עיבוד תלת ממדי במבנים ביולוגיים 76562

דינה שניידמן וברק רווה

<u>הרצאה 1</u>

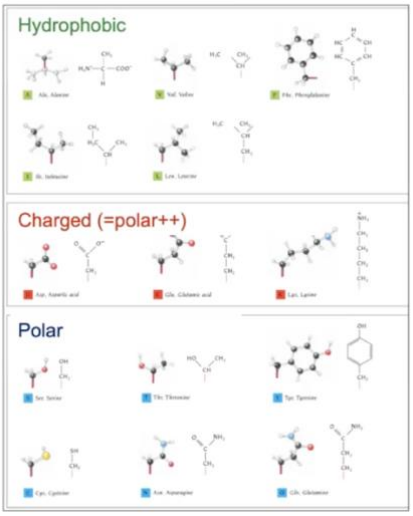
<u>הרצאה 2</u>

<u>הרצאה 3</u>

<u>הרצאה 4</u>

<u>הרצאה 5 – **שיטות אופטימיזציה עבור שיטות קיפול חלבונים חלק א'**</u>

קיפול חלבונים הוא התהליך שבו רצף של חומצות אמינו יוצר צורה תלת ממדית. חלבונים מתקמפלים במילי-שניות שזהו פרדוקס בפני עצמו.



פרדוקס לבנטל

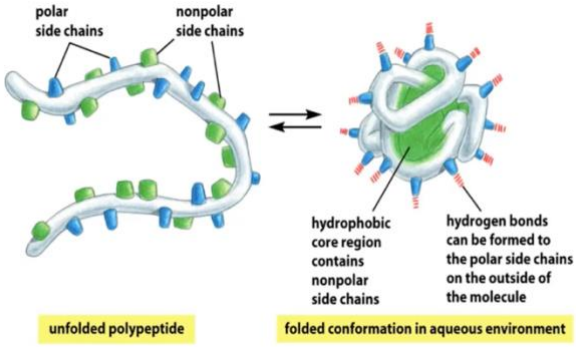
נתייחס אל חלבון קטן שמורכב מ100 חומצות אמינו. יהיו לנו 99 קשרי פפטידים ו198 זוויות דאדרליות ϕ, ψ.

נניח ויש לנו 3 מצבים אפשריים עבור כל זווית דאדרלית, יש לנו 3^{198} אפשרויות לקיפול המבנה, אבל חלבונים מתקפלים מהר – פרדוקס בפני עצמו.

נגדיר מספר הגדרות –

- הידרופוביות – לא נוטות ליצור קשרי מימן עם מים.
- פולאריות – טעונות במטען חשמלי, "אוהבות מים".

יש לנו חומצות אמינו הידרופוביות ויש לנו חומצות אמינו פולאריות, והרי חלבונים מורכבים מחומצות אמינו. מה יקרה אם חלבון שמכיל חומצות אמינו הידרופוביות ופולאריות? חלבונים הרי חיים במים, על כן החומצות ההידרופוביות ירצו להיות בפנים והפולאריות ירצו להיות "בחוץ", ולכן בהסבר פשוט נאמר שזו הסיבה שחלבונים מתקפלים. בפועל כאשר חומצות הידרופוביות מופנות כלפיי פנים האנרגיה של המבנה יורדת, ובאותה המידה כאשר חומצות פולאריות מופנות כלפי מים האנרגיה של המבנה יורדת. מערכות ביולוגיות אוהבות להיות במצב של אנרגיה מזערית – לכן חלבון מתקפל.



<u>בעיית קיפול חלבונים ניתנת להסתכלות בתור בעיה אלגוריתמית.</u>

קלט – רצף של חומצות אמינו.

פלט – מבנה תלת ממדי.

יש לנו 2 גישות מרכזיות למידול החלבונים בקיפול –

1. מידול הומולוגי(קומפרטיבי)
2. מידול ab initio (בלטינית: מהתחלה)

בצורת מידול זו נקבל רק את הרצף וממנו נרצה לקבל את המבנה.

נדבר על המידול ab inition. השיטות תחת המידול כנ"ל מנסות למצוא את המבנה הטבעי(הקיפול הטבעי) של חלבון על ידי סימלוץ התהליכים הביולוגים בקיפול החלבון.

