נושאים בביואינפורמטיקה – עבודה שלישית

כללי: להלן ארבע מטלות, שמתוכן עליכם לבחור אחת. שלוש מטלות הן משימות תכנות והרביעית מטלה של קריאה וסיכום. כל מטלת תכנות יכולה להיבחר על ידי שני צוותים, מטלת הקריאה – על ידי צוות אחד. אני אוסיף מטלות קריאה או תכנות בהתאם לצורך. שלושת מטלות התכנות הן בתחום כריית המידע ממבני חלבונים וממשיכות את שתי עבודות התכנות הראשונות. הרקע התיאורטי שלהן מופיע בפסקה הבאה. מטלת הקריאה היא מאמר פורץ דרך בתחום הלימוד הממוכן של מבני חלבונים.

רקע תיאורטי למטלות התכנות: מטרתן של מטלות התכנות היא לימוד הפרמטרים של פונקציות אנרגיה מהמבנים. פונקצית אנרגיה היא פונקציה המקבלת כקלט מבנה ומחזירה מספר, המייצג אספקט מסוים של המבנה, ומאופיין בכך שבמבנים נכונים (למשל מבנים שהתקבלו בניסוי) הוא מקבל ערכים נמוכים. גישה מקובלת (הייתה אופנתית – יצאה מהאופנה – ועכשיו חוזרת) ליצירת פונקציות כאלו, היא להחליט מראש על פונקציה

$$E = \sum e_i(g_i(X), S_i)$$

כך ש: X הוא המבנה, g_i הוא מאפיין (feature) של המבנה (למשל זוג זוויות במפת רמצ'נדרן), S_i תת הרצף הרלוואנטי למאפיין (למשל טיפוס ח"א) ו e_i פרמטר הנלמד בעזרת הנוסחה:

$$e_i = -log\left(\frac{f^O(g_i(X)|S_i)}{f^E\big(g_i(X)\big)}\right)$$

כאשר $f^O(g_i(X)|S_i)$ היא השכיחות הנצפית (Observed) של המאפיין $g_i(X)$ בהנתן תת הרצף S_i, ו- $f^E\big(g_i(X)\big)$ היא השכיחות הצפויה (Expected) של המאפיין הזה במצב יחוס (reference) תיאורטי. יש גישות שונות לחישוב ערך היחוס, שכמובן יוצרות פונקציות אנרגיה שונות. בעבודה זו השתדלתי לבחור תמיד את אופן החישוב הפשוט ביותר. בעיה נוספת עולה כאשר המונה הוא אפס, ואז $e_i = \infty$, זה גם לא יציב נומרית (קשה לעשות חישובים) וגם ייתכן שהמאפיין הזה פשוט נדיר, ובמקרה לא גילינו אותו כי בסיס הנתונים שלנו קטן מדי. הפתרון הוא להוסיף תצפיות דמה (pseudo counts) לכל המאפיינים, כך שבמקום שכיחות אפס מקבלים שכיחות קטנה מאוד.

הנחיות כלליות למשימות התכנות:

1. אתם מקבלים בסיס נתונים חדש שימצא בקובץ sampleData22.5.22.zip המבנה של קבצי הנתונים זהה לעבודות הקודמות, אבל יש בהם יותר מפי 10 מבנים. בידינו בסיס נתונים גדול פי מאה בערך, אבל העבודה עליו צורכת פי 100 יותר זמן והתוצאות הן, איכותית, דומות.
2. הקובץ part3_utils מכיל מחלקות ושיטות לשרותכם. בפרט יש בו מחלקה Histogram ששימושית מאוד בעבודה זו.
3. אתם מקבלים שיטות main שמטפלות בכל מה שטפל לעיסוק בנתונים, ובפרט אופן ההצגה של הנתונים.

