך לשורש, ועלים שלא עוברים דרך קודקוד i בדרכם לשורש. אם ל-i יש שני צאצאים j, k ונשים את העץ מעל i נוכל לחשב את העלים שמתחת ל-i ומתחת ל-k ולחשב אותם באופן בלתי תלוי.

$$UP_{i\to l} = f\left(UP_{j\to i}, UP_{k\to i}\right)$$

$$UP_{i\to j} = f(UP_{l\to i}, UP_{k\to i})$$

תמיד שני up יוצר up אחד חדש. בהינתן עץ נוכל לחשב את הנראות בעזרת תכנון דינאמי. כמו כן, איפה אנחנו יודעים את הטופולוגיה ואורכי ענפים? בהינתן טופולוגיה חישוב אורכי הענפים זו בעיית אופטימיזציה (מזכיר EM).

מציאת העץ הטוב ביותר זו בעיה שהוכחה כ-NP קשה. מה בכל זאת נוכל לעשות? היוריסטיקות, נבנה עץ "לא מושלם". היוריסטיקה אפשרית היא להתחיל עם עץ של 4 מינים, להחליט איפה להוסיף את החמישי ומשם להתקדם. הבעיה עם זה היא שאחרי שמיקמנו את המין החמישי הוא לא יזוז גם אם מיקמנו אותו בצורה לא מיטבית.

היוריסטיקה נוספת היא להתחיל עם עץ כלשהו ובהמשך לשנות מיקומים - גזימת ענפים וחיבורם במקום אחר. כשננתק ענפים נחבר אותם בצורה אקראית (לכן היוריסטיקה), ונמשיך עם האופציה רק אם החיבור החדש משפר את הנראות. מזכיר מעט Gradient Descent, אבל כאן אנחנו מדברים על עולם דיסקרטי ולא רציף. עץ התחלתי טוב הוא neighbour joining שהרצנו על המרחקים.

אלגוריתמים למציאת MSA

יש הרבה אלגוריתמים לכך. זו בעיית היוריסטיקה, ואת ה-MSA הטוב ביותר נמצא כאשר אנחנו יודעים מה העץ. נשתמש באלגוריתם רקורסיבי לבנות MSA במעלה עץ כשבכל שלב אנחנו משווים דברים קרובים יחסית. אם יתנו לנו MSA נבנה עץ טוב, ואם יתנו לנו עץ טוב נמצא MSA - בעיה של ביצה ותרנגולת. אלגוריתמים טובים משחקים באיטרציות בין השניים.

אחד הדברים שאפשר לעשות בהינתן עץ זה לשאול מה האב הקדמון המשותף במשפחה של אנזימים שעושים דברים שונים (בנקודת הזמן הנוכחית).

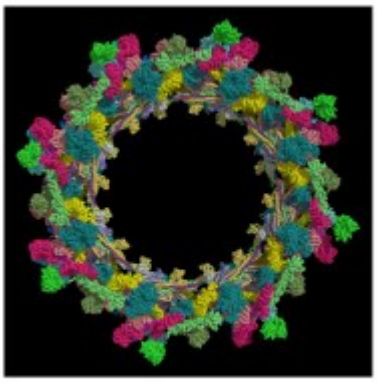
שבוע 9 – הרצאה 17 (הרצאת אורח)

הרצאת אורח: פרופ' דינה שניידמן, חיזוי מבנה חלבונים

מבנים מולקולריים

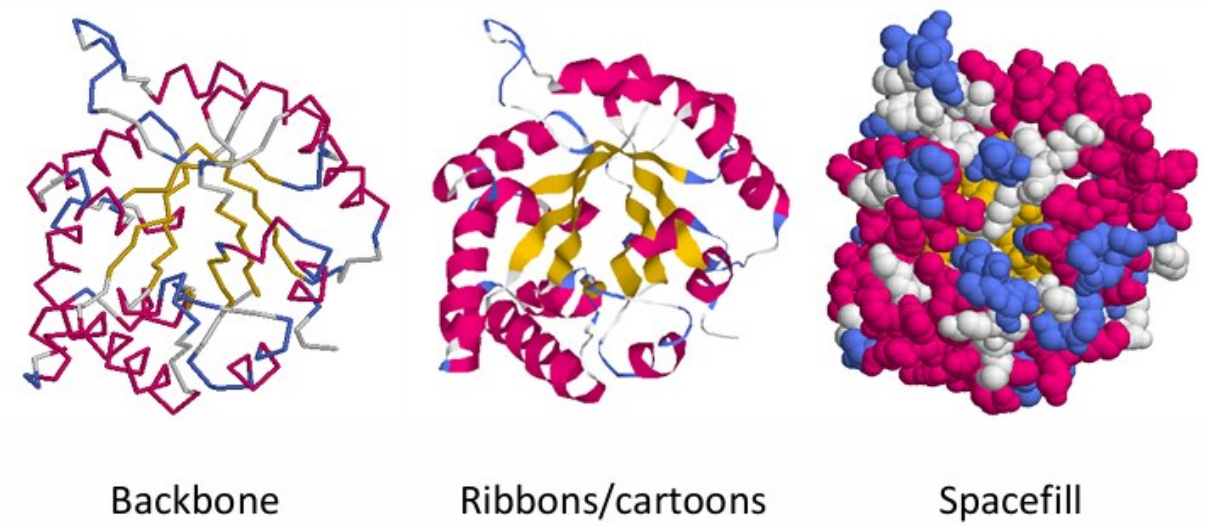
מוטיבציה

למה אנחנו אוהבים מבנים חלבונים או מולקולריים באופן כללי? להבין מה הפונקציונליות של המכונות, איך הן פועלות, ריצופים, אבולוציה של המכונות (נוסף חלק, ירד חלק), פיתוח של תרופות, חלבונים חדשים. כמו כן, מבנה הרבה יותר שמור מרצף, וזה מוביל אותנו להבנה טובה יותר של הפונקציה. דוגמה:



NPC inner ring
(Kosinski et al. 2016)

ישנם ייצוגים שונים למבנה:



הייצוג שראינו בהתחלה הוא מסוג spacefill. כל אטום מיוצג על ידי כדור, זה ייצוג שקל להבין – חומצות אמינו כוללות בממוצע 10 אטומים וכל אטום מיוצג על ידי כדור במבנה. יש קואורדינטת x,y,z לכל אטום. בצורות

האחרות רואים או את ה-backbone או את ה-ribbons שבאמצעותו מזהים את ההליקסים שרואים ב-backbone ובעזרת החלקה הופכים אותם לנעימים יותר לצפייה. מדובר בדרכים שונות לצייר את אותו המבנה.

סקאלה

תא: $1 \times 10^{-6}\ m$ (מיקרונים).

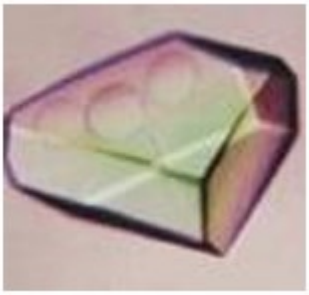
ריבוזום: $1 \times 10^{-9}\ m$ (כ-20 ננומטר) – מדובר במכונה גדולה בתא.

חלבון: $1 \times 10^{-10}\ m$ (כ-20 אנגסטרום).

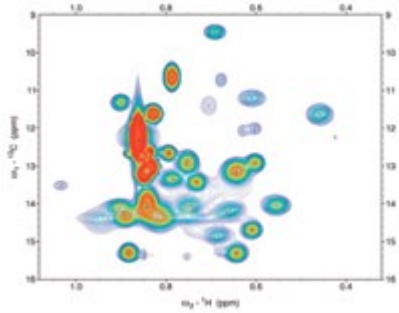
שיטות לדטרמינציית מבנים

לא ניתן לראות את המבנים האלה במיקרוסקופים רגילים, השיטות הנפוצות לראות אותם הן קריסטלוגרפיה בעזרת X-Ray (שיטה בת 60+ שנה) ספקטרוסקופיית NMR או מיקרוסקופיה אלקטרונית.

X-ray crystallography



NMR spectroscopy



Electron Microscopy

עבור קריסטלוגרפיה: צריך בעצם לסדר חלבון ב"צנצנת" וזה מגביר את הסיגנל, לכן אפשר לראות באמצעות *X-Ray*. הקושי כאן הוא לגבי קריסטלוגרפיה: לנקות את החלבון, להפיק אותו בכמות מספקת ולסדר אותו בגביש.

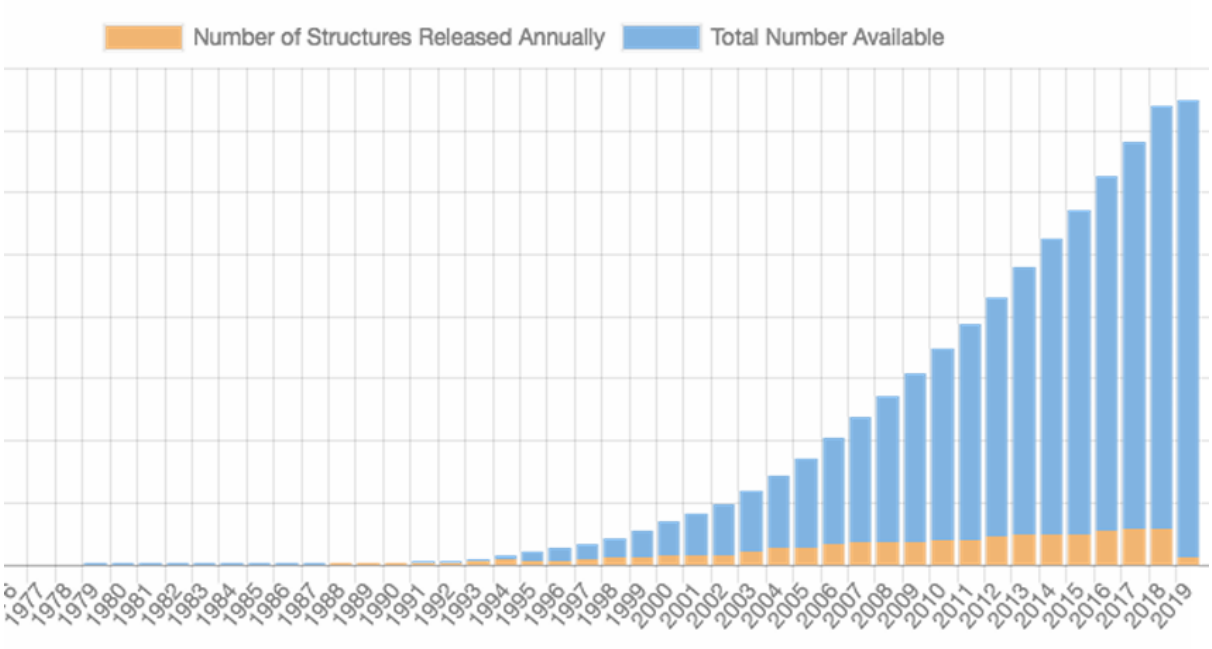
ספקטרוסקופיית NMR: בשנות ה-90 הופיעה ספקטרוסקופיית NMR שמוגבלת בגודל (מתאימה לדברים יחסית קטנים), ודברים דינמיים – לא קל כי צריך לפענח את הגרף שראינו מעל.

מיקרוסקופיית אלקטרון: פריצת הדרך הייתה בערך ב-2014 – resolution revolution, הרזולוציה הפכה מספיק גבוהה כדי להסתכל על מבנים. כבר יותר מ-10 שנים המיקרוסקופים משתדרגים בזכות detector חדש וגם image processing.

כמה מבנים פתורים יש היום?

לפי מאי 2024 – 219,140. הכל נכנס למאגר שנקרא PDB (Protein Data Bank), שיש בו המון יתירות (בהורדת היתירות נגיע לכ-50/60 אלף מבנים). אם רוצים לעשות ML חייבים המון דאטה שהמאגר הזה מספק לנו.

מספר המבנים גדל בצורה לינארית:



לפני הרבה שנים כבר התחילו לחשוב איך אפשר לנצל את כל הרצפים האלה לטובת structure prediction. ה-raw data מאפשר לחזור ולחשב מחדש מבנים כשמתפרסמת שיטה חדשה, זה לא טריוויאלי בכלל שקיים DB כזה.

Protein Structure Prediction

הגדרת הבעיה

בהינתן רצף, נרצה להגיע למבנה: קואורדינטות שייצגו לנו מבנה תלת ממדי.

Structure prediction – לגלות את צורת המבנה, Folding – איך ממצב פתוח החלבון מתקפל למצב סגור. היום לא נדבר על בעיית ה-folding (היא לא כל כך פתורה, לעומת structure prediction. הדאטה הוא של מבנים מקופלים).

נתמקד באיך החלבון יודע את המבנה התלת ממדי שלו ומתקפל אליו.

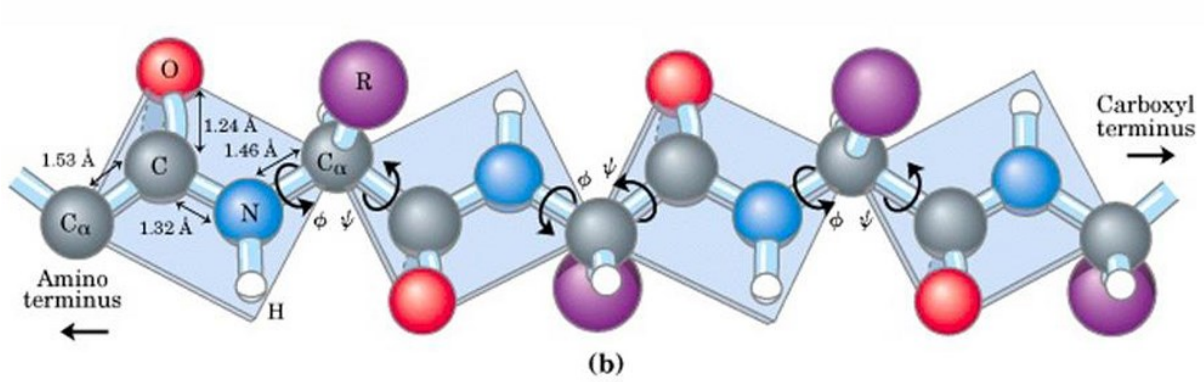
Anfinsen Experiment, 1961

הניסוי הראה שאפשר לקחת חלבון, לפתוח אותו – והחלבון ידע להתקפל חזרה לבד בתוך תמיסה בלי עזרים נוספים. כלומר – כל האינפורמציה שהוא צריך נמצאת בתוך הרצף. זה אומר לנו שאנחנו יכולים לעבוד על הבעיה רק בהינתן רצף כקלט.

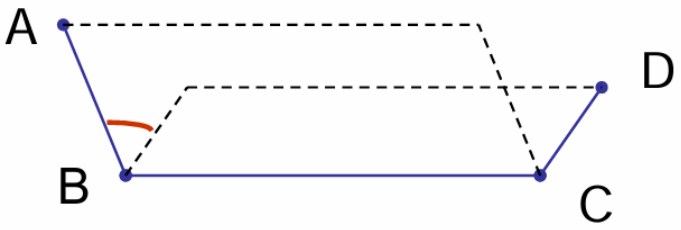
דרגות חופש

ב-backbone יש בכל חומצה אמינית את המשולש (N, C_α, C) – על ה-C_α "תולים" את ה-side chain שמפריד בין חומצות אמינו. בין כל זוג כזה יש torsional angle שמוגדרת על ידי ארבעה אטומים.