+ יתרונות בגישה – יותר כללי, ניתן להרצה גם כאשר אין מודל הומולוגי מוכר לחלבון.

– חסרונות בגישה – הנכונות והסימוכין שלנו על המידול שמסתפק איננו גבוה.

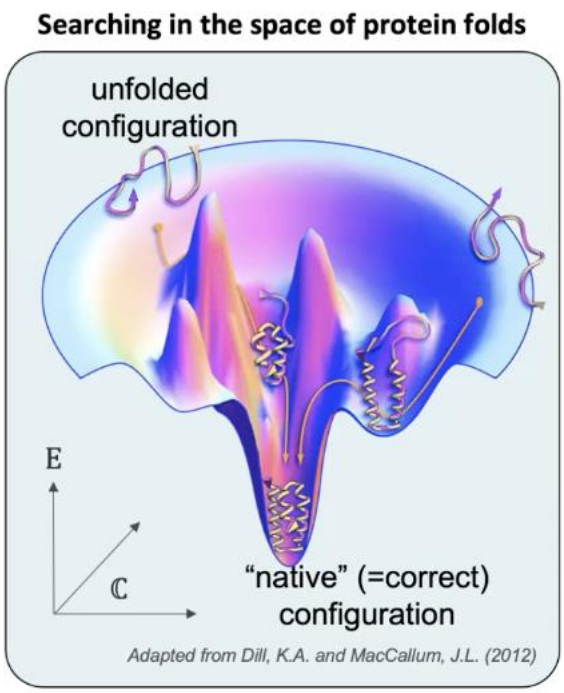
הגדרה – Configuration space

מרחב ווקטורי $\mathbb{C}$ עם סט של קואורדינטת שמתאר את החלבון(או מערכת כלשהיא). ווקטורים $c \in \mathbb{C}$ ממפה מספר מערכות(רובוט, חלבון, פוזיציה של גוף האדם).

הקואורדינטות של $c \in \mathbb{C}$ בד"כ נקראות "system variables" כלומר משתני מערכת. ניתן לקרוא להם גם פרמטרים או קואורדינטות כלליות.

חלבון עם 1000 אטומים, ממד מרחב הקונפיגורציות שלו במרחב קרטזי הוא 3000, כי עבור כל אטום יש תצוגה של 3 קואורדינטות. נזכיר כי יש לנו 6 דרגות חופש שנוכל להוריד(כי לא משנה אם מזיזים או מסובבים את החלבון) ולכן נוכל לדייק ולומר 2994 קואורדינטות בלתי תלויות.

- למדנו על תצוגות שונות אפשריות של חלבונים, ביניהן ייצוג קואורדינטות פנימיות(זוויות דאדרליות). נניח ויש לנו 200 חומצות אמינו, לכל חומצת אמינו זווית ϕ, ψ, ולכן ממד מרחב הקונפיגורציות הוא 398.



נרצה להגדיר את בעיית ה-ab initio בתור בעיית חיפוש.

קלט – רצף החלבון s, פונקציית אנרגיה $E : c \to \mathbb{R}$ ממרחב הקונפיגורציות לממשיים. כל קונפיגורציה תקבל ערך סקלרי.

פלט – קונפיגורציה $c \in \mathbb{C}$ שמספקת לנו מינימום עבור $E(c)$.

למעשה אנחנו רוצים למצוא מינימום לפונקציית האנרגיה הפוטנציאלית של החלבון המקופל.

למה נרצה למזער את האנרגיה של הקונפיגורציה? כי מערכות ביולוגיות אוהבות להיות באנרגיה נמוכה ככל האפשר.

<u>איך נבנה פונקציית אנרגיה?</u>

האטומים מיוצגים בתור "כדורים". נרצה שהם לא יתנגשו, נוכל לעשות זאת בעזרת מינימיזציה של סכום פונקציית אינדיקטור עבור כל זוג.

היינו יכולים לעשות זאת בעזרת מרחק אוקלידי: בעזרת העלאת פונקציית האנרגיה תחת מרחק אוקלידי מתאפס של זוג אטומים(הרי כך הם מתנגשים).

