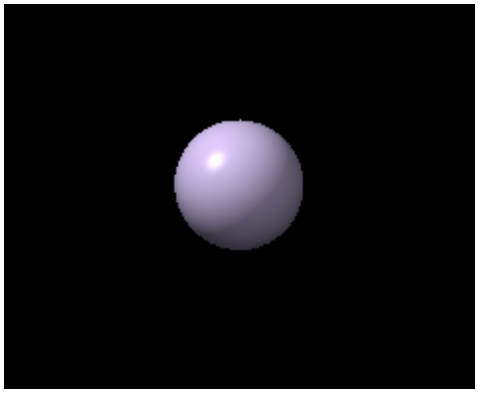
תרגיל 5 - עיבוד מידע תלת ממדי בביולגיה מבנית

רינה קרנאוך ואופק קוה

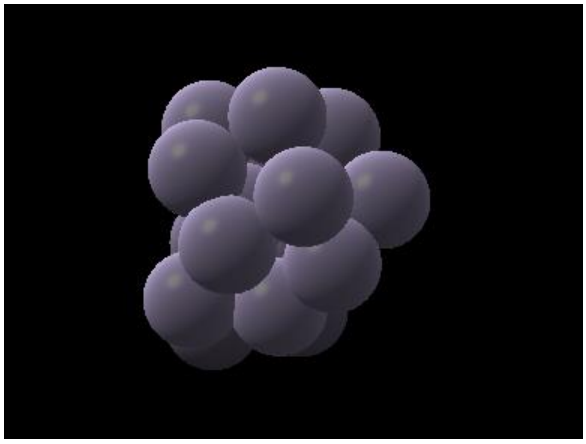
סמסטר אביב 2021

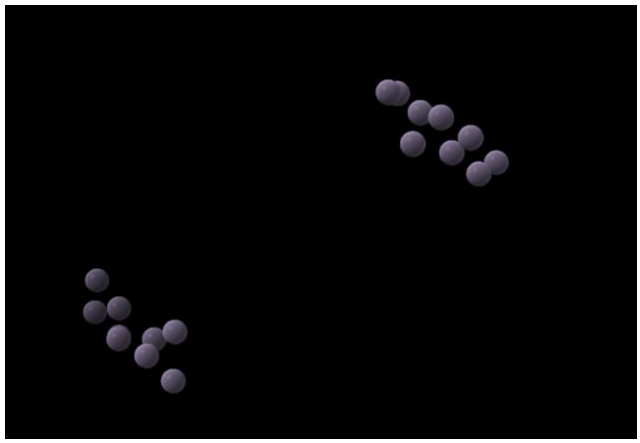
חלק 1:

A0:



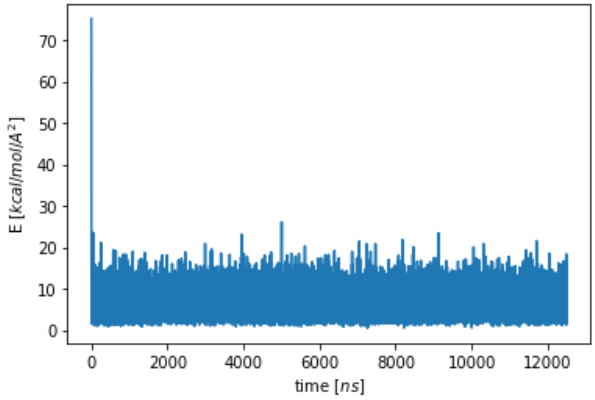
בסרט תחילה אנחנו צופים בחלקיק הroot.
לאחר frame יחיד אנחנו כבר רואים את 2 השרשראות שהוספנו, אבל בתור גוש חלקיקי.
כבר לאחר מספר פריימים רואים ממש 2 שרשראות נפרדות בעלות תנועה מתמדת, מתרחקות, בעלות אינטרקציה בין החלקיקים שלהם.
בנוסף אנחנו רואים כי השרשראות מתרחקות אך עד רף מסוים, משמע יש כוח ששומר על משיכתן. בנוסף החלקיקים בתוך השרשראות עדיין שומרים על קירוב בינם לבין עצמם אך עדיין יש תנועה מתמדת גם בתוך כל שרשרת.