יש במולקולות דרגות חופש, בכל חומצה אמינית אנחנו מסתכלים על 3 אטומים מקובעים ועוד אחד מסתובב שמאפשר את דרגת החופש. יש עוד 2 זוויות עיקריות, בשרטוט ϕ, ψ:



יש זווית שלישית שנקראת ω אבל היא תמיד 180°. את זווית הסיבוב בין המישורים נגדיר כך:



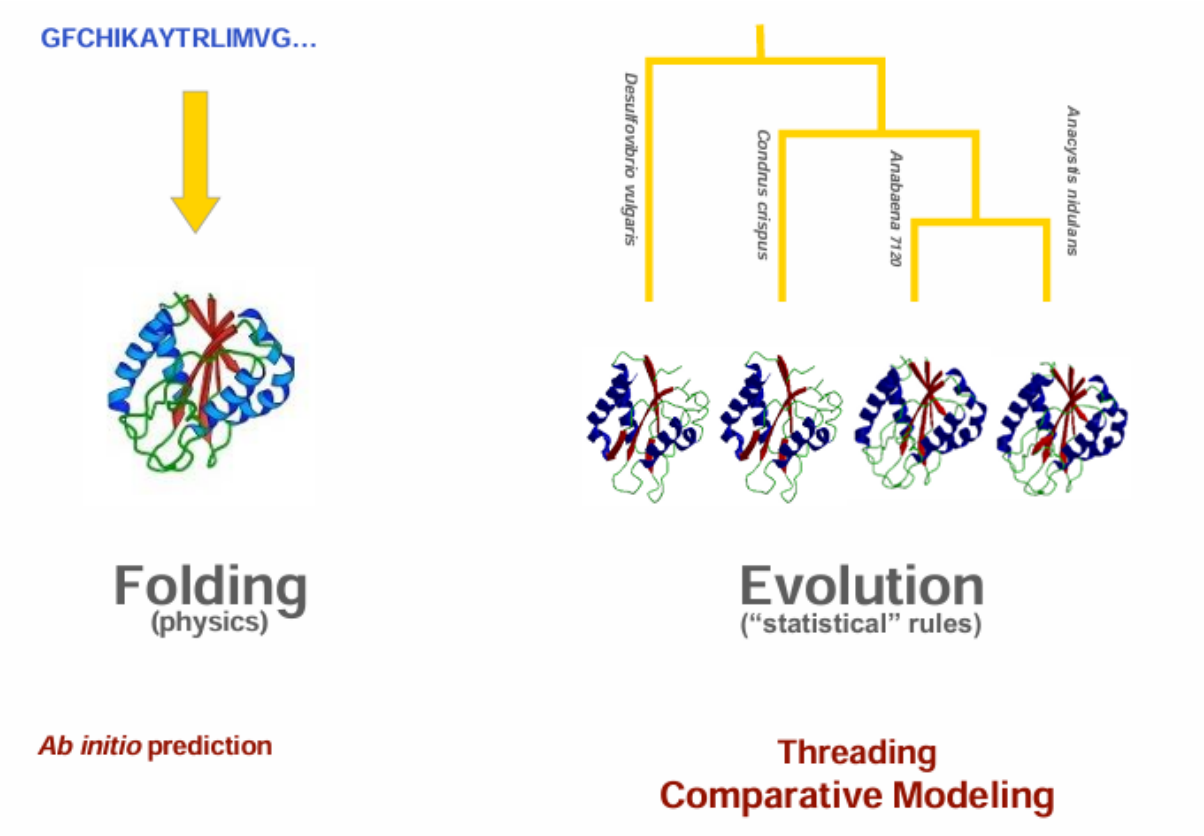
<u>יש כמה דרכים לחשב את המבנה:</u>

- אנחנו יכולים לייצג את הווקטור בעזרת קואורדינטות x,y,z של N, C_α, C (9 פרמטרים באמצעותם נחשב את הזויות ϕ, ψ)
- אפשר גם להסתדר עם הקואורדינטה של C_α בלבד בשילוב הכיוון (6 פרמטרים)
- רק בעזרת הזויות ϕ, ψ (2 פרמטרים)
- שימוש במרחקים בעזרת מטריצת המרחקים בין כל זוג אטומים (ייצוג בזבזני אבל קל להמרה)

עקרונות החשיבה על הבעיה

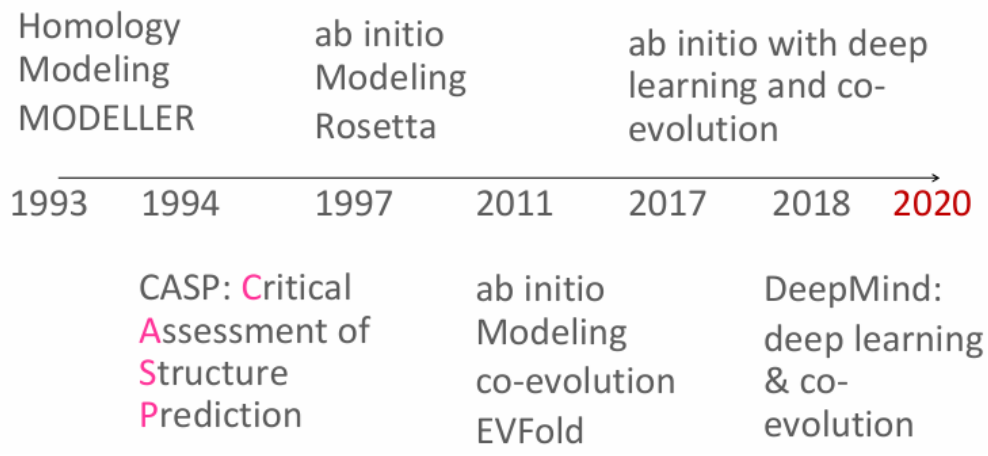
היו כאלה שאמרו שהקיפול נובע מכוחות פיזיקליים, אפשר לחשב אותם ("Ab initio prediction").

אחרים אמרו שיש תהליך אבולוציוני ובאורגניזמים שונים נקבל מבנה דומה לאותה פונקציונליות, אפשר לקבל את התוצאות בעזרת כלים סטטיסטיים. ההנחה הזו שרצף דומה מתקפל למבנה דומה (Threading" Comparative Modeling") אומרת שאפשר למדל על ידי העתקה של מבנה אחר.

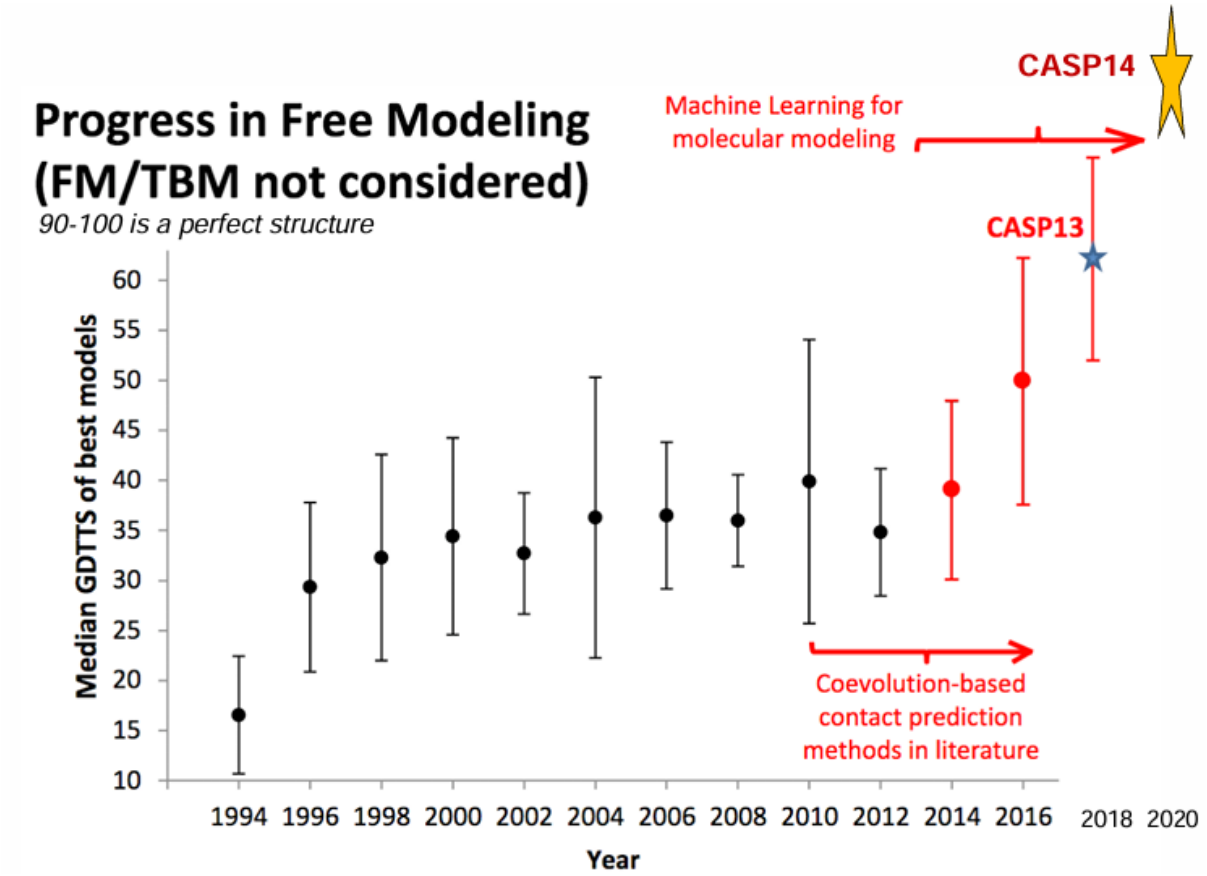


היו בשנות ה-70 אנשים שידעו לקפל חלבון ופרסמו מאמרים, היה רצון לבדוק האם באמת יודעים לקפל על ידי ניסוי עיוור לגמרי – תחרות שנקראת CASP שרצה עדיין כל שנתיים וביקשו ממי שפתר מבנים לא לפרסם אותם במשך זמן התחרות, ה-CASP Predictors קיבלו רצף וכמה הצעות למבנה ומארגני התחרות יכלו להשוות את התוצאות.

היסטוריית אלגוריתמי Folding



אנחנו נסקור את מרבית ה-timeline הזה.

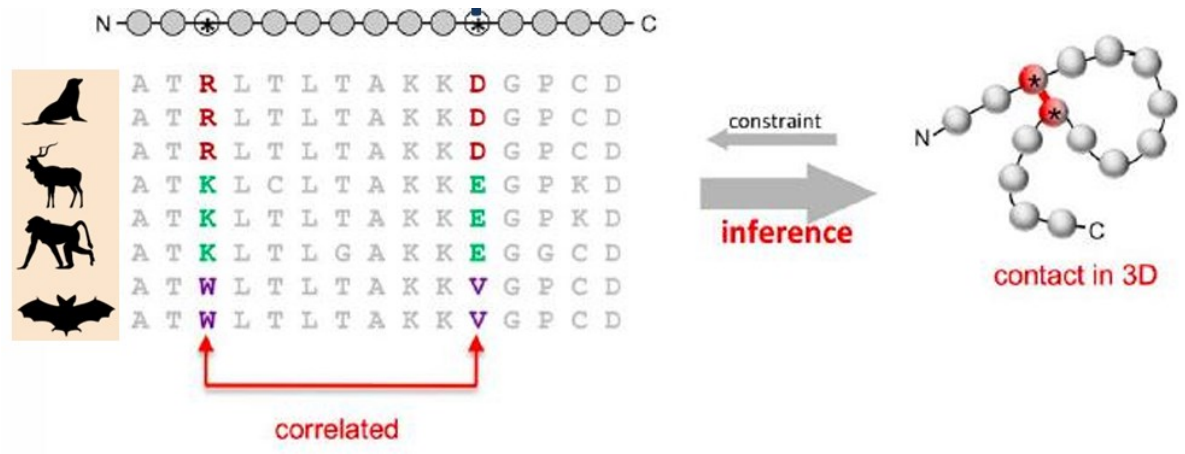


ציר ה-X מראה את השנים שבהם היה CASP (פעם בשנתיים), ציר ה-Y הוא מדד לרמת ההצלחה, כאשר בין 90 ל-100 זו רמת דיוק של קריסטלוגרפיה.

הקפיצה בין 94' ל-96' נובעת משיטות ab initio ואז תקופה ארוכה ללא שיפור. פריצת הדרך החלה בשנת 2010 עם Coevolution.

Coevolution

Sequencing הפך להיות מאוד זול, ופתאום יש גישה להמון רצפים. זה מאפשר להסתכל על שמירות ועל coevolution – יש לנו רצף שאנחנו רוצים לקפל. הולכים למסד הנתונים ופשוט שולפים את כל מה שדומה לסט המטרה באמצעות MSA (Multiple Sequence Alignment). העמודות החשובות הן אלה שמשתנות ביחד, כלומר co-evolve. הקורלציה אומרת שכנראה יש לחץ אבולוציוני ב-contact in 3D.

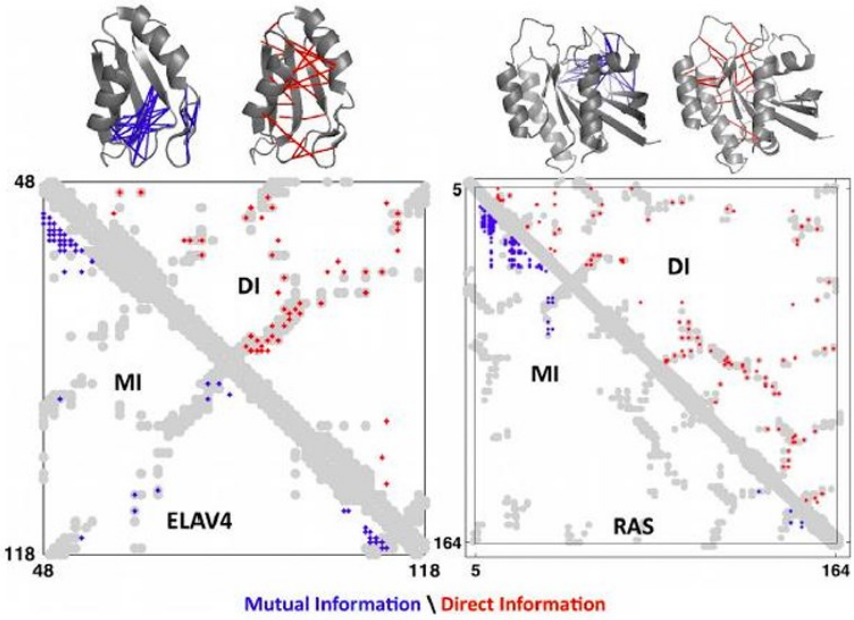


היה צריך שיהיו מספיק רצפים שעברו ריצוף כדי שהשיטה הזו תעבוד. ב-2011 יוצאת השיטה הראשונה שמצליחה לבצע זאת. הם השתמשו בווריאציה של Mutual Information שמחשבת את ה-frequency של כל עמודה בנפרד ואת ה-frequency המשותף שלהם. למשל בדוגמה למעלה ה-frequency של R הוא $3/8$ ושל R ו-D ביחד גם $3/8$. לעומת זאת יש כאן חומצות אמינו שלא נמצאות בכלל (תדירות 0).

הבעיה העיקרית היא טרנזיטיביות – mutual information הוא טרנזיטיבי אבל זה גורם לכך שלא נוכל לדעת מי ב-contact בגלל קורלציה ומי בגלל טרנזיטיביות (יש סיבות נוספות שהורסות את הסיגנל אבל זו העיקרי). מה שעשו בשיטה הזו היא Direct Information, מה ששונה הוא שכשמסתכלים על 2 עמודות לא מסתכלים עליהן בנפרד אלא בקונטקסט הכולל, וזה עוזר לנקות את הסיגנל.

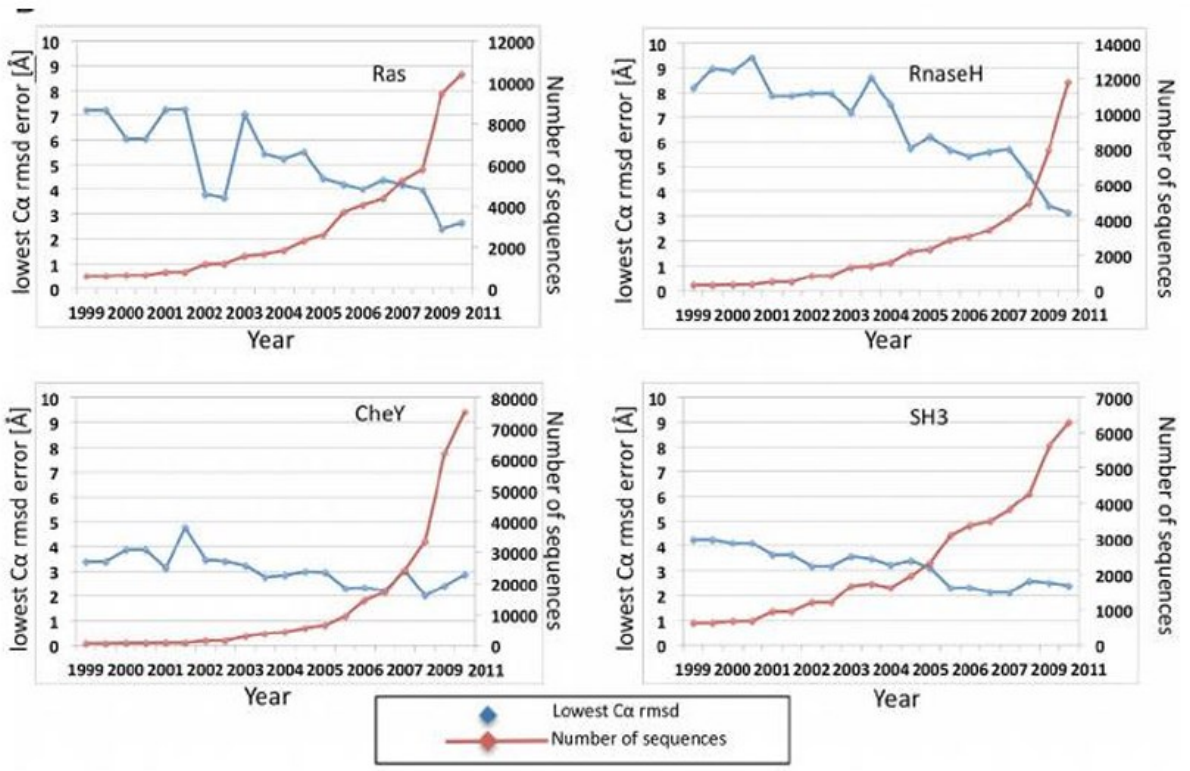
שיטות בהמשך

היו שיטות רבות שהגיעו אחרי, הכל סביב אותה שיטה. הצגנו את ההבדל בין Mutual Information ל- Global Information:



באפור מופיעים ה-contacts הנכונים של החלבון הכחול – האלכסון האפור זה השרשרת כלומר כל מה שקרוב by definition. אנחנו רואים ש-**mutual information** מוצא דברים לא בהכרח נכונים (בפינות של הדיאגרמה) ולא בהכרח עוזרים לנו לקפל – לעומת **direct information**.