מטלה 1 – פונקצית אנרגיה של מפת רמצ'נדרן:

המאפיין $g_i(X)$ הוא זוג זוויות הפיתול ϕ,ψ של שייר i . S_i הוא טיפוס ח"א של שייר זה ו-

$$f^E\big(g_i(X)\big) = \frac{1}{n^2}$$

n, הוא פרמטר שקובע את הרזולוציה שבה אנחנו דוגמים את המרחב. בעבודה זו n=36 כך שהמרחב של 360X360 מעלות מחולק לסריג של 10X10 מעלות.

$$\forall(\Phi,\Psi) \ \in \{(-180,-180),(-170,-180),\dots.(170,170)$$

$$g_{\Phi,\Psi} = \sum e^{-\alpha\left((\mathrm{D}(\Phi,\,\varphi_i))^2+(\mathrm{D}(\Phi,\psi_i))^2\right)} + \varepsilon$$

כאשר ε הוא ספירת הדמה, ו $\mathrm{D}(x,y)$ היא פונקציית המרחק בין x ל- y. שימו לב שזוויות הן היא מחזוריות כלומר המרחק מ- 180 ל- 180- הוא אפס.

אתם מקבלים קובץ: part3_Ramachandran_energy_parameters ובו שיטת main וכמה שיטות עזר. עליכם לכתוב את הקובץ my_part3_Ramachandran_energy_parameters.py ובו השיטות:

1. calculate_ramachandran_maps(types_data, sequences, masks, n_coordinates, ca_coordinates, c_coordinates)

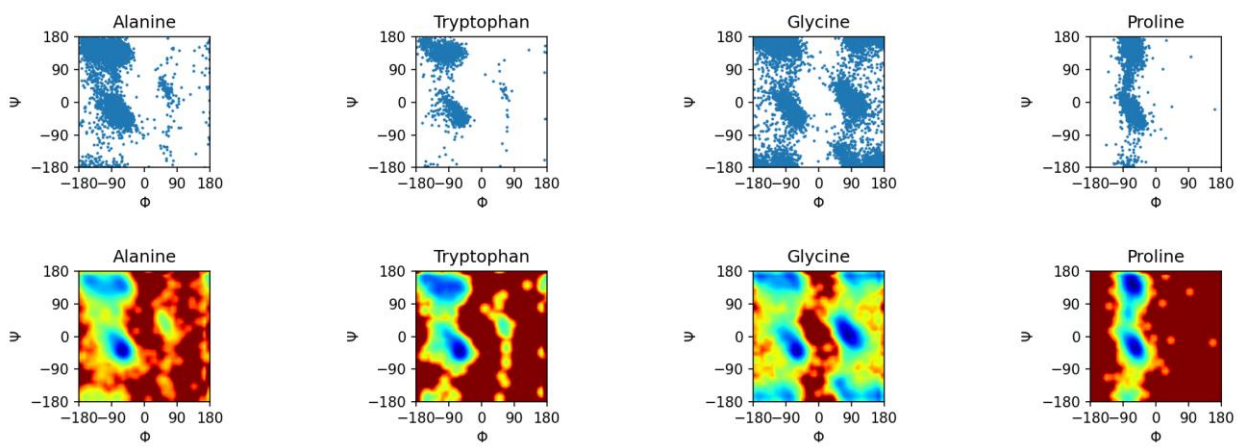
זו שיטה שאתם מכירים מעבודה 2

2. calculate_energy_parameters(2amachandran_energies, types_data)

הפרמטרים הם מבני נתונים המוגדרים ב part3_utils. אתם מקבלים את amachandran_energies2 ריק ובסוף הריצה של השיטה הוא יכיל ערכי אנרגיה לכל אחד מקודקודי הסריג, עבור כל ח"א בנפרד.

אתם מוזמנים (רצוי) להוסיף עוד שיטות עזר כדי לשפר את הקריאות.

בסוף הריצה צריכה להופיע על המסך התמונה הבאה:



צבע כחול מיצג אנרגיה נמוכה, ואדום גבוהה.

