תרגיל בית 3- סטטיסטיקה ופילוגנטיקה

שאלה 1:

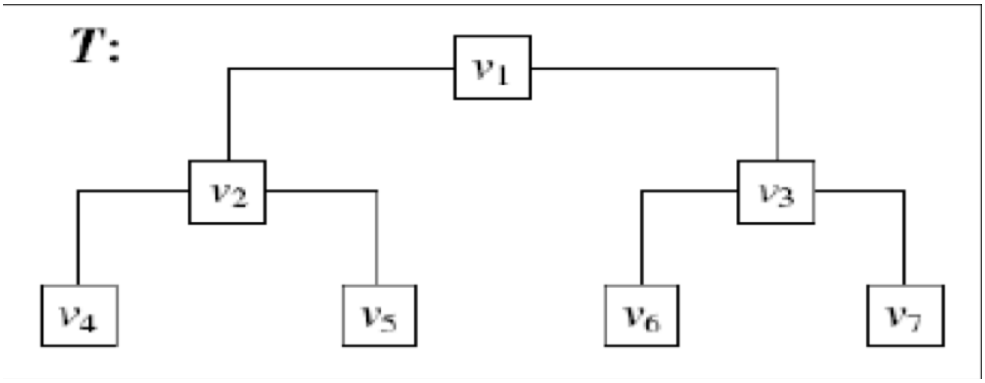
סטודנט להנדסה ביו רפואית רוצה לבנות ממיין שמטרתו היא לזהות סכרת על סמך 100 גנים.
הממיין ממיין 75% מהנבדקים אשר חולים בסכרת כחולי סכרת ו-30% מהנבדקים הבריאים כחולי סכרת. הסיכוי להיות חולים בסכרת הוא 25%.

א. חשבו את ה- sensitivity וה- specificity של הממיין.
ב. מה הם ה- $PV+$ וה- $PV-$?

שאלה 2:

בהינתן טופוגרפיית העץ הנתונה:

א. תנו דוגמה כך שבשורש תיתכן יותר מהשמה אחת וכן הציגו את כל העצים האפשריים המתקבלים תחת הנחת אלגוריתם פיץ'. כמו כן, מהו הציון הפרסימוני המתקבל?
ב. תנו דוגמה כך שבשורש תיתכן השמה אחת בלבד וכן הציגו את כל העצים האפשריים המתקבלים תחת הנחת אלגוריתם פיץ'. כמו כן, מהו הציון הפרסימוני המתקבל?

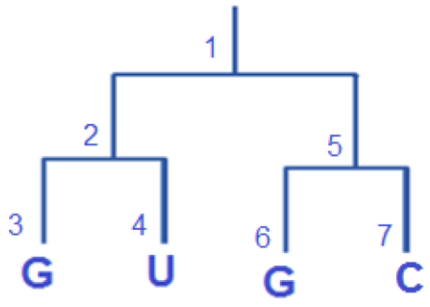


שאלה 3:

בהינתן מטריצת הניקוד הבאה:

	A	C	G	U
A	0	1	2	1
C	1	0	1	2
G	2	1	0	1
U	1	2	1	0

חשבו את ההשמות של הצמתים הפנימיים עבור העץ הנתון באמצעות אלגוריתם sankoff.
פרטו את כל החישובים וכן ציינו מהו ציון העץ.



<u>שאלה 4:</u>

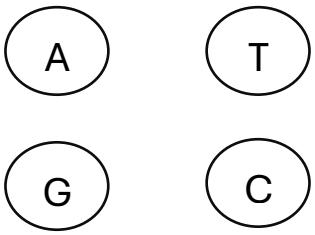
נניח כי קיים יצור היפותטי שבו עבור כל אות בדנ"א שניתן לראות יש לה תלות באות הקודמת בלבד.