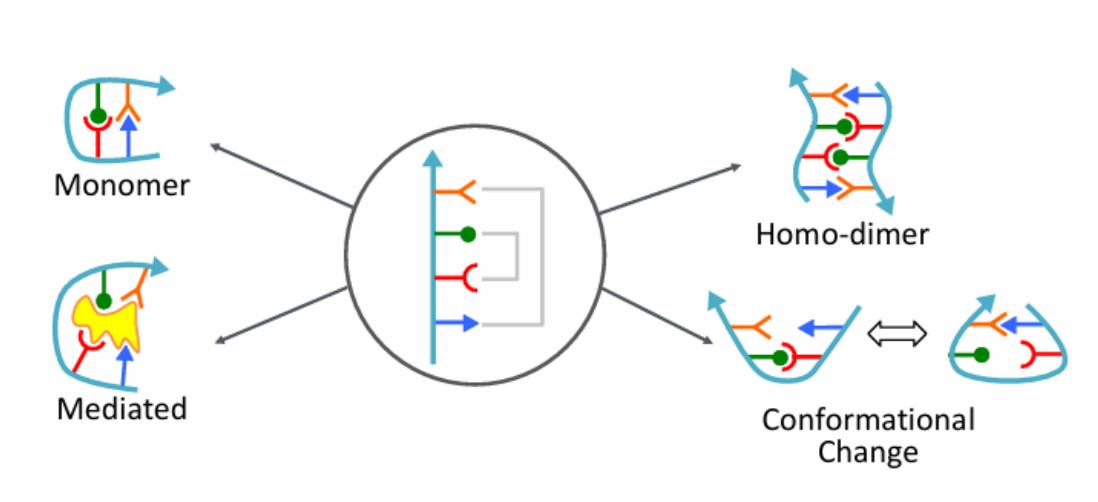
היו לוקחים MSA, מחשבים ממנו רשימה של contacts ואז הולכים לכלים אחרים שהיו זמינים והוסיפו להם את ה-contacts עם אילוצי מרחק ומקבלים קיפול. בהתחלה זה עבד רק אם הייתה כמות עצומה של רצפים הומולוגיים, הנה התלות:



באדום רואים את כמות הרצפים הזמינים, בכחול רואים עד כמה הקיפול מדויק – כמעט בכולם הדיוק משתפר.

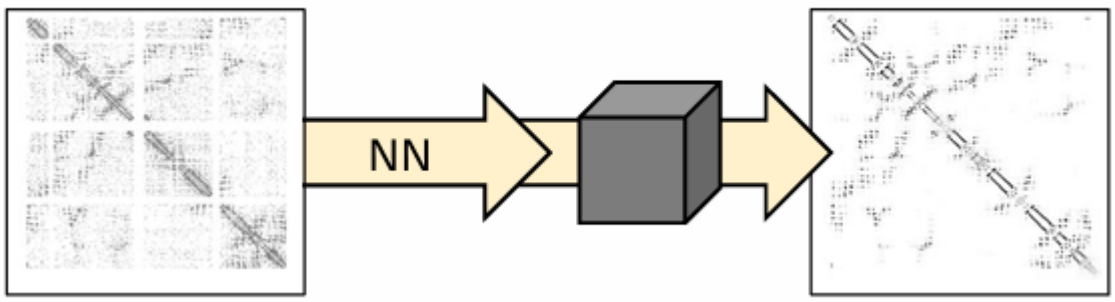
המקור של Contacts

היה קשה להוציא סט נקי של Contacts מדאטה של קואבולוציה, למרות שניסו לפתח הרבה שיטות. קודם כל בגלל בעיית הטרנזיטיביות. הסיגנל נובע מכמה מקומות: טרנזיטיביות (שעדיין הייתה, למרות שניקו), יכול להיות שיש מישהו באמצע שעוזר לחבר, אולי זה דימר, ייתכן שהוא קיים בכמה מצבים ואנחנו מנסים למדל רק אחת מהן.



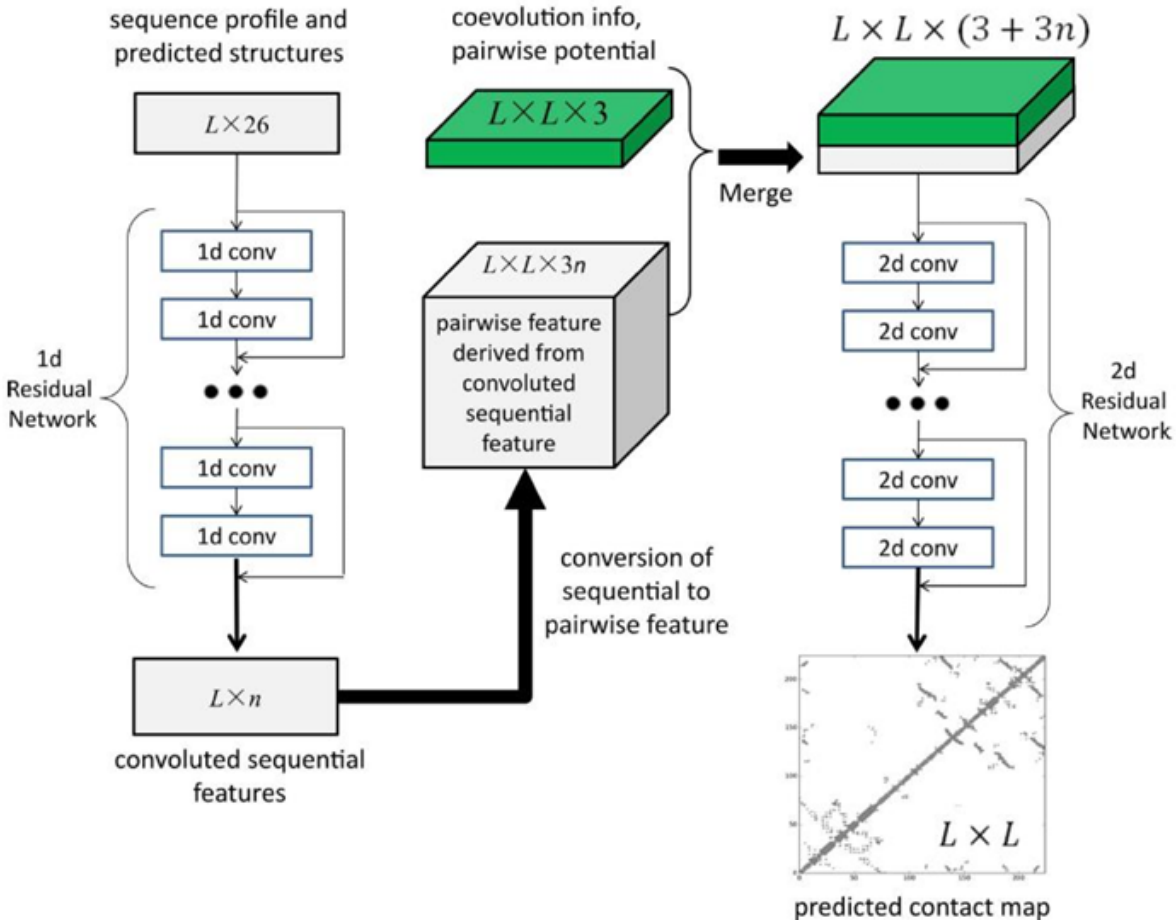
פתרון הבעיה – ניקוי בעזרת רשת נוירונים, 2017

היו כל מיני שיטות ניקוי, למשל להסתכל על מבנים קיימים ולבדוק מה מתאים לסט ה-contacts שלנו. בחירת קומבינציות של פרדיקציות של contacts, המון sampling.. בסופו של דבר שיטה שעבדה היא לנסות להסתכל על פרדיקציה של contacts – ייצוג של החלבון על ידי מטריצת מרחקים ושמים בכל ניקוי contact איזו הסתברות. לוקחים דאטה רועש ומעבירים אותו למודל נקי יותר בעזרת רשת נוירונים:



מה זה דרש? לקודד את הרצף בעזרת 1-hot encoding, מקבלים $L \times 20$. בפועל הקלט הראשון הוא $L \times 26$ כי הוסיפו מידע על מבנה שניוני ומידע נוסף. הקלט השני הוא מטריצת המרחקים עם עומק 3: $L \times L \times 3$, מטריצה אחת היא רק mutual information, שניה היא CCMpred שמנקה, והשלישית pairwise potential עד כמה ח. אמינית א' אוהבת להיות על יד ב'.

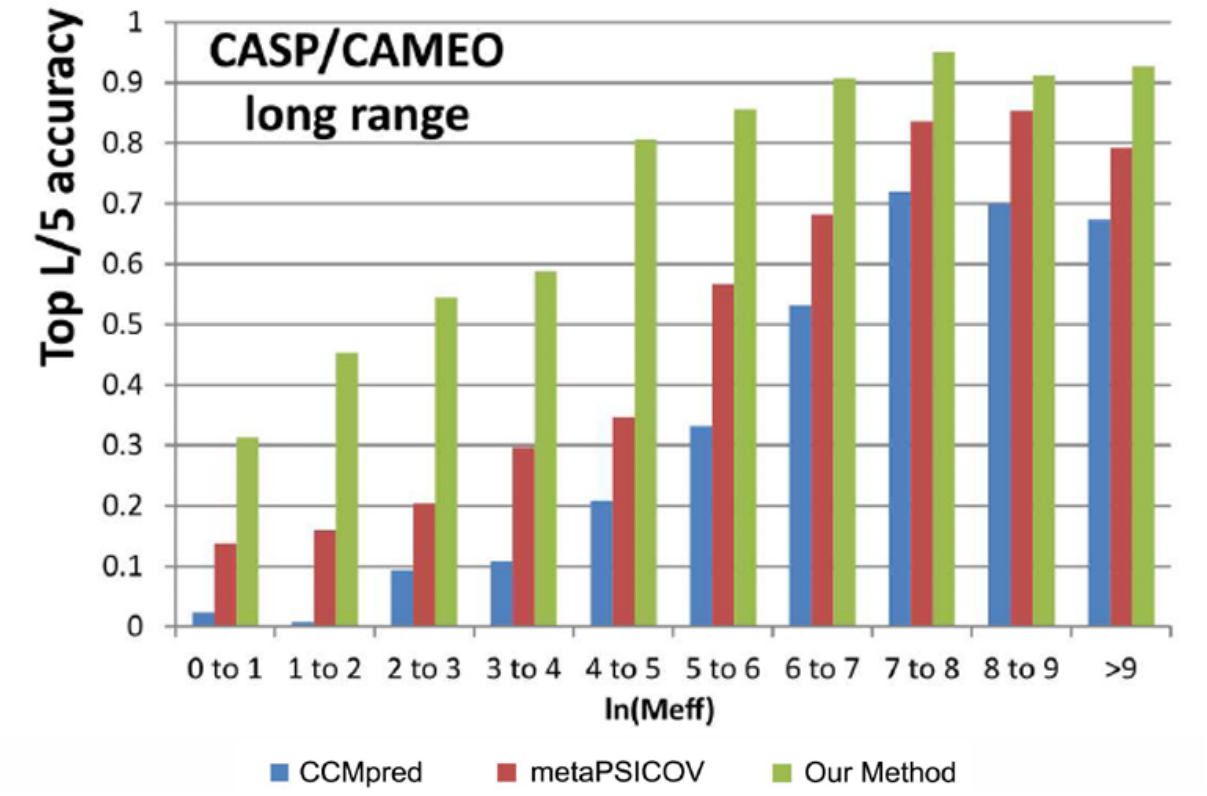
חיבור שני הקלטים מתבצע על ידי העברת הרצף מממד $L \times 26$ דרך residual network שמוציאה פלט באורך $L \times n$. לאחר מכן הופכים לדו-ממד על ידי concatenation, מייצרים $L \times L$ ואז לוקחים את וקטור במקום i, j ומשרשרים אותם ובתוספת הממוצע מקבלים $L \times L \times 3n$. עכשיו אפשר להוסיף לקלט השני:



קיבלנו תמונה דו-ממדית עם הממדים $L \times L \times (3 + 3n)$ (הערך השלישי $3 + 3n$ זה הפיצ'רים שלה). שוב עובר דרך Residual Network ומוציאים פרדיקציה שהיא contact map.

על מה נתאמן? על ה-PDB, אבל הלמידה היא גם מ-MSA ומהשיטה לחישוב קואבולוציה עושה איתה.

תוצאות:

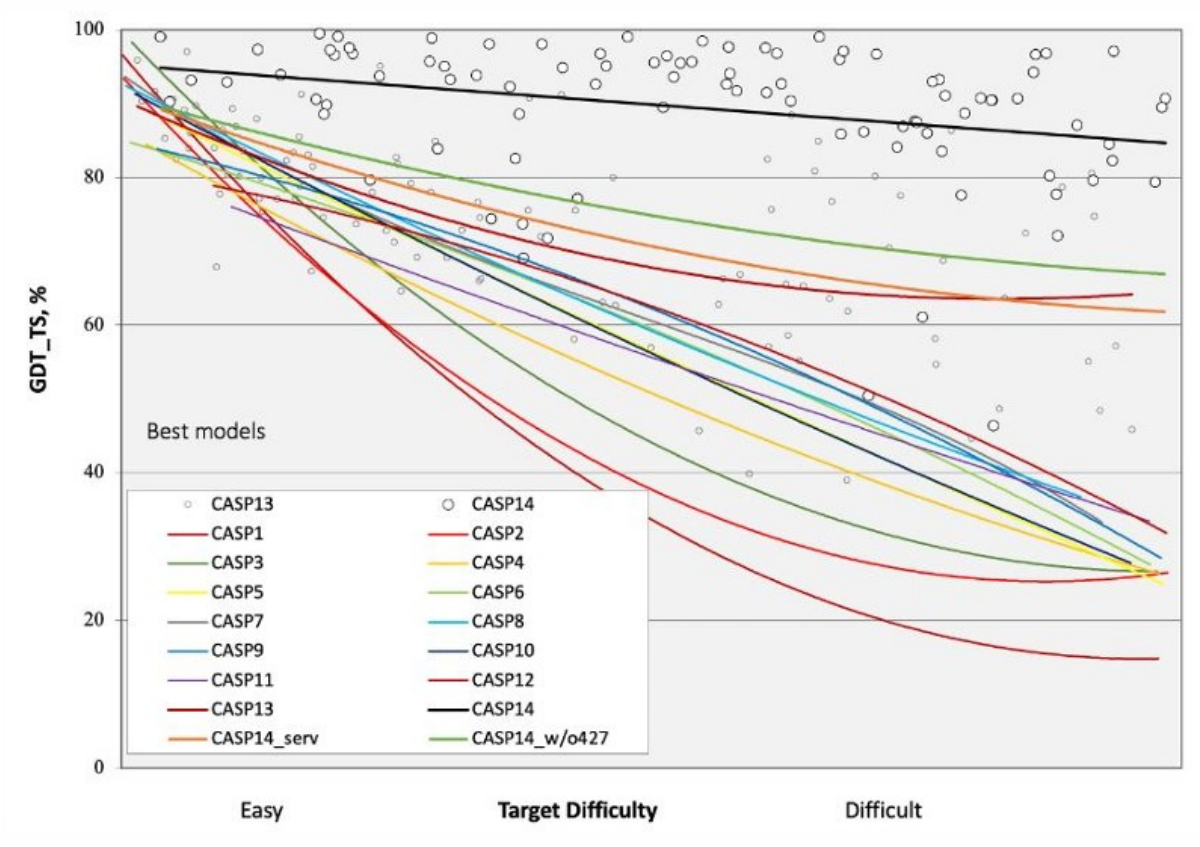


השיטה שתיארנו עובדת בצורה הטובה ביותר בכל המקרים. בציר ה-X היחידות הן $\ln(\mathrm{Meff})$ – הבינו שהיכולת לקפל חלבון תלויה בגודל MSA, כמה רצפים דומים יש. זה לא תלוי רק בכמות, אלא גם באיכות – לא רחוקים מדי ולא קרובים מדי. המדד הזה מחשב את המספר האפקטיבי של רצפים בשונות מסוימת. בציר ה-Y יש את המטריצה שהיא חלק מהשיטה, ובודקים דיוק של L (אורך הרצף) חלקי 5, 20 ה-contacts הכי חזקים ובודקים כמה מתוכם נכונים. השיפור משמעותי בעיקר באזור בו אין הרבה רצפים, ה-MSA הוא דל.

AlphaFold, 2018

השיטה דומה - Res Net, ולומדים מרחקים בעזרת קלסיפיקציה קטגוריאלית במקום contact (שהוא בינארי – יש או אין). רוצים גם לתת כיוונים ל-side chains ומנסים לנבא את הזוויות האלו.