<u>מטלה 2 - פונקצית אנרגיה של מגעים בין פחמני בטא בחלבון.</u>

המאפיין $g_i(X)$ הוא סכום המגעים של שייר i , כאשר המגע מוגדר על ידי גאוסיין רחב על המרחק בין שני פחמני בטא

$g_i(X) = \sum e^{-\alpha D_{i,j}^2}$ כאשר $D_{i,j}$ הוא *המרחק בין פחמני בטא של שייר* i ו- שייר j.

S_i הוא טיפוס ח"א של שייר i. ו- $f^E(g_i(X))$ הוא הערך הממוצע של $g_i(X)$ של כל השיירים בבסיס הנתונים בלי להתחשב בטיפוס שלהם.

<u>הערות:</u>

1. לגליצין אין פחמן בטא. בעבודה זו נחליף אותו בפחמן אלפא (פתרון די סטנדרטי).
2. בלי קשר למבנה, כל שייר נמצא במגע עם עצמו ועם שני שכניו מימין ושני שכניו משמאל. אנחנו נתעלם מהמגעים האלו שמטשטשים את ההבדלים בין טיפוסי הח"א.

אתם מקבלים קובץ contact_energy.py עם שיטת main. עליכם לכתוב את הקובץ my_contact_energy.py שבו השיטות הבאות:

1. get_dm_and_cm(ca_coordinates, cb_coordinates, sequence, mask, alpha)

הפרמטרים מתייחסים לקואורדינאטות, רצף ומסכה של חלבון בודד. הפרמטר alpha נועד לנוסחת הגאוסיין למעלה. השיטה מחזירה מפת מרחקים (כמו בעבודה הראשונה) ומפת מגעים הנגזרת ממנה.

2. get_aa_contacts(contact_map, sequence, mask)

השיטה מקבלת מפת מגעים מחלבון ומחזירה שני טנזורים באורך 20. הראשון, מכיל סכומי מגעים ממוצעים לכל טיפוס ח"א בחלבון ממוינים מקטן לגדול. אם טיפוס מסוים אינו נמצא בחלבון, יוצב בטנזור זה הערך 1-. הטנזור השני מכיל אינדקסים לרשימה האלפבתית של שמות הטיפוסים כדי שנדע איזה סכום שייך לאיזה טיפוס.

3. get_contact_histogram(ca_coordinates, cb_coordinates, sequences, masks, aa_types, alpha)

השיטה מקבלת את הקואורדינטות, רצפים ומסכות של כל החלבונים, מבנה נתונים של טיפוסי ח"א ואת הפרמטר אלפא מנוסחת הגאוסיין.

השיטה מחזירה היסטוגרמה ומילון (dictionary) של היסטוגרמות. את ההיסטוגרמה יש ליצור על ידי הפקודה

```
histogram = Histogram(0, 25, 50, 200)
```

המשתמשת במחלקה המוגדרת ב part3_utils. ההיסטוגרמה הזו תכיל את ההתפלגות הכללית של שכיחויות סכומי המגעים בבסיס הנתונים.

המילון, מכיל 20 היסטוגרמות (אחת לכל טיפוס ח"א) שנוצרו על ידי הפקודות

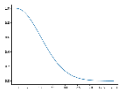
```
for aa_type in aa_types:
    histograms[aa_type] = Histogram(0, 25, 50, 1)
```

ומכילות את ההתפלגויות של כל ח"א בנפרד.

הפרמטר האחרון בבנאי של ההיסטוגרמה, הוא הערך של ספירת הדמה. הערכים של פרמטר זה בהיסטוגרמות השונות מבטיחים שהאנרגיה של מצבים שכלל אינם נצפים (יותר מדי או פחות מדי מגעים) תהיה גבוהה.