A1:

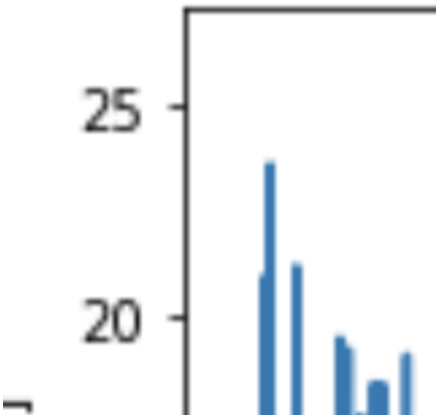
הוספנו קוד וקיבלנו את הגרף –



A2:

בשלבים ראשונים של הסימולציה קיבלנו כי הערך האנרגטי שלה מאוד גבוה. בהתחלה כל החלקיקים צמודים אל השורש, על כן אחד האילוצים שבודקים את המרחק בין מרכזי הרדיוסים מקבל ערכים גבוהים עבור כל הנקודות(כל המרכזים קרובים מאוד.

A3:



נחליף את הקבוע של הthreshold להיות 1.2359 (ניסינו הרבה מספרים), נשים לב שבעדכון הראשוני כבר אין לנו ירידה –

```
The mean energy is 6.10
The std-deviation of the energy is 2.55
```

קיבלנו את התוחלת וסטיית התקן –

אנחנו מעוניינים להבין ה יקרה לסטיית התקן של האנרגיה אם נעלה את הטמפרטורה של הסימולציה.

הנוסחא לתנועה בראונית הינה –

$$X_{n+1} = X_n + \frac{\Delta t}{k_B T} D f(X_n) + \sqrt{6D\Delta t} \cdot R$$

אם נעלה את הטמפרטורה נקבל ערך מוסף קטן יותר של $Df(X_n)$, משמע נזוז פחות (תוספת קטנה יותר של $\frac{\Delta t}{K_B T} Df(X_n)$,כלומר נישאר באותו המיקום על כל השונות תרד. לעומת זאת אם נוריד את הטמפרטורה השונות תעלה, על כן בהתאמה סטיית התקן (הרי סטיית תקן הינה שורש השונות). בנוסף גם הרצנו יחדיו עם שינוי טמפרטורה עם $T = 100$ וקיבלנו –

```
The mean energy is 6.58
The std-deviation of the energy is 6.51
```

ובנוסף עבור $T = 600$ קיבלנו –

```
The mean energy is 10.80
The std-deviation of the energy is 4.14
```

A5:

כעת נתבונן שוב על הנוסחה של תנועה בראונית –

$$X_{n+1\,T\to 0} = X_n + \overbrace{\frac{\Delta t}{K_b T} cT \cdot f(X_n)}^{\frac{\Delta t}{K_b} f(X_n)} + \overbrace{\sqrt{6cT\Delta t} \cdot R}^{\to 0}$$

על כן המיקום ישאף ל –

$$X_{n+1} = X_n + \frac{\Delta t \cdot f(X_n)}{K_b}$$

כפי שראינו השורש נעלם, כי $D = cT$ ו T שואף ל0.

A6:

במקרה בו *T=0* אז האלגוריתם שנשאף אליו הוא מונטה קרלו.

A7:

אנחנו רואים כי הגרפים שלנו למרחקים בין החלקיקים דומים לגרפים של שינוי מניות בstock market. היות ותנועות של stocks מה גם תנועות החלקיקים בתנועה בראונית הן בלתי תלויות, נקבל כי כאשר $T \to \infty$ נוכל "להפעיל" את חוק המספרים הגדולים ולקבל התפלגות נורמלית של מיקומי ה"נקודות"(stocks or particles, לא משנה) בהתאמה, על כן גם stocks וגם חלקיקים מקבלים גרפים דופים לאורך זמן מחוק המספרים הגדולים.