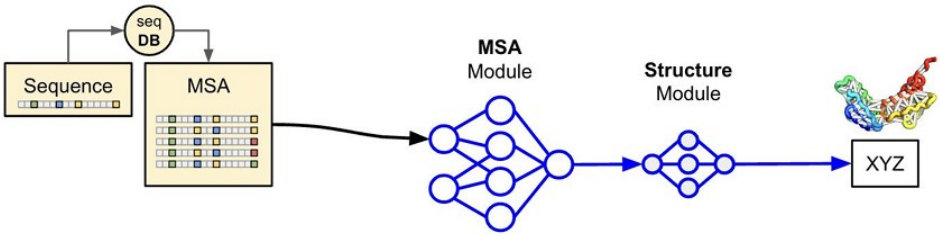
היו 3 קבוצות לפחות שהגיעו עם אותו רעיון לתחרות CASP, אבל AlphaFold הצליחו בצורה הטובה ביותר:



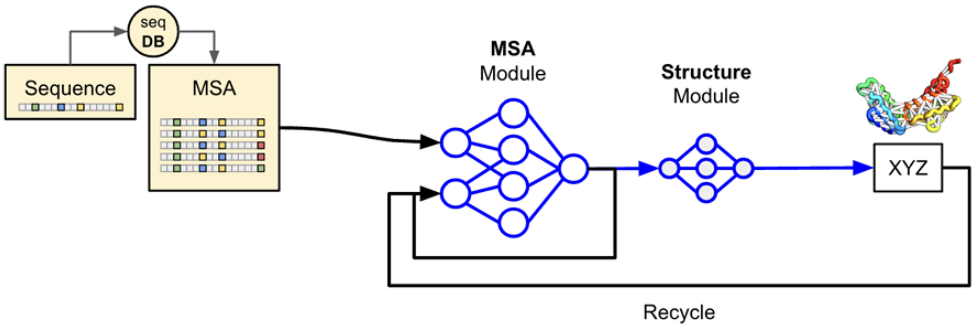
בציר ה-X יש סיווג targets של CASP לפי רמת קושי. בהמשך סיווג המטרה הפך להיות לפי עומק MSA. רואים שב-CASP14 יש קבוצה שהצליחה משמעותית בשחור.

AlphaFold2, 2020

עד עכשיו יש תהליך סדרתי: בונים MSA ומחשבים coevolution בעזרת אוסף שיטות, מכניסים לרשת שמוציאה contact prediction ומשם לשיטה נוספת כדי לקפל בעזרת ה-predicted contacts.

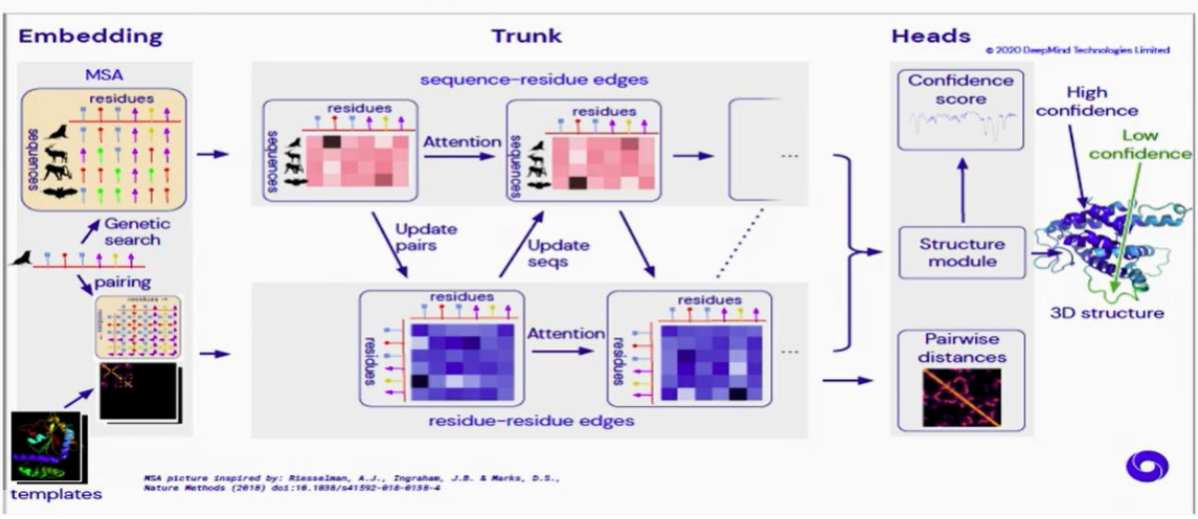


ב-AlphaFold2 הבינו שצריך לחבר בין הדברים. צריך מודול שעושה contact prediction ואחד שמקפל מיד אחריו – הכל באותה רשת. זה מה שאפשר לשפר את הדיוק.



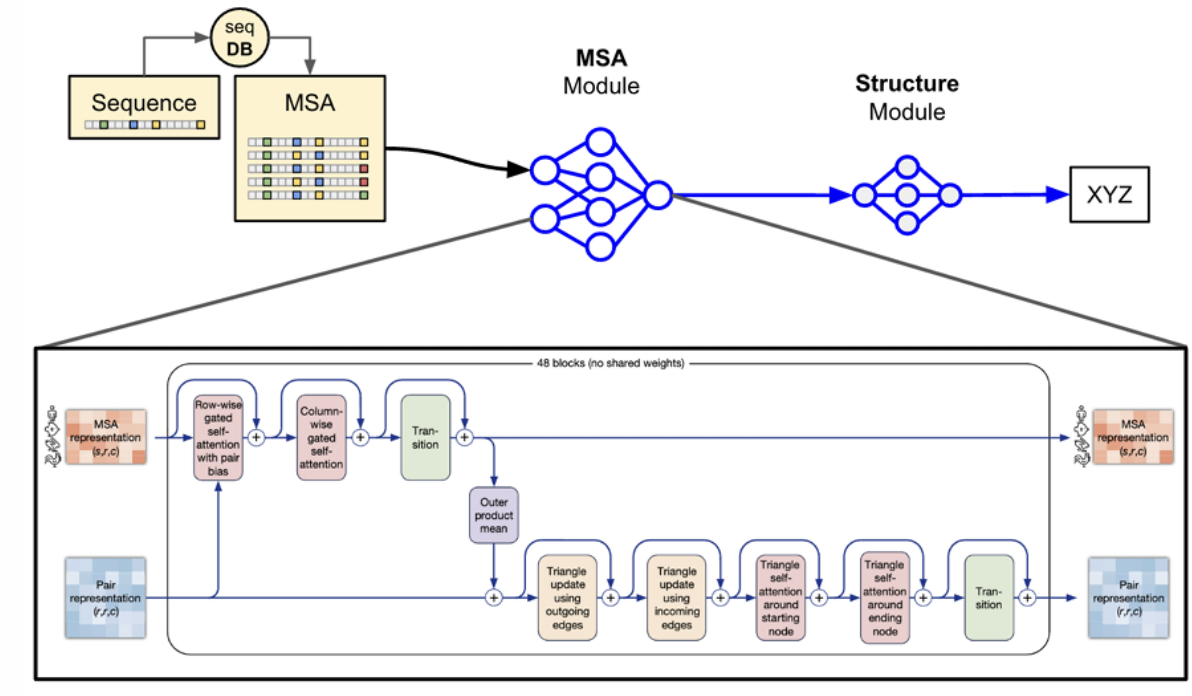
ברגע שיוצא prediction ראשוני אפשר להכניס אותו חזרה לרשת וזה עוזר לבנות מבנים הרבה יותר מדויקים.

איך נראית הרשת החדשה?



החלק הראשון הוא MSA, אליו מכניסים מבנים דומים אם מצאנו כאלו אך זה לא חובה. החלק הראשון Trunk קוראים לו Evoformer כי הוא לומד על האבולוציה/קואבולוציה. יש חלק שלומדת מה-MSA ויש את הייצוג של ה-contacts וכל החלקים מדברים אלו עם אלו. בסוף מוציא ייצוג של רצף ומרחקים, והם עוברים ל-structure module שלוקח את החלבון ונותן גם ציון confidence. כעת נרד לפרטים בכל חלק.

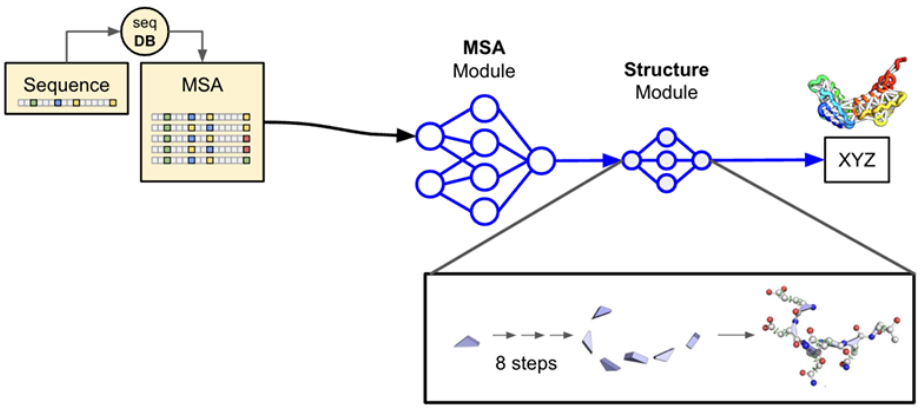
Evoformer



מקבל רצף MSA והמטרה שלו היא לקחת representation ראשוני ו-pair representation ולהוציא ייצוג של רצפים ו-contacts. סה"כ 48 בלוקים שמטרתם ללמוד את ה-coevolution בעזרת attention שהוא column-wise, אבל גם רוצים ללמוד את הקשר בין הרצפים לכן מגיע קודם row-wise-attention. משם עוברים ל-contact prediction בעזרת מה שחישבנו קודם. כשרוצים לנבא מרחקים חייבים שהם יהיו קונסיסטנטיים, וחייב להתקיים אי-שוויון המשולש. המטרה של כל ה-triangle update הוא לקחת את מה שיוצא מה-MSA ולהפוך אותו לקונסיסטנטי בעזרת אי-שוויון המשולש.

מודול לבניית המבנה

עד עכשיו השתמשו בשיטות קיפול ישנות עם אילוצי מרחק, אנחנו רוצים משהו שיוכל לקפל את החלבון בתלת ממד.



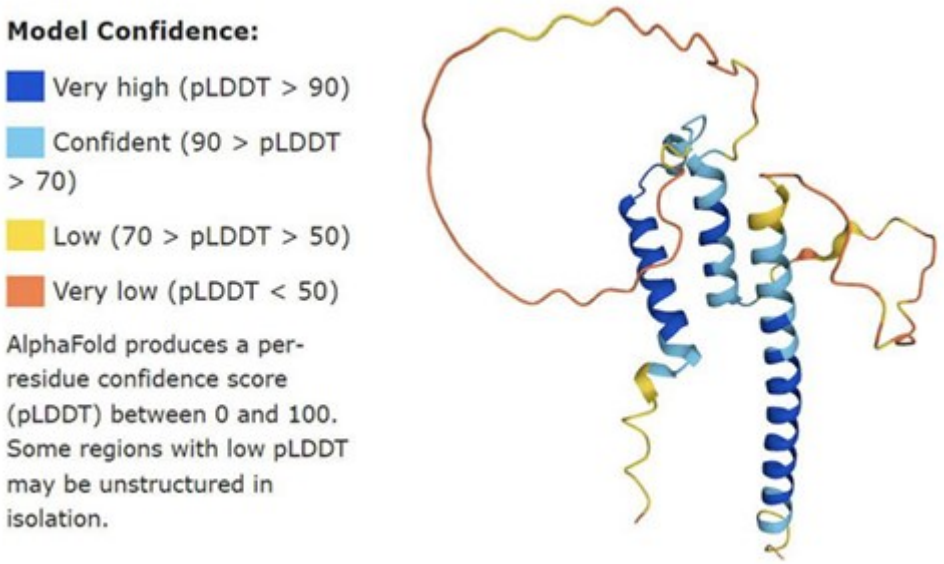
מייצגים כל חומצה אמינית בעזרת המשולש שלה (N, C_α, C), וצריכים למקם את המשולשים במרחב. הקושי כאן הוא שאנחנו צריכים להיות אינווריאנטיים להזזה ולסיבוב, אי אפשר לעבוד עם x,y,z – נכון יותר לעבוד עם ייצוג זוויות או מטריצת מרחקים. עיקר העבודה הוא סביב ה-Loss שהוא מבוסס מרחקים. על כל משולש בונים reference frame (RF) שבעזרתו אפשר להחליט איפה המשולש ממוקם במרחב.

בהתחלה כל המשולשים יושבים ב-0, והרשת לומדת למקם אותם בצורה קונסיסטנטית עם ה- predicted distance. לכל משולש ננבא את ה-RF, סיבוב והזזה במרחב שיסדרו את המשולשים. ה-Loss מחושב לפי L2 של המרחק המחושב והמרחק הנכון.

באמור, המודל החדש הוביל לשיפור במבנה שהצליחו לחזות, עד שב-cycle השלישי כבר מגיעים לרמת דיוק של 90% ואלו ערכים שמתקבלים בעזרת קריסטלוגרפיה.

Confidence

כל מבנה שיוצא מהמודל נצבע בצבעים שונים שממחישים את רמת ה-confidence:



כחול מראה על רמת confidence גבוה, כתום על נמוכה. הסיבה לתוספת הייתה שבמקור ה-PDB באימון היה מאוד קטן, ולכן חשבו לעשות פרדיקציות להמון רצפים אחרים וניקח את אלה שהפרדיקציה שלהם הכי טובה, ואז נעשה data augmentation כלומר אימון על דאטה הרבה יותר רחב. בשביל זה צריך שיטה שתגיד איזו פרדיקציה טובה ואיזו לא. בהמשך פשוט גילו שלכלי שימושים טובים נוספים.

המודל התאמן שבועיים על GPU והוא דורש גישה למשאבים משמעותיים. הפרדיקציה היום לוקחת מספר דקות אבל יכולה גם להגיע לשעות, תלוי בגודל של החלבון שחוזים.

סיכום

בהתחלה structure prediction נעשה בעזרת רצף, הכנסנו MSA ושיטות שניקו את הסיגנל מה-MSA ונתנו contact prediction אותם הפכנו לאילוצי מרחק, מהם קיבלנו קואורדינטות. בהמשך השיטות התפתחו ונהפכו למורכבות יותר.

[CASP12, 2016] השיפור הראשון הגיע מהכנסת רשת שלוקחת את ה-contact prediction והופכת אותו לאילוצי מרחק נקיים יותר.

[CASP13, 2018] בהמשך בעזרת רשתות נוירונים AlphaFold השיג שיפור משמעותי.

[CASP14, 2020] AlphaFold2 הצליח להחליף את אי הסדר ובעזרת עיבוד של MSA או sequences מוציא קואורדינטות.

כשהמאמר פורסם ב-2021 הוא יצא ביחד עם DB של למעלה מ-200 מיליון חלבונים מקופלים, מאוד קל לגשת אליו ולשלוף את זה.

[CASP15, 2022] את ה-structure prediction הגדרנו בהתחלה כמודל שלוקח רצף ומוציא קואורדינטות בלי ה-MSA. זה אפשרי בעזרת protein language models שמורידים את הצורך ב-MSA.

שבוע 10 – הרצאה 18

סרטן וגנטיקה

סרטן

מה זה סרטן? שם לקבוצה מאוד גדולה של מחלות שמה שמאפיין אותה היא תאים של הגוף יוצאים מכלל שליטה וגדלים בצורה לא מבוקרת.

אם יש תאים שמסודרים, בשלב מסוים אחד מהם מתחיל לגדול, להתחלק ולדחוף את האחרים כך שהוא לוקח את משאבי הגוף וגם פוגע ברקמה שהוא נמצא בה.

גידול סרטני - מסה של תאים מהמקור הסרטני שנהפך לגוש, וייתכן שיתישבו במקומות אחרים בגוף. בסוף, הסרטן מפריע לתהליכים בגוף ומוביל לכשל מערכתי.

בקונטקסט של הקורס - סרטן התחיל בתא אחד שקרה לו משהו שהוביל לשושלת תאים שנקראים בהמשך "סרטן". המחלה עוברת אבולוציה בתוך הגוף. כל אדם שסובל מסרטן זה מקרה שבו תא רגיל עבר אבולוציה לתא סרטני. משמעות הדבר היא שהיו שינויים ב-DNA של התאים. כמעט כל המקרים של סרטן הם מקרים של בעיה בבקרה על איכות החלוקה של התאים ושמירה על ה-DNA בצורתו המקורית.

אם נצייר את עץ השושלת של התאים, בשלב מסוים מופיעות בעיות, כל בעיה במקום אחר. בשלב מסוים אחד מהצאצאים פיתח מספיק בעיות כדי שיהפוך לתא עם יכולת סרטנית. כשיש "מספיק בעיות" – התא ממשיך להתחלק למרות שהסביבה אומרת לו לא, ולא מת למרות שהסביבה מנחה אותו למות. זה מוביל לחלוקה לא מבוקרת.

אם כל מופע של סרטן זו אבולוציה בלתי תלויה, איך יש סוגי סרטן דומים אחד לשני? סרטני מעי דומים כי הם התחילו בתא מסוג מסוים, אבל היו צריכים לשבור כמה "מנעולים" כדי לפרוץ החוצה כתא סרטני. דמיון בין המחלות נובע מזה שצריך לפרוץ את אותם המנעולים, ותלוי בסוג הרקמה הרשימה חוזרת על עצמה.

במחקר הסרטן חוקרים את המוטציות שקורות לגנומים ומנסים להבין מה זה אומר על הסרטן, כך נוכל להבין אילו מנגנונים מולקולריים עובדים בסרטן. יש סוגי סרטן שהדרך שלהם לדאוג שהם ממשיכים להתחלק זה שאחד מהרצפטורים שמקבל סיגנל להמשך חלוקה נתקע עם מוטציה שגורמת לו להיות פעיל כל הזמן. דרך נוספת היא שהתא מתחיל להפריש את הסיגנל שמעודד תאים להתחלק, ואז קולט את אותו סיגנל ובעצם נותן לעצמו אור ירוק להמשיך להתחלק.