נניח ויש לנו 2 מרכזי כדורים –

$$B_1 : (c_1, c_2)$$
$$B_2 : (c_3, c_4)$$

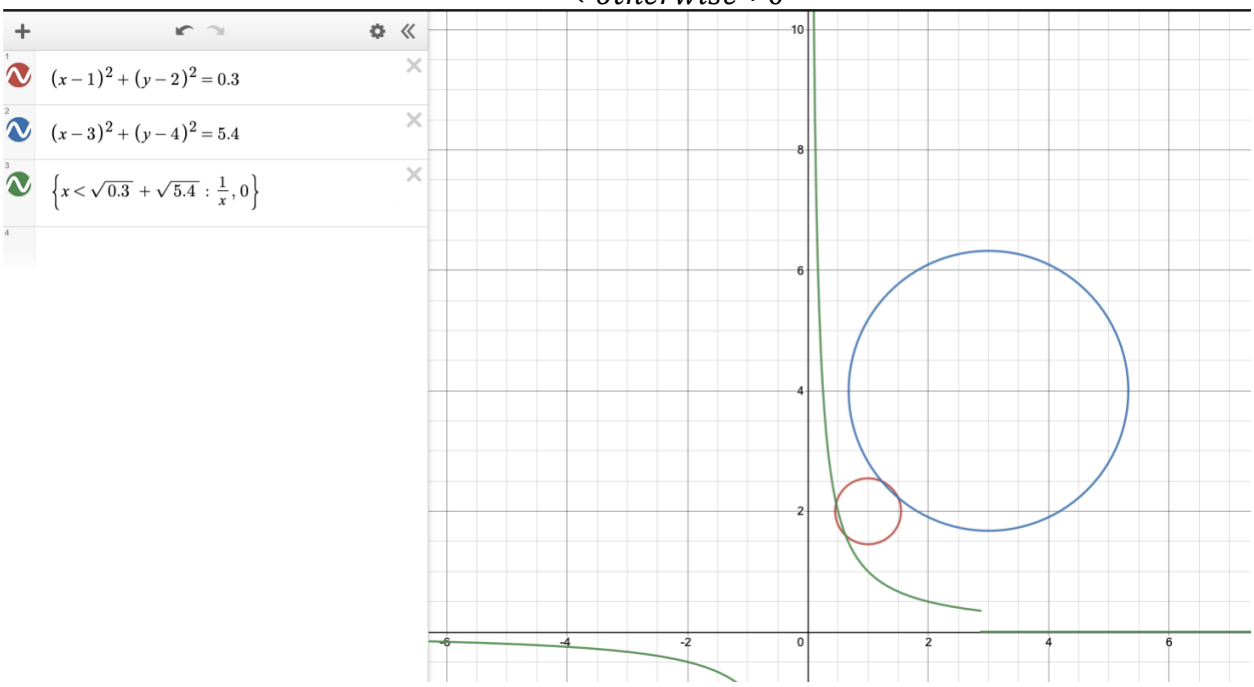
נרצה פונקציה רציפה $E(c)$. נתבונן על המרחק בין 2 המרכזים –

$$d = dist(B_1, B_2)$$

נוכל להסתכל על הגרף – $E(c) = \frac{1}{d}$, גרף רציף ומונוטוני יורד, הרי נרצה לקבל "מינימום" לפונקציה כדי שהכדורים לא יתנגשו.

נניח ורדיוסי הכדורים הם r_1, r_2 בהתאמה. נרצה להגדיר את E כך –

$$E(c) = \begin{cases} d < r_1 + r_2 : \frac{1}{d} \\ otherwise : 0 \end{cases}$$



כאשר במקרה כנ"ל אמנם d קבוע וכך נראת פונקציית האנרגיה תחת קונפיגורציה ספציפית, אבל מיקומי מרכזי הכדורים אינם קבועים כמובן כי יש לנו מצבור הקונפיגורציות.

<u>בפועל נבנה את פונקציית האנרגיה בעזרת</u> –

- שיקולים פיזיקליים
- סטטיסטיקה
- למידת מכונה

ניתן להראות כי בעיית קיפול החלבונים היא בעיה *NP* קשה אפילו ב*2D*, אבל באופן מעשי אנחנו יכולים לפתור את הבעיה באמצעות מרחבים אנרגטיים.

שיטות מקובלות הן *Molecular Dynamics, Monte Carlo* וגם בעזרת למידה עמוקה.