הסיכוי לראות את האות 'A' ולאחר מכן את האות 'A' הוא 0.5, האות 'A' ולאחר מכן את האות 'T' הוא 0.5, הסיכוי לראות את האות 'C' ולאחר מכן את האות 'C' הוא 1, הסיכוי לראות את האות 'G' ולאחר מכן את האות 'G' הוא 0.2, הסיכוי לראות את האות 'G' ולאחר מכן את האות 'A' הוא 0.4, הסיכוי לראות את האות 'G' ולאחר מכן את האות 'T' הוא 0.1, הסיכוי לראות את האות 'T' ולאחר מכן את האות 'T' הוא 0.2 והסיכוי לראות את האות 'T' ולאחר מכן את האות 'A' או האות 'C' הוא 0.25.

א. מהי מטריצת מעבר ההסתברויות של המודל? מלאו לפי הטבלה הבאה:

	A	C	G	T
A				
C				
G				
T				

ב. ציירו את שרשרת מרקוב המתאימה למודל לפי האיור הבא:



שאלות קוד (על שאלה 5א' יש לענות ידנית ב-PDF)

<u>שאלה 5:</u>
ברצונכם לחשב האם ישנה בממוצע העדפה סטטיסטית להופעת זוגות נוקליאוטידים מסוימים ברצפי ה- UTR5 של אי-קולי.
ניתן שתהיה חפיפה בין זוגות כלומר, מסתכלים על כל מסגרות הקריאה.
כתבו פונקציה שנקראת **find_significant_nt_pairs(sequences)**, שמקבלת רצפים, מחשבת פי-ואליו אמפירי של כל זוג נוקלאוטידים ומחזירה אילו זוגות נוקלאוטידים הם בעלי העדפה סטטיסטית.

- קלט: sequences- רשימה של רצפי נוקלאוטידים מסוג string.
- פלט: רשימה המכילה str של כל זוגות הנוקלאוטידים עם העדפה סטטיסטית הסיגנפיקנטית (רמת מובהקות 0.05).
  במידה ואין זוגות כאלו- החזירו רשימה ריקה.

- יש להתחשב במקרה שבו יש זוג נוקלאוטידים שלא קיים ברצפים המקוריים אבל קיים ברצפים הרנדומיים, או להפך.
- אין צורך לבדוק את תקינות הקלט.

עקבו אחרי השלבים הבאים:

א. **נסחו ידנית** (בקובץ ה-PDF עם שאלות 1-4) את שאלה הפי-ואליו האמפירי שיש לחשב על מנת לבדוק האם ישנה העדפה לזוגות נוקלאוטידים מסויימים ב- UTR5.
ב. צרו 100 רנדומיזציות לכל רצף ברשימה. למשל אם הרשימה sequences מכילה 50 רצפים, יתקבלו 5,000 רצפים רנדומיים- 100 רצפים לכל אחד.
ג. חשבו את ממוצע ההופעות של כל זוג נוקליאוטידים אפשרי (סך הכל 16 זוגות) ברצפים שברשימה sequences (סך כל מספר ההופעות שכל זוג נוקלאוטידים מופיע לחלק למספר הרצפים).
ד. בצעו חישוב זהה עבור הרצפים הרנדומיים שיצרתם. לכל רצף ברשימה sequences יש 100 רנדומיזציות, לכן לכל זוג נוקלאוטידים יש לחשב 100 ממוצעים.
ה. השתמשו בשאלה הסטטיסטית שכתבתם בסעיף א' והחזירו רשימה של זוגות הנוקלאוטידים עם העדפה סטטיסטית מובהקת.

דוגמת הרצה:

```
sequences = ['ATCGATATATGCATC', 'ATCCATATATAT', 'ATCCTTATCCTTATATATATAT']

nt_sig = find_significant_nt_pairs (sequences)

print(nt_sig)
```

➢ ['AT', 'TA']

שאלה 6:

א. כתבו פונקציה שנקראת **calc_pssm(sequences, windowStart, windowEnd)** המחשבת את הסיכוי לראות כל נוקליאוטיד בכל פוזיציה בחלון.