להבין אילו מהדרכים האלו פעילות חשובות להבין את התנהגות הסרטן. ככל שחוקרים יותר אפשר לטפל בסרטן בצורה נקודתית יותר ויעילה יותר, כמו למשל לכוון תרופה לרצפטור.

גנטיקה של סרטן

נגדיר שאלה חישוביות: Mutational profile tumor.

ריצוף של ה-DNA של התאים. מקבלים רצפים והמטרה היא להשוות tumor מול healthy מאותו אדם. נרצה להשוות ולתת מפה של השינויים. אילו מוטציות יכולות לקרות כשעושים תהליכים כאלה?

סוגי מאורעות גנטיים

1. **מוטציה נקודתית** - אות אחת בהעתקה הועתקה לא נכון והוחלפה. זה קורה לא מעט, בייחוד אם מנגנוני הבקרה של שכפול תא מפסיקים לעבוד.

2. **copy number aberration** – חתיכה מהגנום הלכה לאיבוד או עברה שכפול. זו בעיה הרבה יותר חמורה בבקרה. יש מנגנונים שונים שיכולים להוביל לזה, אבל הרבה פעמים במקום שני העתקים (כמו אצל אדם תקין) בסרטן איבדנו אחד מההעתקים או שכפלנו אותה. מה קורה כשמאבדים אזור מהכרומוזום? פגיעה בבקרה. הזכרנו שני גנים: tumour suppressor (אחראיים על בקרה – מחיקתם מסירה מנגנוני הגנה) ו-oncogenes (אחראיים על חלוקה והישרדות תאית – הגברתם מגבירה פעילות סרטנית). הגנים לא נמצאים כדי לעודד גידול אלא כי יש פונקציה ביולוגית שהייתה כנראה מספיק חשובה כדי לשמר אותה באבולוציה, הבעיה במקרה הסרטני זה שהוא מעניק לסרטן יתרון.

3. **Fusion / Rearrangement** – דיברנו על שינוי של אובדן והכפלה, וזה כמו cut & paste. זה אומר שאת ההכפלה צריך למקם בגנום, אז יש אפקט שיש לנו העתק שלישי, וגם את האפקט שבגנום רגיל נכנסה באמצע חתיכת מידע שנכנסה באופן מלאכותי למקום הזה. דווקא בנקודות התפר של ההדבקה קורים אירועים מעניינים כמו היווצרות גנים חדשים שהם שרשור של שני גנים. אלו אירועים נדירים, אבל דווקא אלו האירועים הדרמטיים. לעתים הדרך להגיע לסרטן זה שני חלבונים שהזיזו אחד כך שיהיה צמוד לשני כך שייצר סיגנל A וגם יקלוט את אותו הסיגנל. כדי שזה יקרה צריך גם שאירוע נדיר יתרחש, וגם שהאירוע יהיה יותר סביר ממה שהיינו מצפים באקראי.

הגישה בגדול היא שמרצפים דגימה של הגידול ודגימה של רקמה בריאה, ונכנסים לבעיה חישובית שנקראת Variant Calling. בהתחשב בזה שמקבלים ריצוף בין שני מקורות ביולוגיים נרצה למצוא את ההבדלים ביניהם.

בעיות

1. **שיוך המקור** – לא תמיד תהיה תשובה אחת מאיפה הגיעה חתיכה סרטנית.
2. **בעיות ריצוף** – חוסר התאמה בקריאה אחת אולי לא מספיק כדי לשכנע אותנו שיש מוטציה, צריך להיתקל במספיק העתקים שחוזרים על עצמם עם אותה שגיאה כדי להשתכנע שהשגיאה אמיתית.

3. **בעיית דגימה** – אנחנו שולטים רק בצורה חלקית אילו רצפים נקבל, גם אם נרצף פי 10 במספר הבסיסים יהיו אזורים בגנום שלא יופיעו ברצפים או יופיעו רק מספר מועט של פעמים. זה יכול לקרות עקב "חוסר מזל" בדגימה או פשוט כי הם נעלמו.

איך מוצאים שינויים שאנחנו סומכים עליהם זה תחום שלם שלא נתעמק בו בקורס.

TCGA - The Cancer Genome Atlas - פרויקט בינלאומי שאסף מהרבה מרכזים רפואיים דגימות של ביופסיות ותאים רגילים, עליהם ערכו ריצופי גנום ושיטות נוספות. מדובר במסד נתונים גדול מאוד.

נניח שיש לנו את כל ההבדלים בין הרקמה הבריאה לרקמת הגידול. מה עושים עם זה? מה לומדים מזה? יש ביד רשימת מוטציות ואנחנו רוצים לקבל מזה תובנות ביולוגיות:

1. אפשר לחקור בכיוון Known mutations.
2. De novo analysis.

בסוגי סרטן רבים יש מוטציה שגורמת לכך שלא ייווצר חלבון תקין שאחראי על בקרה, למשל הכנסת STOP באמצע האזור שמקודד לחלבון. דרך אחרת לגרום נזק לחלבון התקין היא להחליף אותיות כך שהחלבון לא מתקפל כשורה, כאן קשה להבחין האם המוטציה היא כזו או לא. צריך להבין האם המוטציה גרמה לחלבון תקין, חלבון לא פעיל או למשהו אחר שלא ידוע לנו (שונות עם השפעה לא ידועה - VUS variant of unknown significance).

"לחפש מתחת לפנס" זה להסתכל על גנים שאנחנו יודעים שהם חשובים (פעולות שמשנות את tumor supressors או oncogene). היום יש כ-600 מוטציות ידועות. בטיפול המודרני לחולי סרטן רוצים להגיע למצב שכמעט תמיד מרצפים את הביופסיה, מרצפים ומקבלים אילו אזורים דומים לאזורים הידועים.

כמה מוטציות רואים כשאנחנו מרצפים? תלוי. יש סקאלה של בין פחות מ-1 לבין 10^6 ואחרים בין 100 לבין 10^6. יש סרטנים "שקטים" או "קרים" עם מעט מוטציות, ואחרים "חמים" עם המון מוטציות. יש סוגי סרטן יותר מועדים לפורענות מבחינת מוטציות ויש כאלה שפחות. בסרטן שהצליח לנטרל הרבה בקרות זה בשלב מסוים פוגע גם בסרטן עצמו כי הוא לא מצליח לשכפל את מה שחשוב לו, ומעט מדי מוטציות אומר שאפשר יחסית לשמור על תפקוד תקין.

נניח שיש 3,000 מוטציות, האם כולן משמעותיות? לא, יש כאלו ללא השפעה. המוטציות שתא סרטני אוסף בדרך, חלקן לא רלוונטיות בכלל לתהליך הסרטני אבל הן יישאר איתו בשימור ה-DNA, לכן זה שרואים מוטציה לא אומר בהכרח שהיא חשובה. אחת הבעיות היא להפריד בין bystander\passenger לבין driver mutations כלומר כזו שנותנת לתא הסרטני יתרון בתחרות מול הסביבה שלו לגדול מהר יותר מהסביבה שלו.

טיפול

איך נטפל בחלבון הבעייתי? נניח שיש לנו תרופה נגדו, כשנפעיל סלקציה חזקה התאים שישרדו הם אלו שהחלבון שהתרופה שלנו תוקפת יקבלו יתרון. אם החלבון שתקפנו הוא passenger התא לא ייפגע, אבל אם הוא driver התא הסרטני יצטרך למצוא חלופה.

איפה מחפשים את הבעיה?

אפשר להסתכל רק על known mutations, אך הבעיה עם זה היא שאנחנו מתמקדים רק במוטציות על חלבונים, ולא מסתכלים על אזורים שלא מקודדים לחלבונים, ואלו הם בין היתר אזורי בקרה. למשל הרבה פעמים החלבון תקין ומה שלא תקין זה הגורמים שבהם הוא מתבטא (כך למשל הסרטן מתחמק ממערכת החיסון). המוטציות מהסוג הזה הרבה יותר קשות לפענוח.

קריאה להרחבה: The Emperor of all Malady, One Renegade Cell.

זיהוי ודיאגנוזה

דיברנו על מוטציות ועל זה שהרבה מהן לא רלוונטיות. שאלה שמטרידה בכל זאת - גילוי הסרטן. כאשר באים לגלות סרטן יש כמה בעיות.

Diagnosis / Detection

צריך להפריד בין זיהוי נוכחות של סרטן לבין דיאגנוזה. למשל בנאדם חולה עבר ניתוח וטיפולים, לשארית חייו תמיד תרחף מעליו השאלה האם הסרטן חזר. בשלב הזה אחת הבעיות היא איך מזהים כמות מינימלית של המחלה שעדיין נשארה: minimal residual disease.

נחלק את העולם לשיטות כמו MRI וכו', אבל לשם כך צריך שיהיה משהו לא תקין ברמה שרואים, או שמחפשים חתימה של סרטן ברמת הדנ"א. שיטה כמו cell free DNA בשימוש כשהגידול קטן, לעומת מקרים בהם תאים שהשתחררו ומופיעים בזרם הדם - אבל זה קורה בעיקר כשיש הרבה מהגידול.

בגלל שאלו עניינים של רצף אפילו מולקולות בודדות יכולות להיות דגל אדום, לשם כך צריך מעט שגיאות ויכולת זיהוי ברצף קטן, כי cell free DNA מופיע מעט מאוד בדם. הדרך לנצל את זה היא שיש מעט מאוד מוטציות.

Diagnosis \ Early detection

כאן אנחנו לא יודעים שהבנאדם היה חולה בעבר, אז אין לנו טביעת אצבע של הסרטן שאנחנו מחפשים. מחשבה הגיונית אבל אולי נאיבית היא שאם מוצאים מוטציה מסוימת ברצפי ה-DNA זה אומר שאפשר לזהות עדות מספקת כדי לעשות early detection של נוכחות סרטן. איפה זה נופל? במשך השנים גילו שכשלוקחים גוש סרטני ומרצפים רואים המון הבדלים מבריא. אבל, אם לוקחים רקמה בריאה ומחלקים אותם לקבוצות קטנות של תאים, נגלה המון מוטציות, חלקן כאלו שיכולות להיות מזוהות עם סרטניות למרות שהתא לא סרטני, כי לא עבר את כל התנאים כדי להפוך להיות סרטני. הרבה אנשים מסתובבים עם תאים "דפוקים" בכל

מיני בעיות, אבל הגוף מאוד עמיד לשינויים כאלה. כשמחפשים ברגישות מאוד גדולה מוצאים את המוטציות האלה, ומוצאים אותן הרבה בתאים של מערכת הדם. רוב ה-DNA החופשי בדם בא מתאי דם לבנים שיש להם אפקט שמעט תאי גזע שמהם הן נוצרות. יש הגברה טבעית של מוטציות לא סרטניות, אלא רק יכול להפוך להיות סרטני אם יקרו עוד 3 תקלות רעות באותו התא - בעצם מדובר ב-false positive.

שבוע 10 – הרצאה 19 (הרצאת אורח)

הרצאת אורח: ד"ר מתן , Transcriptomics and single-cell RNAseq

אבסטרקציה וספציפיקציה

רעיון שכדאי לשים לב אליו בביולוגיה – שני צדדים של מטבע, אבסטרקציה וספציפיקציה.

The Abstraction (A Story or Model)	The Observable (The Physical Object & Event)
The "Gene" (A concept of heritable information)	**The DNA Molecule / Protein Sequence** (Physical chemical structures)
Metabolic "Pathway" (e.g., Glycolysis diagram)	**Single Enzymatic Reaction** (Substrate colliding with enzyme active site)
"Species" (Taxonomic category)	**Individual Organism** (That specific bacterium or bird)
Cell Cycle "Phases" (G1, S, G2, M)	**Cell Division** (The physical act of cytokinesis)
"Genotype" (Conceptual genetic makeup)	**"Phenotype"** (The visible trait or measured behavior)

בתוך מחקר ביולוגי יש המון אספקטים שהם רעיונות אבסטרקטיים, כמו למשל **גן**, מדובר באבסטרקציה: יש מולקולות DNA והחלטנו לקרוא לאזור מסוים גן בעקבות חלבון, סדר גודל של 20,000 גנים שמקודדים לחלבון (70-80 אלף לגנים בכלל). אפשר גם להגיד שמין הוא אבסטרקציה, או מחזור של התא..

החשיבה של בידוד חלקים אבסטרקטיים מחלקים אובייקטיביים עשויה לעזור.

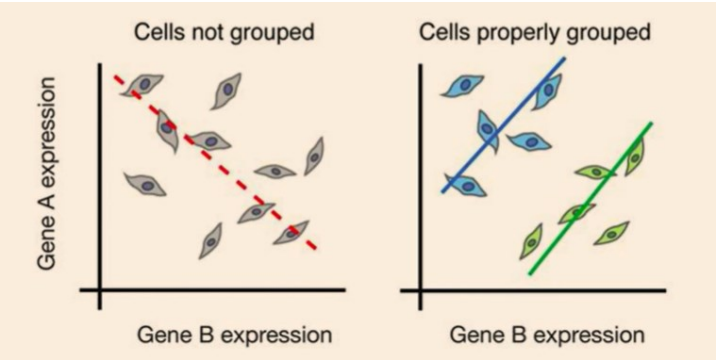
אנטומיה של mRNA

ב-DNA יש אזור שנקרא אקסון שהוא מקודד לחלבון (אם נתמקד בגנים שמקודדים לחלבון) ויש אזורים יותר רגולטוריים. בעולם אידיאלי, האם היינו רוצים למדוד RNA? בעולם אידיאלי היינו רוצים למדוד את החלבונים, אבל מדידה של RNA קלה יותר.

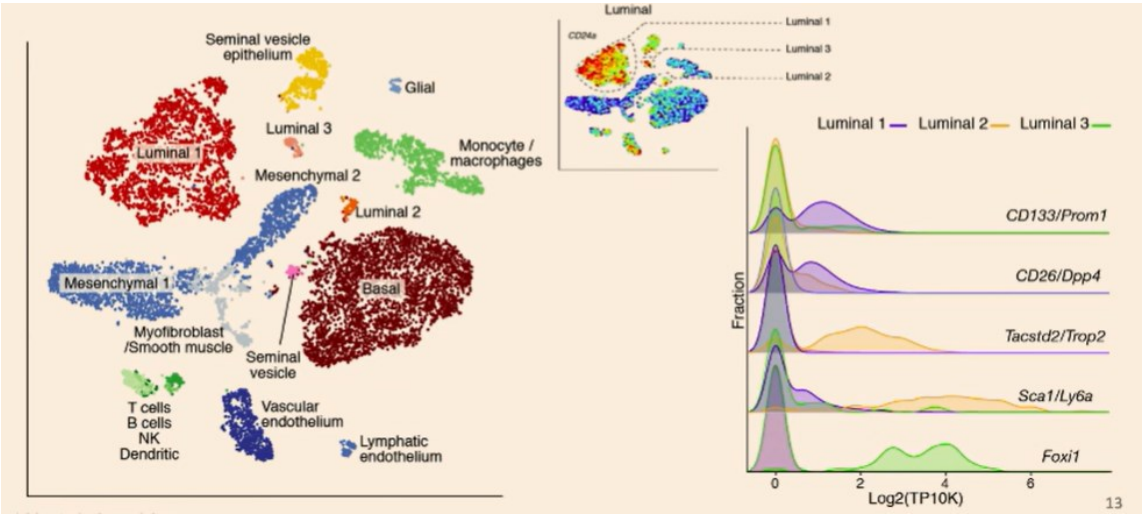
היסטורית ה-sequencing קרה ב-Bulk, כימתו את המולקולות שראו – פעם היינו מסתכלים על התמהיל של כל התאים ביחד והיום אפשר להסתכל על התאים הבודדים.

איך זה שימושי?

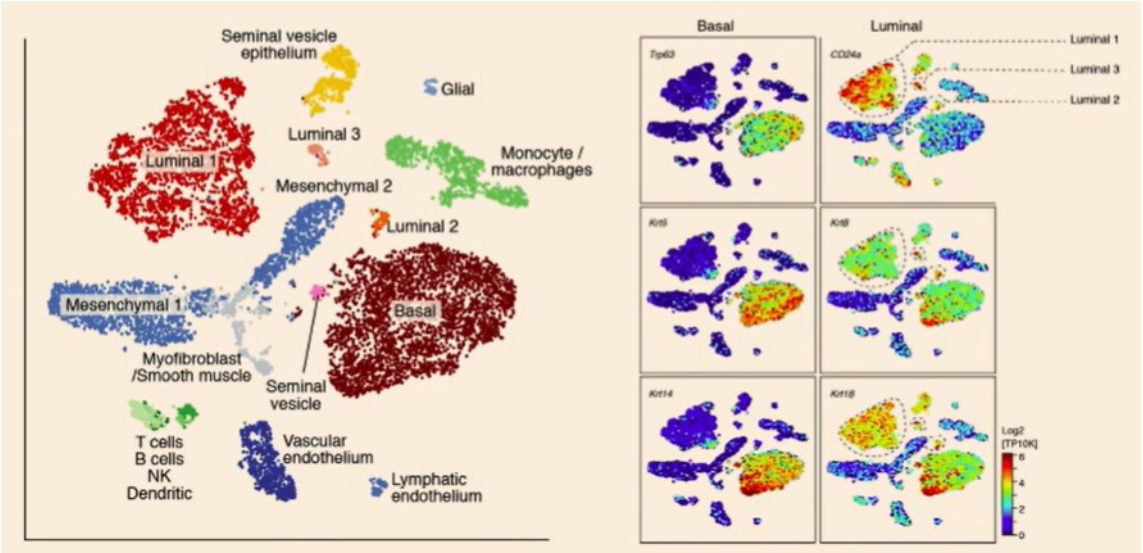
(1) **פרדוקס סימפסון** – כשמסתכלים על מיצוע ולא מחלקים לקטגוריות אפשר לפעמים לראות דפוסים לא נכונים, כמו למשל:



(2) **שונות בתוך התאים** – כשמסתכלים על תאים, כל אחת מהנקודות מייצגת תא בדו-ממד (הורדת ממד), הורדת הממד הזו מסייעת לנו לייצג מידע. אפשר לראות הטרוגניות מרשימה עם סוגים שונים של תאים.