אנחנו נדון **בדינמיקה מולקולרית** בהמשך הקורס, בקצרה אנחנו מנסים לסמלץ את החוקים של ניוטון. אנחנו מנסים לסמלץ על פי החוקים של ניוטון את תזוזת האטומים, אך הבעיה בפועל היא שחלבונים תמיד נתקעים בנקודות מינימום לוקליים. איך נתקעים במינימום לוקליים? מגרילים קונפיגורציה רנדומית ובודקים את האנרגיה שלה. ננסה לשפר את האנרגיה של החלבון עד שנגיע ל-"מבנה הנכון", אבל במקרה כנ"ל אנחנו עלולים להיתקע במינימום לוקלי.

לעומת זאת **נדון ב-*Monte Carlo*** בשיעורים הקרובים. באופן מקוצר אנחנו דוגמים קונפיגורציות לא בהכרח בעזרת חוקי הפיזיקה. אנחנו מסמלצים את הקיפול של החלבון אבל לא בצורה שמנסה לחכות את הפיזיקה.

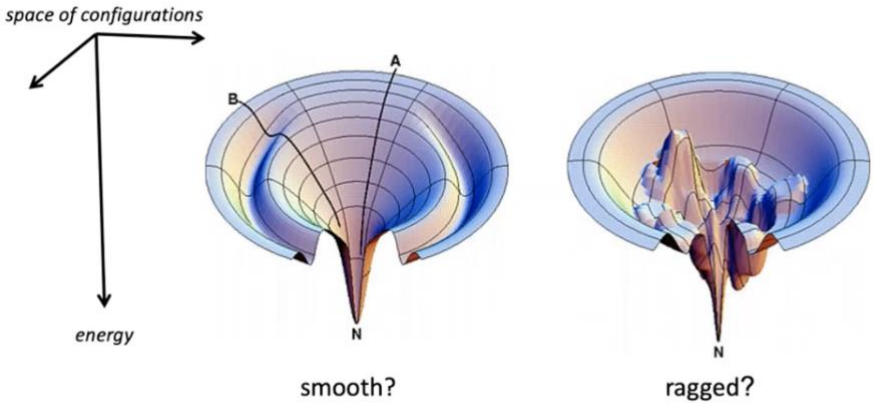
השיטה יותר פרקטית, ונפוצה גם במידול הסתברותי, באסיאני ותחומים נוספים. יש לה פחות סיכוי להיתקע במינימום לוקלי אך לא בהכרח שלא תיתקע.

<u>שיטות מינימיזציה לוקלית</u>

- *גרדיאנטים ונגזרות*

נרצה שפונקציית האנרגיה שלנו תהיה בקורלציה לפונקציית האנרגיה החופשית. בנוסף נרצה שהפונקציות יחלקו את אותו מינימום גלובלי אנרגטי, מה גם נצטרך גם אסטרטגיית דגימה מבין כל הקונפיגורציות הקיימות. בהרחבת פונקציות לממדים גבוהים, נרצה לסרוק את כל הקונפיגורציות בכמה ממדים שיש לנו ולמצוא את הקונפיגורציה האופטימלית.

משטחי אנרגיה הם לא חלקים – הם מפוספסים, אז קשה לנו למצוא את המינימום הגלובלי.



החספוס יגרום לנו להיתקע בנקודות מינימום לוקליים מאוד מהר, בעזרת ההיתקעות שלנו נרצה לשפר את מבנה החלבון ולא רק לקפלו.

כזכור גרדיאנטים והסיאנים מכלילים את הנגזרות הראשונה והשנייה עבור פונקציות רב ממדיות.

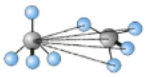
בהינתן פונקציית האנרגיה שלנו $Energy = E(x_1, y_1, z_1, \dots, x_n, y_n, z_n)$. אם פונקציית אנרגיה שלנו רציפה, בהנחה על רציפותה נוכל לגזור אותה ולהוציא ממנה את ההסיאן והגרדיאנט.

<u>עבור מה ישמש הגרדיאנט?</u>

דוגמא – אנרגיית ואן-דר-ואלס בין זוגות אטומים עבור $O(n^2)$ אטומים –

Example:

Van der-Waals energy between pairs of atoms – O(n²) pairs:



$$E_{VdW} = \sum_{i,j} \frac{A}{\bar{R}_{ij}^{12}} - \frac{B}{\bar{R}_{ij}^{6}} \Rightarrow \frac{\partial E_{VdW}}{\partial R_{ij}} = \frac{-12A}{R_{ij}^{13}} + \frac{6B}{R_{ij}^{7}}$$

$$R_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

מימין חישבנו את הנגזרת של פונקציית האנרגיה של ה*VDW* בין שני אטומים j, i, כאשר R_{ij} המרחק בין שני אטומים, ו *A,B* קבועים.