A9:

$$X_{n+1} = X_n + \frac{\Delta t}{k_B T} D f(X_n) + \sqrt{6D\Delta t} \cdot R$$

a. כפי שהסברנו קודם, מתפלגת גאוסיאנית מחוק המספרים הגדולים.

b. זה לא גאוסיאן מדויק היות והצעדים לא באמת בלתי תלויים לאורך הדרך, הרי X_{n+1} כן מוגדר על ידי X_n בנוסחה, הרי אחד הrestraints של תזוזת החלקיקים היא הbrownian dynamics מה שיוצר חוס תלות.

חלק 2

א. נוכל לייצג חלבונים בעזרת "מבנה ראשוני" משמע רצפי חומצות אמינו או בעזרת "מבנה שניוני" משמע הליקסי אלפא ורצועות בטא.

ב. **אלגוריתם *MCMC* עם קריטריון מטרופוליס-הייסטינגס –**

- עבור n מספר נבחר מראש של איטרציות:
  - נציע רנדומית צעד מהקונפיגורציה הנוכחית אל קונפיגורציה *C* אחרת שניתן להגיע אליה(תחת הצעדים האפשריים שאנחנו מאפשרים), כאשר הרנדומית נמדדת מהתפלגות מוצעת מראש.
  - נחשב את ההסתברות p על פי קריטריון מ.ה והנתונים הנוכחיים, כלומר נחשב –

$$\frac{\left\{\begin{matrix} probability\ from \\ proposed\ to\ current \end{matrix}\right\}}{\left\{\begin{matrix} probability\ to \\ proposed\ from\ current \end{matrix}\right\}} \cdot e^{-\frac{\Delta E}{K_B T}}$$

ונבחר את p להיות המינימום מבין 1 לערך.

  - *T* הטמפרטורה, K_B קבוע בולצמן ובנוסף השינוי הוא השינוי באנרגיה ממעבר מהמצב המוצע לנוכחי(או ההפך).
  - נקבל את הצעד המוצע בהתפלגות $Ber(p)$.

לסיכום: בכל שלב אנחנו מרנדמים קונפיגורציה אלייה נוכל להגיע מהקונפיגורציה הנוכחית, מקבלים את הצעד באופן רנדומי בהסתברות שמחושבת על ידי קריטריון מ.ה שלוקח בחשבון את ההסתברות לצעד מהמצב הנוכחי למוצע וההפך, במכפלה באקספוננט בחזקה שלוקחת בחשבון את שינוי האנרגייה, הטמפרטורה וקבוע בולצמן, כל זאת כדי שעבור n איטרציות מספר ששואף לאינסוף נקבל שאיפה להתפלגות בולצמן כדי לקבל באינסוף את הקונפיגורציה מהאנרגיה המינימלית.

ג. עבור כל שכבה נדון מה תפקידה ברשת הנוירונים –

a. **שכבת הקונבולוציה** – מבצעת קונבולוציה על שכבת הinput בעזרת מימד יחיד, כדי ליצור "מתיחה" של המידע כoutput,

בקיצור: מותחת את המידע בעזרת קונבולוציה, אבל משמרת את הממד הנוכחי.

b. **שכבת הdropout** – שכבה שבאופן רנדומי מספקת משקל אפסי אל משקולות מסוימות ברשת, מה שמונע overfit.

c. **שכבת האקטיבציה** – מפעילה פונקציה שאיננה לינארית על הinput אל השכבה. האינטרס מאחורי שכבה זו הוא שהמידע "בטבע" איננו לינארי ברובו.

ד. נדון ב*GD* וברשתות נוירונים:

a. step-size – הערך שמוכפל בגרדיאנט הפונקציה, על מנת להגיע למינימום לממזער הפונקציה f אותה מנסים למזער. כלומר- הגרדיאנט מוכפל בגודל הצעד, זהו הערך שנחסיר מהערך הנוכחי לערך החשוד הממזער הבא , כאשר באינטואיציה היא שצעד קטן במורד הגרדיאנט יספק לנו את x הממזער: באיטרציה t כלשהיא של האלגוריתם, נחשב:

$$x^{(t+1)} = x^{(t)} - \eta_{t+1}\nabla(f\left(x^{(t)}\right)$$

כאשר f הפונקציה אותה ממזערים ואנו מחפשים את הממזער, כאשר η_{t+1} מספקת לנו את גודל הצעד בכיוון הירידה של הגרדיאנט שעוזרת לנו למזער את ערך הפונקציה בעזרת הנקודה.