(3) **צעדים ניסויים רבים בניסוי יחיד** – אפשר לזהות אוכלוסיות בתוך רקמה, לבודד אותם, לכמת, לזהות את האוכלוסיות השונות ואז לשאול את האוכלוסיות שאלות.



ייתכן שחלק מהתאים לא מעניינים לצורך הניסוי (כמו למשל תאים של מערכת החיסון) ובעבר הם היו נכנסים כחלק מהכימות הכולל – היום אפשר להבדיל אותם מהשאר ולא לכלול אותם בחישוב.

לסיכום: אפשר לזהות בבת אחת חלקים שונים ולשאול עליהם שאלות. הטכנולוגיה הזו התפתחה מאוד מהר ולאחרונה לא חל שינוי משמעותי.

סוגים שונים של שיטות single-cell

בסך הכל המטרה היא למצוא דרך לסמן מאיפה הגיעו מולקולות DNA. איך כל שיטה עושה זאת?

Plate-based/Sorted

הטכנולוגיות הראשונות שהבשילו היו מתאים בעזרת מכשיר יקר שעם לייזרים ומיקרופלואידיקה מיקם תאים בבאריות. בכל צלחת עם בארית היה תא. בעזרת אינזימטיקה היו מקבלים כימות.

בשיטה זו ההפרדה פיזית.

מיקרו-פלואידיקה

בעזרת מיקרופלואידיקה כלאו בתוך טיפות שמן את התא יחד עם סימון שלו.

בשיטה זו ההפרדה בעזרת שמן.

Micro/Nano-well

הרחפה של תאים על גבי צלחת עם באריות קטנות כך שבממוצע תא נלכד רק בבארית אחת, ואז מקבלים בכל אחת מהבאריות תא יחיד ואת הסימון שלו. היתרון של הטכנולוגיה זה שהיא מאוד זולה, ללא מכשור יקר. אפשר לקחת את הערכה ולשטח, לבצע ניסוי ולחזור למעבדה.

בשיטה זו ההפרדה בצלחות.

Split/Pool Combinatorial

בסוף מצאו שיטה לייצר את הברקודים המזהים בתוך התאים. יש דרכים אינזימטיות לעשות את זה בצורה קומבינטורית עם הכנסה של ברקודים. השיטה לא לגמרי נוצלה עקב מגבלות ריצוף.

בשיטה זו הברקוד בתוך התא.

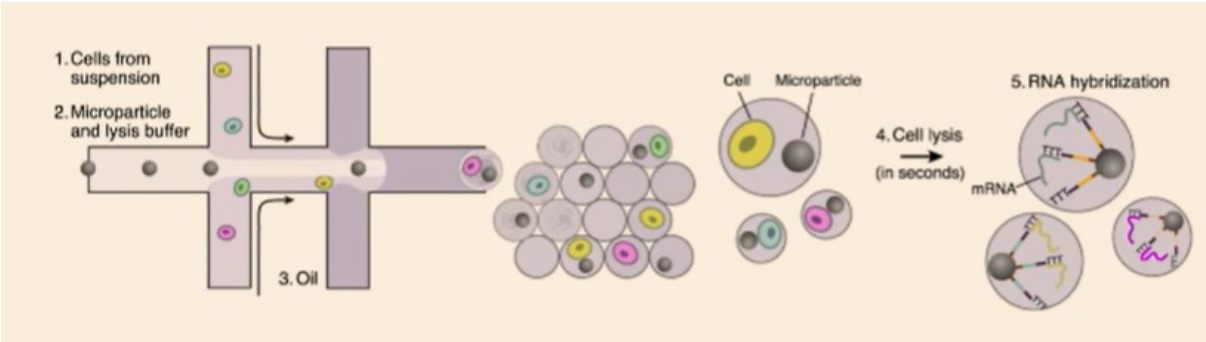
טיפול ברקמות והשהיית תאים

הרבה פעמים נרצה להקפיא תאים כדי להשהות תהליכים, אבל בעת הקפאה המים בתאים מתרחבים ופוגעים בתאים. לפעמים נעדיף שימוש בגרעיני תאים בלבד בעיקר במיקרופלואידיקה, כדי שתאים יעברו בהצלחה. למשל תא עצב יש לו מבניות מסוימת שיכולה להקשות על המעבר, או תא שומן.

המון מהפיתוחים בעולם ה-single cell התמקדו בנושא זה. לכן הרבה פעמים השתמשו בתאי דם שרגילים להיות מופרדים אלו מאלו (בניגוד לרקמה שבה הם תפוסים יחד).

כשמסתכלים על שיטות חישוביות צריך לבדוק האם הסתכלו על תאים שמספיק דומים לתאים הנבדקים.

Droplet based scRNAseq



כאן תאים עוברים בתוך מיקרופלואידיקה שנכלאים בתוך בועית, מה שמאפשר לנו להוסיף ברקוד שיסמן מאיפה התא הגיע. יש חרוז שתופס זנבות poly-A (כי אחוז מאוד קטן של ה-RNA בתא מכיל אותו). המון פרוטוקולים שמסתכלים על RNA-seq משתמשים בתכונה זו.

יש טריק נוסף שנקרא UMI כלומר Unique Molecular Identity. כדי לעשות short-trade sequencing צריך לעשות אמפליפיקציה ואז מקבלים חוסר דיוק בכמות המקורית, ויש כל מיני ארטיפקטים כמו c-content מבנים שניוניים וצרות כאלה ואחרות. הרעיון של UMI זה בעצם במקום לספור העתקי PCR אנחנו נספור את כל הדברים שיש להם אותו UMI ומתמפים לאותו אזור בגנום מולקולה יחידה.

הסטטיסט המספיק הוא: רצף ברקוד תא, רצף UMI, ומיקום מסוים בגנום. כך לא חייב לספור את מספר המופעים, והסט הזה נקרא מולקולה אחת.

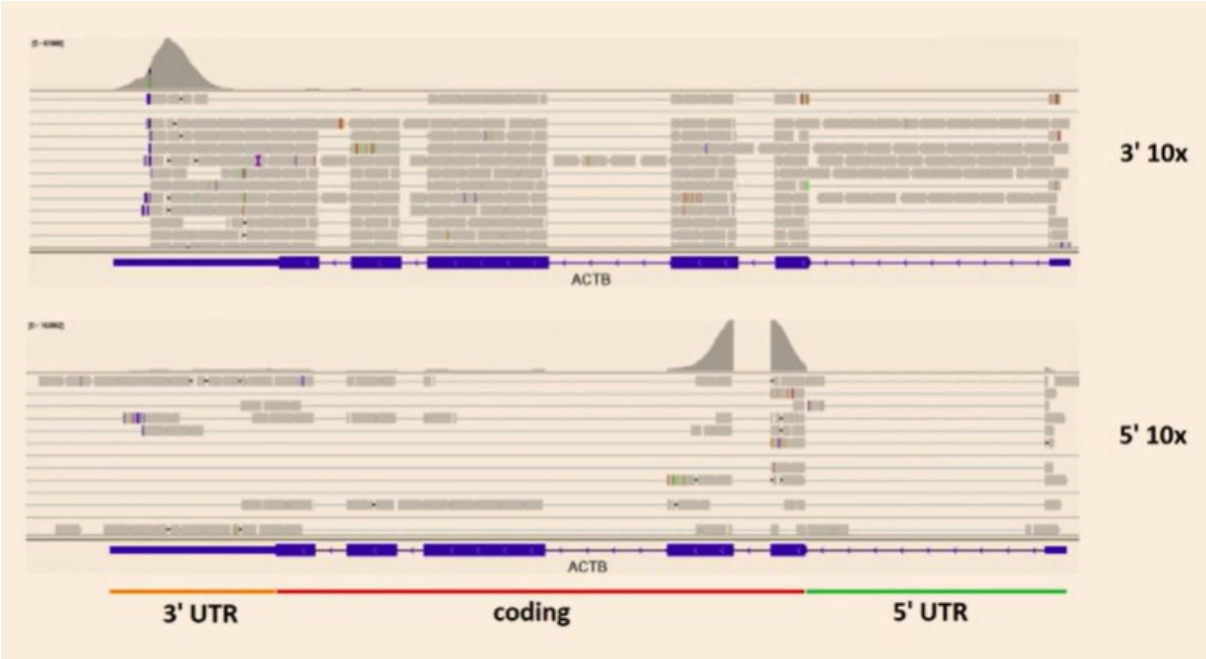
Transcript alignment

היסטורית כשעוד לא היה רצף גנום זמין, הרבה מהשיטות התחילו מ-assembly. יצירת רצף ומיפוי הטרנסקריפטים. לא טריוויאלי למפות לגנום כי בסוף נראה רצף עם פערים. נדרש alignment לגנום שהוא מודע להגדרות הגן או למפות לטרנסקריפטום.

"Digital" Gene Expression

למטריצה שמקבלים בסוף לפעמים קוראים digital gene expression אחרי שקיבלנו cell barcode, UMI, cDNA. מטריצה של גנים מול עמודות של תאים נפרדים.

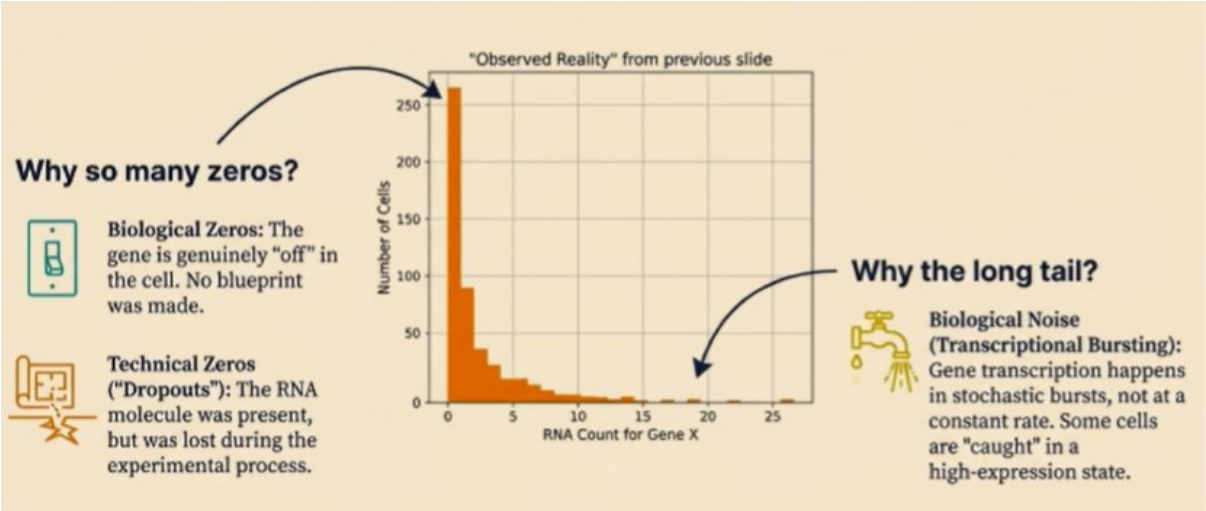
Read bias



ה-plot מנסה להראות איך ב-read יש peak באזורים מסוימים. יש שיטה אחרת שעושה בימות אחיד יותר של הגן, אבל רוב השיטות לא יכולות לעשות full cell RNA-seq וגם להוסיף UMI. המשמעות היא הרבה יותר רעש מדידה כמחיר לכיסוי טוב יותר של הטרנסקריפט.

Sparsity

כאן בציר ה-Y יש את מספר התאים, ובציר ה-X את מספר המופיעים של גן מסוים:

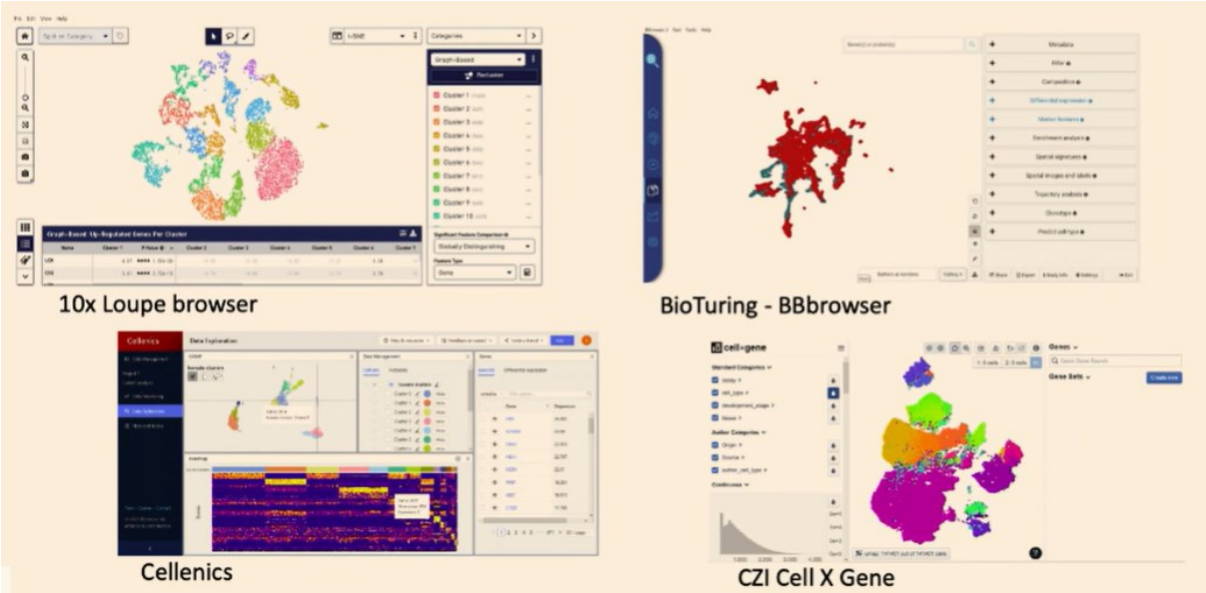


יש שני מקורות עיקריים של אפסים בדאטה – חלקם קשורים למדידה (תופסים בניסוי אחוז מסוים מהטרנסקריפטים שקיימים). הצד השני נובע מכך שמדובר באוכלוסיות תאים שהרסנו והרבה מהגנים מופעלים רק כשיש בהם צורך, כשהאוכלוסייה נהרסה הצורך הזה לא עולה.

Single Cell Analysis

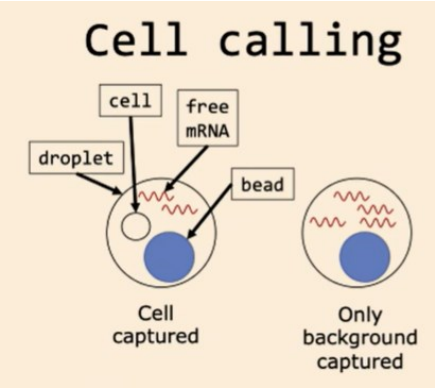
יש שני ecosystems עיקריים, אחד מהם ב-R (seurat) והאחרים ב-python (כמו scanpy או scverse).

יש כלים שונים לאקספלורציה אינטרקטיבית, Loupe browser וכדומה:

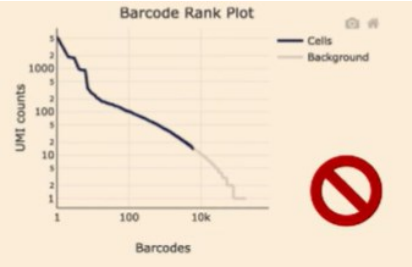


Cell Filtering

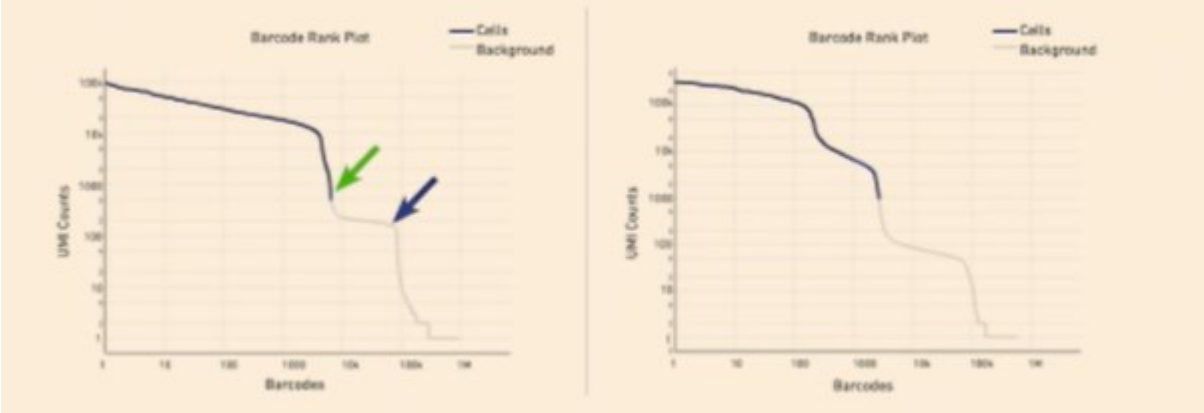
המטריצה בסוף הניסוי היא רשימת מולקולות, ואז אפשר לספור כמה פעמים רואים cell barcode.