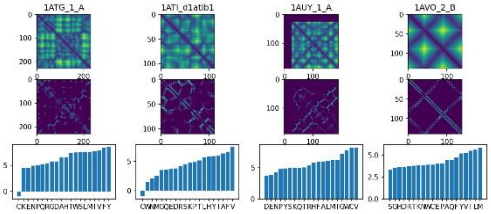
4. get_energy_parameters(contact_histogram, contact_aa_histograms)

השיטה מקבלת היסטוגרמה של סכומי המגעים של כל השיירים בבסיס הנתונים, ומילון שהמפתחות שלו הם שמות טיפוסי הח"א (אות אחת) והערכים הם היסטוגרמות של סכומי המגעים של אותו טיפוס. השיטה מחזירה מילון עם אותם מפתחות, וערכים שהם טנזורים עם ערכי אנרגיה לכל סכום מגעים.

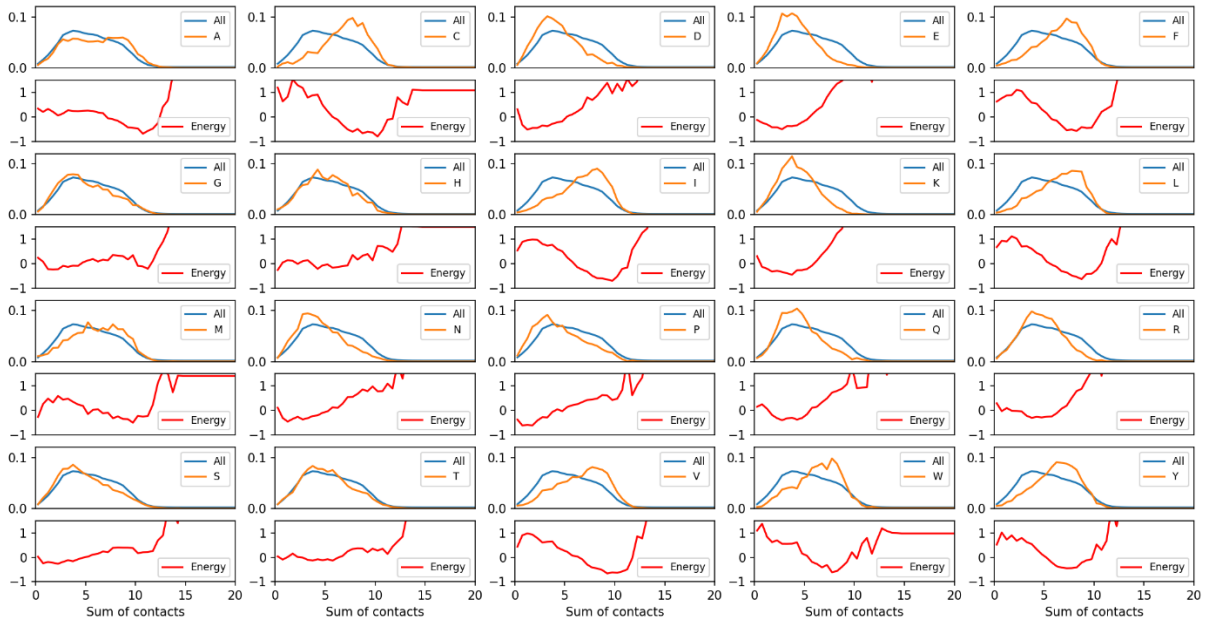


השיטה main פותחת שלושה חלונות צריך לסגור חלון כדי שיעלה החלון שאחריו.
החלון הראשון אינו דורש מכם כל עבודה ונועד רק להדגים את הגאוסיין.

החלון השני, מציג את מפות המרחקים של ארבעה חלבונים, את מפות המגעים שנגזרות מהם, ואת ההתפלגות של סכומי המגעים הממוצעים בהם. שימו לב שהמגעים הטריוויאליים לאורך האלכסון (כל שייר עם עצמו ושני שכניו הקרובים ברצף משני הכוונים) הוסרו.



החלון השלישי, מציג את ההתפלגויות של סכומי המגעים בכל טיפוסי ח"א ביחס להתפלגות הכללית (שורות אי זוגיות) וערכי האנרגיה שנגזרים מהם (שורות זוגיות)



<u>מטלה 3 – פונקצית אנרגיה זוגית על המרחקים בין פחמני בטא בחלבון.</u>

המאפיין $g_i(X)$ הוא המרחק בין זוג פחמני בטא. בהתאם, S_i הוא זוג (לא סדור) של טיפוסי ח"א. $f^E\big(g_i(X)\big)$ הוא התפלגות המרחקים של כל זוגות פחמני הבטא בבסיס הנתונים בלי להתחשב בטיפוס שלהם.