i. מה אם הוא גדול מדי? האלגוריתם עלול לפספס את המינימום הלוקלי בגלל קפיצות גדולות מדי "מצד לצד" שביניהם המינימום שאנו מחפשים.

ii. מה אם הוא קטן מדי? האלגוריתם עלול לשאוף אל המינימום הלוקלי(או גלובלי אם הפונקציה קמורה) מאוד לאט, משמע זמן ריצת האלגוריתם יהיה ארוך.

ה. נכתוב את הקלט והפלט של השיטות מביולוגיה מבנית שלמדנו בקורס.

a. *Protein structural alignment* –

קלט – זוג שרשורי חומצות אמינו s, t מאורכים זהים.

פלט – מציאת אחוז הדמיון בין השרשורים (הגבוה ביותר), ומציאת ההתאמה עצמה בזוג השרשורים שמספק את האחוז כנ"ל(בשימוש בהחלפת אותיות או הכנסה-מחיקה של אותיות).

b. *ab-initio folding* –

קלט – קבלת שרשור החלבון s, מרחב קונפיגורציות אפשרי *C* ופונקציית אנרגיה מ*C*, אל הממשיים, כאשר אין ידע מוקדם על פתירת המבנה.

פלט – קונפיגורציה c ב*C* שמספקת לנו מינימיזציה של פונקציית האנרגיה.

c. *comparative modeling* -

קלט – ידע מוקדם(טמפלטים) של מבנים ידועים מראש, שרשור של החלבון אותו נרצה למדל. לעיתים גם נקבל restraints עבור המבנה, או חוקים פיזיקלים מוגדרים מראש וסטטיסטיקות על מבנים.

פלט – קונפיגורציה שעומדת באילוצים(אם נתנו) ומקבלת אחוז alignment גבוה בין הטמפלטים הידועים מראש אל המבנה שניסינו למדל.

d. k-means clustering –

קלט – דאטה על m נתונים, k כמות קלסטרים אליו נרצה לקלסטר.

פלט – חלוקת הדאטה אל k קבוצות של המידע בגודל m שלנו, בעזרת "מרכזי מאסה" של המידע.

e. protein-protein docking –

קלט – זוגות של מולקולות (רצפטור, ולגאנד) שמוצגם בעזרת המבנה התלת מימדי שלהם(קורדינטות x,y,z).

פלט – חיזוי של המבנה התלת ממדי שלהם בתור יחידה אחת(שמורכבת משתי המולקולות)(בעזרת חישוב טרנספורמציה לאוריינטציה שלהם) ומתן "ציון" להתאמתם לבנה יחיד.

f. molecular dynamics -

קלט – חלקיקים במרחב קונפיגורציה, הגדרה לאינטראקציות והדינמיקה בין החלקיקים(כמובן שגם קבועים שנדרוש לחישובים).

פלט – בכל שלב באלגוריתם נקבל מיקום חדש לאטומים בעזרת חישוב הכוחות והתזוזה שנובעת מהכוחות עקב האינטראקציה והדינמיקה שניתנה על החלקיקים כנ"ל.

g. motion-planning –

קלט – "מפה", מרחב קונפיגורציות. קונפיגורציה התחלתית וסופית.

פלט – "דרך" בין הקונפיגורציות, שעוברת דרך הקונפיגורציות האפשריות, אליהן נעבור ממצב למצב תוך מעבר מקונפיגורציות בעלות אנרגיה נמוכה, תחת אילוצים שנובעים מהמרחב(לא נרצה אנרגיה גבוהה).

h. trRosetta – protein folding using DL

קלט – רצף של חומצות אמינו (*protein sequence*) או *multiple sequence alignment*

פלט – המבנה התלת מימדי הצפוי של החלבון (*predicted 3d structure model*)