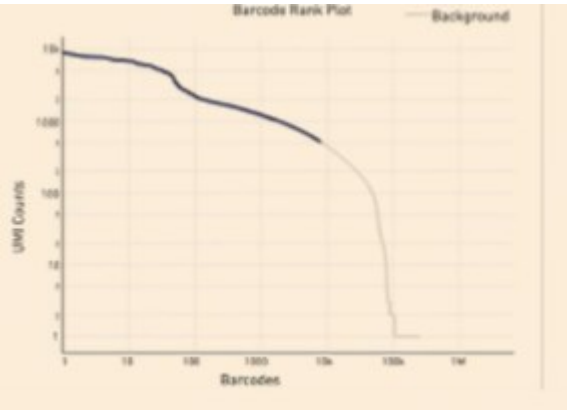
כל נקודה בקו היא תא בודד והתאים מסודרים לפי כמות הברקודים שראינו, כך נוצר קו רציף. בניסוי מוצלח נראה "ברך", ומה שמתחתיו כבר לא נחשב תאים:



בניסוי פחות מוצלח לא נראה את הברך. מה יכול לגרום לנו לכמה ברכיים בריצוף של רקמה? הסיבה היא ביולוגית כמו cell cycle, גדלים שונים, סוגי תאים שונים. ההטרגוניות הזו גורמת למבנה כמו מדרגות:



אפשר למצוא בפרסומים ברך מינורית, כי יש הרבה גורמים שמפריעים לנו להפריד תאים מרעש:



Cell quality filtering

יש פילטרים שונים כמו מספר הגנים השונים, מספר המולקולות וסטטיסטים נוספים. המספרים משתנים בין טכנולוגיה ניסויית אחת לאחרת.

Doublets / Multiplets

אנחנו בשאיפה תופסים תא אחד בכל טיפת שומן (או כל אחת מהשיטות האחרות). Doublets זה לתת לשני תאים את אותו cell barcode, איך עוד אפשר לקבל את אותו ברקוד לשני תאים? אם שניהם נכנסו לאותה בארית/טיפה, אבל אנחנו גם מסתמכים על העניין שבכל ניסוי יש סדרת beads שלכל אחד מהם רק ברקוד אחד ייחודי, אנחנו סומכים שהייצור הוא כזה – ואין שניים זהים. ייתכן מקרה בו יש את אותו cell barcode.

למה doublet זו בעיה?

יש doublet שהם פעמיים אותו סוג של תא (שני תאים מאותו סוג ואותו ברקוד), ואולי doublet משני סוגים שונים לחלוטין. זה ממציא לנו סוג תא חדש שמבטא גנים של שני סוגים, ולא החליטו לאן הוא שייך כי הוא

מבטא קצת מפה וקצת משם – בניסוי לא ברור אם מדובר בתא חדש או בטעות doublet. יש כלים שמנסים לעשות doublet detection.

סוגי "blindspots" מרכזיים

יכולים להיות מקרים של low mRNA, בעיות אופטימיזציה (קשה לעשות אופטימיזציה לכל סוגי התאים בבת אחת כשמסתכלים על רקמה אנושית), תאים שבירים (כמו נוירונים), תופעה שבה אנזימים הורסים RNA וכו'.

נורמליזציה והורדת ממד

לסוגי תאים שונים יש כמות מאוד גדולות ושונות של טרנסקריפטים. השאלה היא האם לא היינו רוצים לעשות נורמליזציה שבה במקום לספור את כמות הטרנסקריפטים האבסולוטית נספור את היחסית. יש מודלים שיודעים לקחת את זה בחשבון. יש tradeoff בין מודל מורכב לבין נורמליזציה.

PCA

אחד הכלים השימושיים הוא dimensional reduction באמצעות **PCA**.

מבחינת אינטואיציה, אם יש לנו ענן של מידע אז PC1 זה הקו הרחב ביותר בתוך הענן, והשני הוא הניצב לו ("מעיכה" של הממד הרחב ביותר בדאטה).

הרבה פעמים מממשים PCA באמצעות SVD. השלב הראשון ב-PCA הוא נרמול, מעבירים את כל הנקודות לראשית כדי לסדר אותן באותו מרחב. אם מרבית הדאטה היא אפסים, ברגע שנרמל אותו הוא יהיה צפוף מאוד ולא נוכל יותר לעבוד איתו בצורה נוחה (יש שיטות שמתמודדות גם עם זה).

Dimentionality reduction

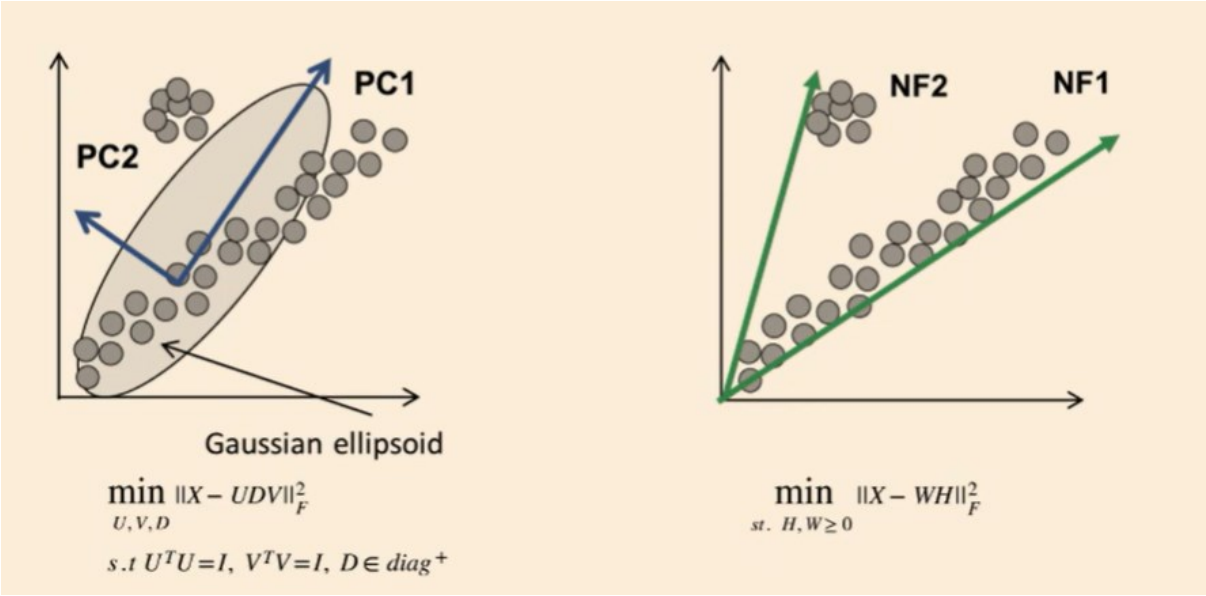
רוצים למצוא טרנספורמציה מממד גבוה לנמוך כדי למצוא ייצוג מצומצם יותר ששומר את התכונות של הממדיות הגבוה. הרבה מהביולוגיה לא מתרחשת בגלל גן יחיד אלא המכלול של הגנים, הורדת ממד מאפשרת אנליזה בעזרת מיצוע תופעות והתבוננות על פעילות משותפת.

Non-negative matrix factorization (NMF)

הרעיון הוא לפתור משוואה עם דרגת חופש ל-H, W ואז היא הופכת לבעיה קמורה עם פתרון יחסית פשוט. הרעיון הכללי לקחת מטריצת נתונים ולמצוא שתי מטריצות אי-שליליות שמכפלתן מייצגת בצורה כמה שיותר נאמנה את המקורית.

אחת המגבלות היא שהפתרונות לא ייחודיים, ויש רגישות לאתחולים. היתרונות הם שאפשר לפתור מאוד במהירות (כפל מטריצות), על דאטה ביולוגי מתקבלות תוצאות טובות.

השוואה ביניהם



כל PC בנפרד הוא חיבור וחיסור של principal component שמייצג את הדאטה, ולכן ה-PC לפעמים ימוקמו מחוץ לדאטה. ב-NMF לרוב נשאר בתוך הדאטה, מה שהופך את המילון (W) להיות יותר אינטרפרטבילי לעומת ה-PC.

tSNE\UMAP

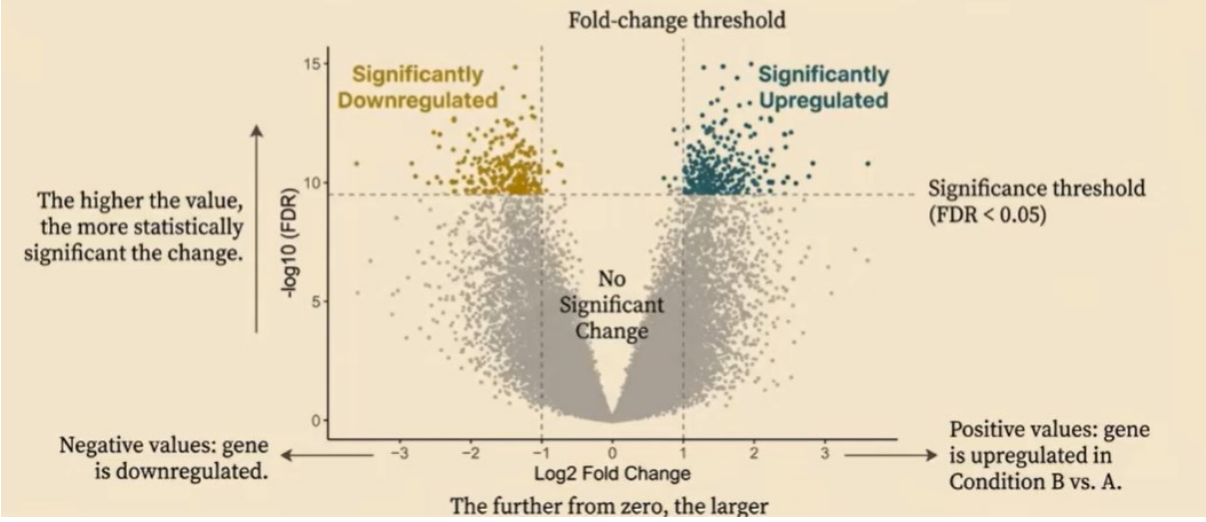
רוצים לקחת דאטה מממד גבוה ולהציג אותו בדו-ממד. אחת הבעיות בדאטה מממד גבוה היא שבממד הגבוה לכל הנקודות אפשר למצוא סט של קואורדינטות שבהן הן דומות או לא דומות, ואז רמת הדמיון יחסית אחידה. כשמורידים את הממד קשה לראות האם ההבדלים מעניינים או לא, למשל ה-PC1 השונות הכי גדולה מעידה על משהו טכני (כמו זיהוי batches) ואם נרצה הבדלים קטנים יותר, היוריסטיקה טובה היא להסתכל על מרחקים לוקאליים במקום גלובליים.

PC זו היוריסטיקה שמגדירה מה זה לוקאלי, אבל בגדול אפשר לצמצם את זה לבעיית Nearest-Neighbors. במקום לייצג את כל המרחקים בדו ממד, נייצג את המרחקים של השכנים הקרובים ביותר בגרף דו-ממדי. לקירוב ראשון זה מה שקורה ב-tSNE וב-UMAP (שתי דרכים להגדיר "שכונה"). הטרנספורמציות מנסות לשמור שהשכנים שלך במרחב הגבוה יישארו שכנים שלך במרחב הנמוך. כל דבר אחר לא מעניין (מישהו שיותר רחוק, כבר לא מעניין כמה הוא יותר רחוק). טיול רציף במרחב הגבוה יישאר כך גם במרחב הנמוך, אבל מה שהיה בממד גבוה קו ישר יכול להפוך להיות מעגל (או להפך).

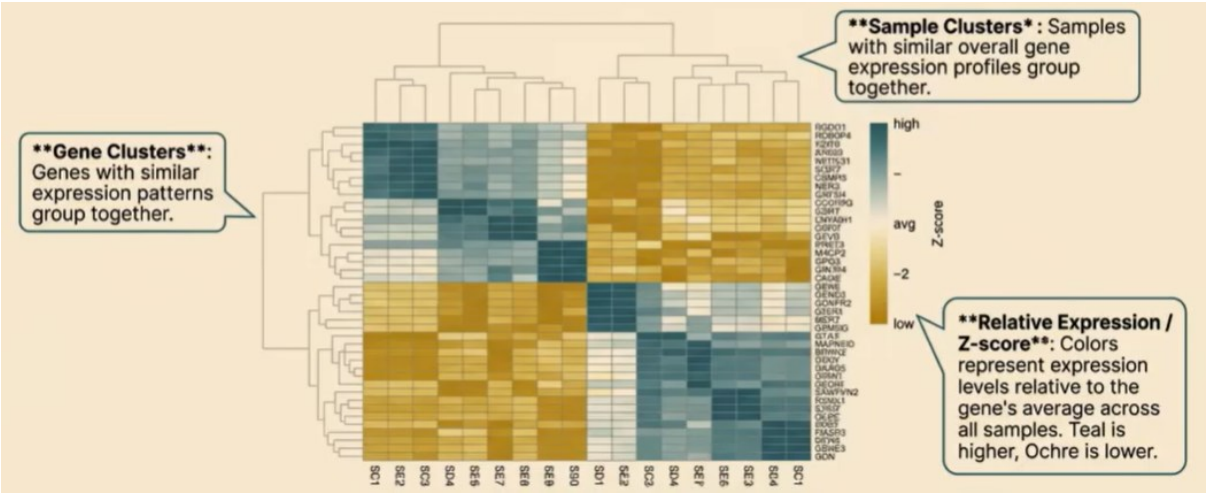
האינטרפרטציה צריכה להישאר בעיקר על המובנים הלוקאליים. יש תופעות כמו שכמעט תמיד doublet יהיה שילוב של סוגי תאים שונים.

מידול הבדלים ב-Gene Expression

אפשר להשתמש במבחנים סטטיסטיים שונים כמו T-test, מידול דאטה עם התפלגות Negative Binomial (NB) ובעצם לייצר מודל regression עם יותר מקום לשונות טכנית וביולוגית. זה מביא מטריצה של gene expression differences, מה שמאפשר להשוות בין מצבים שונים ולראות דברים שנראים מעניינים.



זה נקרא Volcano Plot והוא מנסה לכמת שינויים ב-gene expression ברמת החשיבות. ציר ה-x הוא יחס הממוצעים (log) וציר ה-y הוא ציון סטטיסטי של ההבדל. בהמשך אפשר להוסיף heatmap:



זה חושף מבניות.

איך אפשר להסתכל מתחת לפנס של מה שאנחנו יודעים? יש דאטה עם המון תוצאות ואנחנו רוצים למקם את התוצאות בתוכו. האם אנחנו רואים את ה-gene set הידוע בתוך התוצאות הניסוייות? אחד הקשיים היום היא להשתמש בידע ביולוגי ישן כדי לעשות אינטרפרטציה של ניסויים קיימים, ואולי מודלי שפה יאפשרו את זה. בינתיים, בהינתן סט מסוים של גנים מניסוי differential expression או single cell אפשר לבחור שאלות של over-representation.

שבוע 11 – הרצאה 20 (הרצאת אורח)

הרצאת אורח: ד"ר ניר קליסמן, Protein Mass Spectrometry

ספקטרומטריית מסות

הקדמה

המכשיר של ספקטרומטריית מסות זה הכלי העיקרי לענף שנקרא Proteomics. אנחנו נתמקד בעיקר בחלבונים. השימושים העיקריים הם של high throughput, לקחת סרום של דם או רקמה ולאפיין את כל החלבונים שיש בה (אילו חלבונים, כמה מכל חלבון, אילו מודיפיקציות יש עליהם..). המכשיר מאפשר לנו לעשות את זה.

נתחיל בין הבדלה בין sequencing ל-mass spectrometry. כשאנחנו רוצים לקרוא רצף של ח. גרעין (RNA/DNA) רצים לאורך שרשרת. השיטות היום הולכות וקוראות אות אות במולקולה. רושמות string של מה שהן קוראות. עושים את זה ב-reads קצרים ואז מרכיבים אותם. הדברים הטכניים שתומכים באפשרות לעשות את זה זו העובדה שאפשר לקחת RNA או DNA ולהגביר אותו, מה שמאפשר את המדידה, חלבונים - אי אפשר להגביר. הבעיה השנייה של ריצוף חלבונים לעומת ריצוף DNA הוא שמבבינה כימית הבסיסים מאוד דומים אחד לשני חוץ מהגודל - זה מאפשר להתאים שיטות לכל הספקטרום. לעומת זאת בחלבונים, שגם כן בנויים בצורה של שרשרת לינארית, שרשראות הצד מאוד שונות אחת מהשנייה (במטען, פולאריות, הידרופוביות). זה הופך את הבעיה להרבה יותר קשה. להבדיל מטכנולוגית ריצוף של ח. גרעין שבה אנחנו יכולים לקחת מולקולה בודדת ולקרוא אות אחר אות, לחלבונים אנחנו צריכים להגיע לגישה אחרת לגמרי.