<u>הערות:</u>

1. כידוע לגליצין אין פחמן בטא. בעבודה זו נחליף אותו בפחמן אלפא (פתרון די סטנדרטי).
2. בלי קשר למבנה, כל שייר נמצא במרחק קצר עם עצמו ועם שני שכניו מימין ושני שכניו משמאל. אנחנו נתעלם מהמרחקים האלו שמטשטשים את ההבדלים בין טיפוסי הח"א.
3. בשיטה main יש שורה שמוסיפה להיסטוגרמה הכללית קריאות דמה מיוחדות עבור המרחקים הקצרים ביותר, ערכים אלו מעלים את האנרגיה של עבור מרחקים אלו.

אתם מקבלים קובץ pairwise.py ובו שיטת main. עליכם לכתוב את הקובץ my_pairWise.py שבו השיטות:

1. get_distance_map(ca_coordinates, cb_coordinates, sequence, mask)

שמחשבת ומחזירה את מטריצת המרחקים (כמו בעבודה הראשונה) בין פחמני בטא (או אלפא במקרה של גליצין).

2. filter_range(values, min_value, max_value)

שמקבלת טנזור של ערכים (מרחקים) ומחזירה טנזור שנופו ממנו הערכים שמחוץ לטווח בין ערכי המינימום והמקסימום.

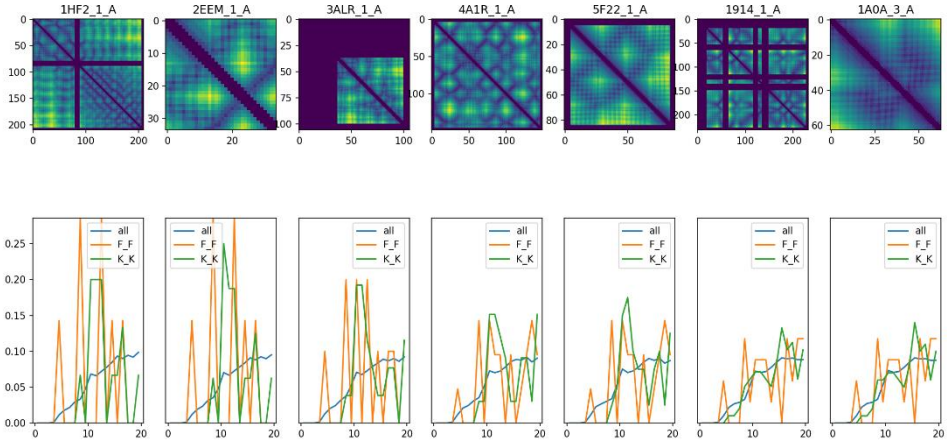
3. get_pair_wise_distances(cb_distance_map, sequence, mask, aa_type1, aa_type2)

שמקבלת מטריצת מרחקים, רצף, מסכה ושני טיפוסי ח"א, ומחזירה רק את המרחקים בין אטומים של טיפוסים אלו.

1. get_energy_parameters(contact_histogram, contact_aa_histograms)