- קלטים:
  - sequences: רשימה של רצפים (מסוג string) לפיהם תחושב המטריצה.
  - windowStart: האינדקס (מסוג int) של הרצף בו יתחיל החלון.
    למשל עבור windowStart=0 המטריצה תחושב עבור חלון שמתחיל בנוקלאוטיד הראשון.
  - windowEnd: האינדקס (מסוג int) של הרצף בו יגמר החלון (לא כולל).
    למשל עבור windowEnd=6 המטריצה תחושב עבור חלון שמסתיים בנוקלאוטיד השישי. כלומר האינדקס האחרון שנכלל במטריצה הוא 5.

← אורך החלון עבור מקרה זה הוא 6.

- פלט:
  - מטריצה מסוג np.array שמספר השורות שלו הוא 4 ומספר העמודות שווה לגודל החלון. כל שורה בפלט מייצגת נוקלאוטיד (לפי הסדר המוצג בטבלה) וכל עמודה מייצגת מיקום ברצף. המטריצה צריכה להכיל מספרים מסוג float המייצגים את ההסתברות של כל נוקלאוטיד להיות בפוזיציה המתאימה. יש לעגל את ההסתברויות ל-5 ספרות אחרי הנקודה העשרונית.

אין צורך לבדוק את תקינות הקלט.

	Position 1	**Position 2**	**...**	**Position [window size]**
A				
C				
G				
T				

*** הטבלה היא רק לצורך הבהרה של הצורה של הפלט, אין צורך (וגם אין אפשרות) לכתוב את הכותרות ב-np.array ***

דוגמת הרצה:

```
sequence_list = ['ACTGACTG', 'ACTGGCTA', 'AGCTCTAA', 'ATTTGCG']

pssm = calc_pssm(sequence_list, 0, 5)

print(pssm)

➢ array([[1.  , 0.  , 0.  , 0. , 0.25],
        [0. , 0.5 , 0.25 , 0. , 0.25],
        [0. , 0.25 , 0. , 0.5 , 0.5],
        [0. , 0.25 , 0.75 , 0.5 , 0.]])
```

<u>הסבר:</u>

- windowStart=0, windowEnd=5, לכן נחשב את המטריצה עבור 5 הנוקלאוטידים הראשונים בכל רצף. אורך החלון הוא 5 ולכן הפלט מורכב מ-4 שורות (עבור 4 נוקלאוטידים) ו-5 עמודות.
- בכל הרצפים הנוקלאוטיד A מופיע במקום הראשון ולכן בשורה הראשונה (שמייצגת את A) ובעמודה הראשונה (שמייצגת את המיקום הראשון) ההסתברות היא 1.
- במיקום הרביעי, בשניים מהרצפים מופיע הנוקלאוטיד G ובשניים מהרצפים מופיע הנוקלאוטיד T, לכן בשורה השלישית (G) והרביעית (T) ובעמודה הרביעית ההסתברות היא 0.5.

ב. כתבו פונקציה **find_best_match(pssm, sequence)** שמוצאת את האינדקס ברצף שיש לו את הסבירות הכי גבוהה להיות תואם ל-PSSM.

- קלטים:
  - pssm: מטריצת PSSM שמתקבלת מהפונקציה בסעיף א'.
  - sequence: רצף מסוג string.
- פלט:
  - מספר מסוג int המייצג את האינדקס.

דוגמת הרצה:

```
sequence_list = ['ACTGACTG', 'ACTGGCTA', 'AGCTCTAA', 'ATTTGCG']

ind = find_best_match(pssm, 'TCGGTCAACTTGTCATGGATT')

print(ind)

➢ 7
```

<u>הסבר:</u>

נסתכל על מטריצת ה-PSSM מסעיף א'. הנוקלאוטיד הכי סביר במיקום הראשון הוא A, במיקום השני C, במיקום השלישי T, במיקום הרביעי G או T, ובמיקום החמישי G.

נסתכל על הרצף שמתחיל באינדקס 7 ובאורך 5 (אורך החלון): 'ACTTG'.