ספקטרומטריית מסות בזיהוי חלבונים

על רגל אחת, מדובר במשקל מאוד מדויק של מולקולות. כל מה שאנחנו יכולים לעשות זה לקחת סט של מולקולות, לשים אותן על משקל ולדעת מה הוא. בעזרת המשקל נרצה לזהות אינדיבידואל אחד מתוך הדוגמה המורכבת. ניקח את האינדיבידואל על המשקל ונמצא את המסה, נלך למילון שממפה את המסות וכך נקבל את השם. כך מזהים אלמנטים בדגימה מורכבת.

מהם שני האלמנטים הכי חשובים בתהליך? על המשקל להיות מדויק - אם המשקל עובד בכפולות של קילוגרמים כששוקלים בני אדם, הרבה אנשים יפלו על אותה משבצת. המשקל הזה הוא מדויק להפליא. הבעיה היא שהוא לא יוכל להבדיל בין שני חלבונים שההרכב שלהם דומה אבל הפריסה הממדית שלהם לא (לוצין ואיזולוצין).

האלמנט השני שאינו טריוויאלי הוא הספר/המילון שממפה כל חלבון למשקלו. פרוטאומיקה רוכבת על ההצלחה של הגנומיקה שמאפשרת לנו למצוא את הזיהוי באמצעות הספר. אם נרצה לעשות פרוטאומיקה על יצור נדיר שאף אחד לא ריצף - לא נוכל לעשות זאת.

מסיבות טכניות אנחנו לא מודדים את המסה של החלבון המלא, כי המסה הזו גדולה מדי לספסיפיקצייה הטכנית של המכשיר. חותכים את החלבון לפפטידים קצרים, כל אחד בין 10 ל-20 ח. אמינו. האם זה לא רק הופך את הבעיה למורכבת יותר? עדיין רצף של 10 או 20 ח. אמינו הוא בדרך כלל ייחודי לחלבון. כן אפשר לזהות חלבונים רק מרמת הפפטידים. בדרך כלל אנחנו משתמשים באנזימים שחותכים את החלבון.

מכשיר להפרדת פפטידים

הצגנו שיטה מאוד ישנה כדי להעביר את הרעיון: אנחנו לוקחים חלבונים במבחנה בתוך תמיסה מימית, וצריך להעביר את הפפטידים למצב שהם מרחפים בוואקום, מה שצריך לעשות זה ליינן אותם (להוסיף להם מטען חשמלי). יש כל מיני דרכים לעשות את זה, באמצעות לייזר שמאדה את הדוגמה, או להעביר את הדוגמה דרך פיה צרה כך שהמים מתאדים והמולקולות מרחפות בגז. בעזרת שדות חשמליים מאיצים אותם ומפרידים לפי מסות, למשל באמצעות שדה מגנטי. למולקולות יש מהירות ומטען ולכן יעברו מסלול מעגלי (כוח לורנס), בהתאם למשקל ולמטען נקבל רדיוסי מסלול שונים ושם המולקולות ינחתו. מודדים את המסה של המולקולות שהיו ומזהים ניצוצות איפה שהייתה פגיעה. מכל המדידות מקבלים ספקטרום של המסה חלקי המטען. הבעיה היא שהספר שלנו הוא רק על המסות, צריך לנטרל את אלמנט המטען. ציר ה-y הוא intensity, סופר כמה מולקולות יש וזה כימות של החלבונים, להבדיל ממצב של ריצוף ח. גרעין שם צריך לעבוד כדי לקבל את הכמות. כל אחד מהעמודות בתרשים היא פפטיד.

אנליזת התוצאות

נעשה zoom in על אחד הפיקים, ומגלים שכל דבר שנראה כמו peak אחד הוא סדרה של כמה peaks עם רווחים קבועים יורדים בעוצמה. כל ה-peaks הם בעצם מאותה מולקולה, ובכל זאת הוא נתן כמה נתוני mass/charge - התשובה היא שזה נובע מהאיזוטופים של אותו יסוד (אטומים שונים של אותו יסוד הנבדלים זה מזה בכמות הניוטרונים). בהקשר שלנו, כל פחמן 100 בביוספרה הוא לא פחמן 12 אלא 13 (7 ניוטרונים ו-6 פרוטונים). מבחינת האורביטליות פחמן 12 ו-13 מתנהגים כמעט אותו דבר מבחינה כימית, אבל עדיין אחד מהם יותר כבד (זה עם ה-7 ניוטרונים). אם ניקח פפטיד ארוך עם הרבה פחמנים (קרוב ל-75 פחמנים שונים), ועכשיו כבר יש סיכוי לא רע שחלקם יהיו פחמן 13, וזה מסביר את ה-peaks השנים. ככל שיש יותר פחמני 13 ה-peak נמוך יותר. ההפרש בין העמודות הוא למשל ½ אז המטען הוא 2. בעזרת זיהוי ההפרשים הזה אפשר לזהות את ה-charge ולהוציא את המסות.

איזוטופים

נעיר שיש הרבה איזוטופים בטבע, אבל הסיבה שמה שראינו נובעת בעיקר מהפחמן היא שצריך שהאטום יהיה בכמות גדולה בפפטיד כדי להוביל לשינוי. למשל לגופרית יש איזוטופ שמופיע 1 ל-20, אבל יש יחסית מעט גופרית בפפטידים.כמו כן, בעזרת ספקטרומטריית מסות זיהו את האיזוטופים בטבע (הרבה לפני שזיהו את הניוטרון).

נמשיך לדבר על פרקטיקה של ספקטרומטריית מסות.

המצב האידיאלי זה שהיינו רוצים שיהיו לנו את המדידות של הפפטידים, ולקחת את הספר שלוקח את המסות ולראות איפה יש התאמה. בכל מקום שיש התאמה אני רוצה לראות מסה שמדדנו ולהשוות אותה לספר.

בעצם מה שקורה הוא שלמרות דיוק המכשיר הוא שלכל מסה שקבענו יש המון ערכים שמתאימים, הטבלה צפופה (טווח השגיאה מאוד קטן).

עכשיו אנחנו עושים תהליך בתוך המכשיר שדומה למה שנעשה בריצוף. מדובר במולקולות אמיתיות ועם שדות אלקטרומגנטיים אפשר לתעל אותן איך שנרצה - לוקחים מולקולות של peak מסוים ומתעלים אותן למקום עם גז דליל. הן מתנגשות במולקולות הגז הדליל (למשל חנקן) ופשוט נשברות. אפשר לקחת את הפפטידים ולקבל אוסף שברים. המכשיר רק מודד מסה, אז נמדוד שוב את השברים. מקבלים שוב ספקטרום שנובע מהפפטיד שראינו קודם רק שהפעם הוא שבור בכל מיני מקומות. הקשר שנשבר הכי בקלות הוא הקשר הפפטידי בין חומצות האמינו. מכל שבר אנחנו נקבל שני שברים: מהקשר הפפטידי לסוף הפפטיד (C-טרמינלי) או מהקשר הפפטידי לתחילת הפפטיד (N-טרמינלי). זה קצת כמו ריצוף, כי עכשיו הבדלי המסות בין שברים עוקבים מאפיינים את האות שנשברה - מזכיר תהליך של קריאת אות אחר אות.

עכשיו יש לנו דרך לשבור את החפיפה מקודם. שוברים את הפפטידים ומוצאים את המסות של השברים, כאן כבר תהיה רק התאמה אחת - מאיזה חלבון הגיע הפפטיד. המסה הכוללת זהה, אבל השברים זהים.

הצגת המכשיר

ראינו את כל העקרונות, נציג את המכשיר:

המכשיר מורכב ממשאבה שתפקידה לקחת את הדגימה במצב נוזל ולהזריק אותה לתוך המכשיר. המכשיר הוא יקר (2.5 מיליון דולר), וזו בעיה. זה מציב רף על הכניסה למחקר במדינות עניות יותר עם אנשים כישרוניים. הערה: אחרי שרכשו את המכשיר, הרצה של הבדיקה היא יחסית זולה.

בעיות במדידה

נדבר על אספקט טכני שקשור למדידה - בעיית טווח דינמי. למשל בתאי דם יש חלבונים שנמצאים בכמויות עצומות. יש בעיה שיש peaks מאוד גבוהים ו-peaks מאוד קטנים. המכשיר מכייל את עצמו לגבוהים ולא יוכל למדוד טוב את הקטנים (אולי אין מספיק עותקים מהשבירה). בעיית הטווח הדינמי קיימת בכל מדידה שעושים, אבל גם בריצוף של ח. גרעין (אם רוצים לעשות סריקת mRNA בתא צריך המון reads ואז המדידה יקרה ואולי בלתי אפשרית, רק כדי למצוא mRNA שנמצא רק בעותק אחד).

יש בעיה שנובעת מסיבוכיות: יש המון מולקולות שונות וסדרות איזוטופיות מתחילות לעלות אחת על השנייה, זה גורם למכשיר להיות לא בטוח איך קוראים חלק מהדוגמאות

יש בעיה של טווח דינמי וגם של סיבוכיות. אם ניקח את כל הפפטידים שנוצרו (בערך 200 אלף פפטידים שונים) ונזריק את כולם בבת אחת למכשיר, הספקטרום יהיה כל כך צפוף שלא יהיה ניתן למדוד. נמתח את המדידה לאורך זמן על יד זה שנפריד את הפפטידים במדד כימי. המדד במקרה הזה הוא בעיקר רמת ההידרופוביות של הפפטיד (LC - Liquid Chromatography). לוקחים את כל הפפטידים ושמים אותם בצינורית. הפפטידים ברובם הידרופוביים ויש הרבה ח. אמינו שיכולות להיתפס אלו לאלו ונעצרות בקצה. לאט לאט נכניס לנוזל בצינורית ממס אורגני (כמו מתנול) שמתחיל לשטוף את הפפטידים הפחות הידרופוביים החוצה. הם נשטפים בשלבים, וכך כל פעם נזריק למכשיר רק חלק מהפפטידים.

אנחנו בעצם משלמים בזמן, אבל כל דקה או מדידה עכשיו פשוטה יותר ואפשר להתמודד איתה.

עכשיו ציר ה-X שלנו הוא הזמן. אנחנו מתאימים בהתחלה את כל הפפטידים על הקולונה של ה-LC, מעלים את הממס האורגני ולאט לאט הפפטידים יוצאים עד שמקבלים מצב בו הקולונה ריקה. בכל נקודת זמן לא זרקנו את כל הסיבוכיות וזרקנו למכשיר, אלא הוספנו רק חלק קטן. יהיו ספקטרומים שיצליחו יותר ובאלה שיצליחו פחות, אבל המדידות בכלל מוצלחות יותר.

מודדים ספקטרום של כל היונים שנכנסים והם כל כמה שניות מתחלפים. הולכים ולוקחים בכל פעם סט פפטידים מסוים, מכניסים אותו למקום שבו שוברים, מודדים את השברים וכן הלאה. עושים בין 15 ל-50 שבירות כאלה ואז מודדים את הספקטרום מחדש. יכולים להיות בין 100 ל-200 אלף אירועי שבירה.

בערך מודדים מדוגמה מרקמה 100 אלף פפטידים, שהם כ-6,000 חלבונים שונים (בערך כמות החלבונים שיש בתא, כל תא מבטא רק חלק מהחלבונים ויש סה"כ כ-20,000). צריך בין 1,000 ל-10,000 עותקים של מולקולה כדי לבצע מדידה. לפעמים אפשר לבצע תהליכי הפרדה כדי לפשט את הסיבוכיות ואז אפשר להגיע ל-8,000 ו-10,000 חלבונים.

בעיות נוספות

באוקריוטים להבדיל מחיידקים הרבה מהגנים עברו הכפלות ושילושים, יש הרבה חלבונים כמעט זהים ברצף ובאים מגנים שונים - וזו בעיה, למרות שברמת ה-DNA יכול להיות שיש שינויים בקוד הגנטי, ברמת החלבון שהוא שמור יותר יכולים להיות שני חלבונים עם אותו רצף ואז לא ניתן להבדיל בין וריאנטים שונים אלא אם כן תהיה דרך למדוד את ההחלפה בחומצות בודדות.

False Discovery Rate (FDR)

איך קובעים מה ה-FDR? מקבלים רשימה של חלבונים שזוהו, עד כמה אפשר לסמוך עליה? זה פרט שחשוב בכל מדידה. בהקשר שלנו זה נעשה בעזרת הזרמת decoys, לוקחים רצפים לא אפשריים שנותנת סדרת פפטידים שמבחינת הרכב ח. אמינו זהות למקורי אבל לא קורות בצורה טבעית, ועדיין חלק מהספקטראות של

השברים מזהות חלק מה-decoys - זה קורה במדידות פחות איכותיות (עם מספר שברים נמוך). מה שעושים זה לתת score לכל התאמה. אנחנו מחליטים איפה אנחנו חותכים את הרשימה ומדווחים אותה ואת מה משליכים. תמיד יהיה tradeoff. אפשר להוריד את הרף ולקבל יותר זיהוי, אבל התוצאות בממוצע פחות נכונות. אנחנו שולטים בפרמטר.

מה אפשר למדוד רק ב-Mass-spect?

זרחון

התהליך שדיברנו עליו מאפשר למדוד בצורה טובה למשל איפה בפפטיד קרתה מודיפיקציה כמו זרחון. לא רק שהמסה הכוללת עכשיו כוללת זרחן, אלא גם כשנשבור יש שברים עם זרחון ושברים ללא זרחון. אנחנו יכולים להשוות ולדעת באיזו ח. אמינית קרה הזרחון ומה הרמה שלה. יש מדידות שמוצאות רק את הפפטידים המזורחנים בעזרת שיטות כימיות.

מדידת אינטראקציות בין חלבונים

זה בדרך כלל נעשה על ידי מדידות Pull-down, שמים תג מולקולרי ומושכים אותו עם משהו כבד (מגנטי או מסה גדולה) שיכול להתחבר לתג המולקולרי ומושכים מהתמיסה החוצה. יש הרבה אינטרקציות לא ספציפיות שיכולות להידבק, וחייב לנקות את הרקע הזה כי הם יכולים להיות בכמות לא זניחה.

דוגמה - חקר חלבון DUX4: רצו לבדוק מי החלבונים שנתפסים לקצה ה-C-טרמינלי. רוצים לנקות את החלבונים שלא נדבקים לקצה ה-C-טרמינלי או לתא. הניסוי הוא כפול, פעם אחת מושכים את החלבון המלא ופעם אחת את מה שמעניין.

יש כאן volacno plot - בציר ה-X היחס בעוצמה של חלבונים שיצאו ב-DUX4 לעומת DUX4C. ציר ה-y הוא עד כמה אפשר לסמוך על התוצאה, p-value. אנחנו רוצים שאינטרקטורים אמיתיים יהיו גם קיצוניים בכיוון DUX4 וגם גבוה (עם ערך p-value גבוה). מה זה אומר p-value?

אנחנו מדברים על מצב שיש כמה מדידות ומקבלים במדידה אחת ערך מסוים ובמדידה אחרת ערך אחר. הערכים לא יוצאים תמיד אותו הדבר. ההפרש בין המדידות מעיד לנו כמה אנחנו בטוחים בכל דבר. נמדדו 3 חלבונים ששונים בכמות הרעש בין המדידות. בחלבון הצהוב אנחנו יותר בטוחים שההפרש מתאר את המצב, ולכן נגיד שה-significance של B גבוה יותר משל A ומשל C. ה-p-value מייצג מה השונות במדידה, לבין ההפרש בין המדידות השונות. ככל שיש פחות שונות על ערך יותר גבוה, הביטחון יותר גבוה. צריך לחזור כמה פעמים.

Meta-Proteomics for Microbiome Studies

אפשר לקחת דוגמת צואה ולנסות לאפיין מתוכה microbiom של אדם או עכבר. עושים היום מטה-פרוטאומיקה שמבוסס על ריצוף DNA בדוגמה, ומהרצפים אפשר לאפיין אילו מיקרואורגניזמים יש. אפשר

לקבל את הרכב הבקטריות במקרוביום, יש עקרונות מידיים לשיטה הזאת: זה לא רק נותן את הפפטידים אלא גם את ה-hosts, כמו למשל פפטידים של המוגלובין שיכולים להעיד על דימום - את זה אפשר לגלות רק ברמת החלבון ולא ברמת ה-DNA. כמו כן, אפשר לזהות את הדיאטה. התוצאה היא טבלאות עשירות שאפשר למצוא בהן פפטידים שונים ממקורות תזונתיים שונים או מבקטריות.

האם אפשר להשתמש בפפטידים שנמצאים טבעית במעי? המעי דומה מאוד למבחנה שאיתה מייצרים את הפפטידים (חום, פרוטיאזות טבעיות) - מנסים לסנן פפטידים טבעיים ולמדוד אותם. אפשר לראות כמויות של בקטריות מסוימות בכמה מדידות זמן, ושני גרפים: תוצאות DNA Sequencing ואחת של המסות - התוצאות זהות מבחינת trends אבל לא בערכים. בריצוף לקח חודשים לעומת בדיקות לפי מסה שהיה אפשר לזהות תוך שעות (קריטי בנושא כמו מעי של פגים שבתוך שעות ספורות יכול לפתח זיהום שעלול להוביל למוות).