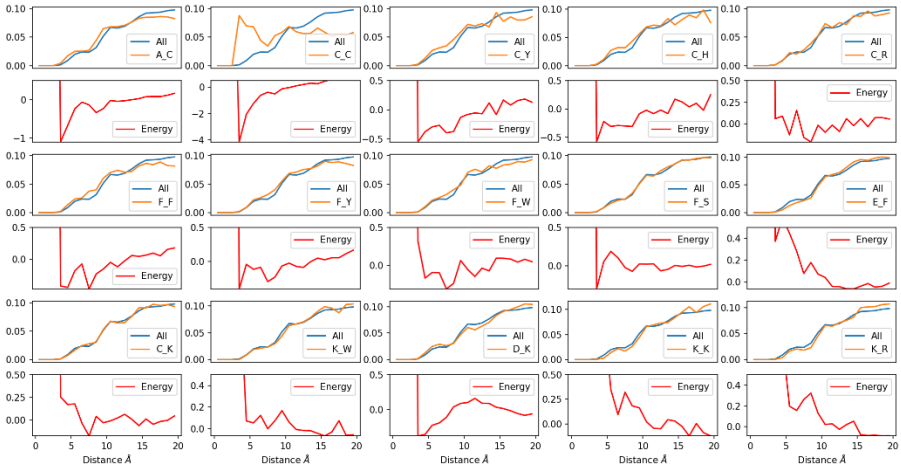
השיטה מקבלת היסטוגרמה של המרחקים בין פחמני בטא של כל זוגות השיירים בבסיס הנתונים, ומילון שהמפתחות שלו הם זוגות של שמות טיפוסי הח"א (X_Y) והערכים הם היסטוגרמות של המרחקים בין אותם טיפוסים. השיטה מחזירה מילון עם אותם מפתחות, וערכים שהם טנזורים עם ערכי אנרגיה לכל מרחק.

השיטה main פותחת שני חלונות, צריך לסגור את החלון הראשון כדי שהשני יעלה (זה לוקח כמה רגעים כי נדרשים הרבה חישובים).



החלון הראשון, מציג את מפות המרחקים של שבעה חלבונים, אחרי שהסרנו מהם את המרחקים ה "טריוויאליים" אלו של האלכסון הראשי ושני אלכסונים מעליו ומתחתיו. מתחת לכל מפת מרחקים יש היסטוגרמות של מרחקי פחמני בטא (כל זוגות השיירים בכחול, פניל-אלנין עם פניל-אלנין בכתום, וליזין עם ליזין בירוק). היסטוגרמה הראשונה נוצרה מהמפה השמאלית ביותר, ושאר ההיסטוגרמות נוספו בהדרגה משמאל לימין. שימו לב איך ככל שיש יותר נתונים הרעש יורד.

החלון השני מראה 12 דוגמאות (מתוך 210) לזוגות של ח"א. לכל זוג כזה אנחנו מראים את התפלגות המרחקים שלו בבסיס הנתונים לעומת התפלגות המרחקים הכללית (שורות אי זוגיות) ואת ערכי האנרגיה המחושבים מהיחס שבין שתי ההתפלגויות (שורות זוגיות).



מטלה 4 – מטלת קריאה: מאמר המתאר את השיטה AlphaFold2.

1. קראו את המאמר
Highly accurate protein structure prediction with AlphaFold, Jumper et. Al

לנוחותכם המאמר נמצא באתר הקורס. באתר נמצא גם הנספח הטכני (Supplementary information) של המאמר. אינכם צריכים לפרט ברמה שלו, אבל הוא עשוי לעזור לכם להבין דברים שנשמעים אבסטרקטיים מדי במאמר.

2. כתבו סיכום קצר של המאמר. הסיכום יכלול:

א. מהי הבעיה שהשיטה מנסה לפתור (פסקה)?

ב. מהם מקורות המידע העיקריים שבהם השיטה משתמשת, ומהו ההיגיון הביולוגי שמאחוריהם (2 עד 3 פסקאות) ?

ג. הסבירו את תמונה 1. התמונה מורכבת מבלוקים המחוברים בחיצים. לכל בלוק, מהו הקלט (מה משמעות הממדים) ומהו הפלט. מה משמעות החיצים החוזרים (recycling) ?

ד. תמונה 3a, מפרטת בלוק מתוך תמונה 1. בחרו בלוק פנימי אחד והסבירו אותו בקיצור: קלט, פלט, והקשר ביניהם. הסבירו את זרימת המידע ההדדית בין שני המסלולים (הייצוג של העמדת הרצפים המרובה והייצוג הזוגי).