מדובר על הוצאת גרדיאנט באופן אנליטי – על פי הכללים, אך נוכל להוציא גם מבחינה נומרית את הגרדיאנט שלנו.

נוכל להעריך על פי ההגדרה של נגזרת בהכללה גסה. מבחינה אינטואיטיבית –

$$f'(x) \approx \frac{\big(f(x+1) - f(x)\big)}{x+1-x} = f(x+1) - f(x)$$

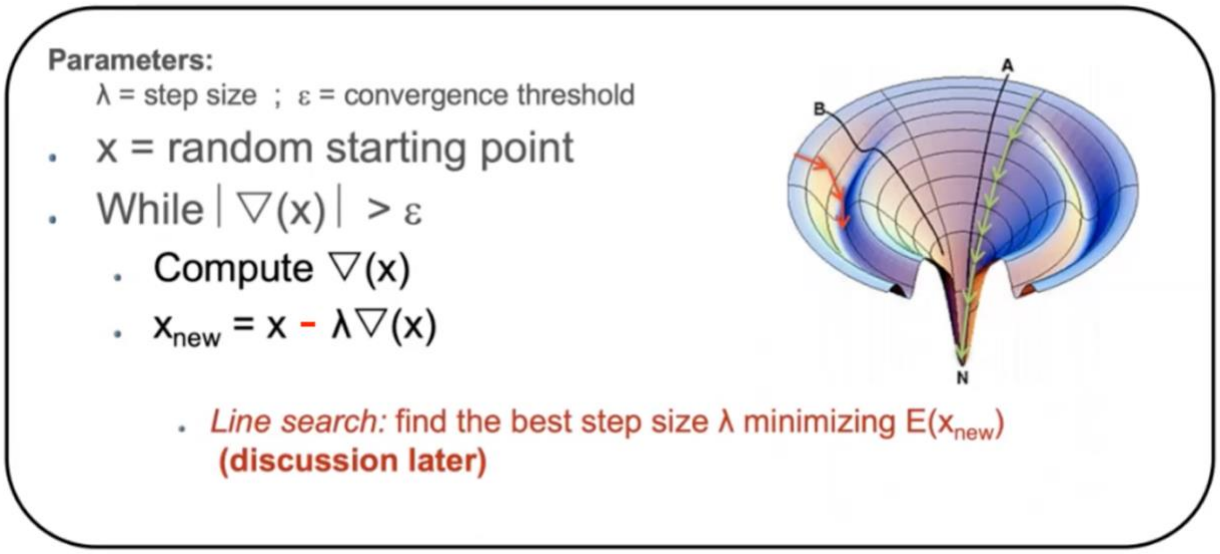
<u>הגרדיאנט</u>

שיטת ה"gradient descent" הינו האלגוריתם הבסיסי ביותר למציאת מינימום אנרגטי תוך שימוש בחישוב גרדיאנטים.

הנחות היסוד שלנו הן שיש לנו –

1. מרחב קונפיגורציות
2. פונקציית אנרגיה דיפרנציאבילית
3. מימוש שמאפשר לגזור את פונקציית האנרגיה בשביל הוצאת גרדיאנט

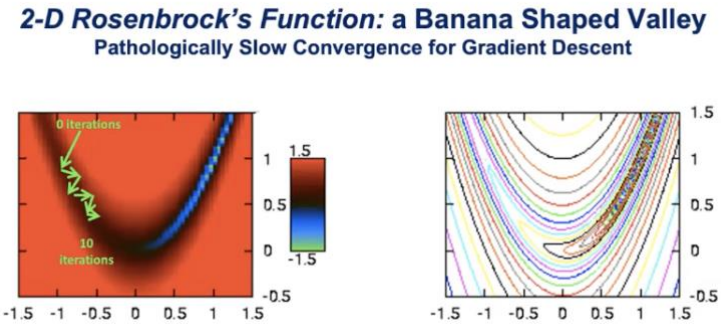
במקרה שבו יש לנו את הנ"ל האלגוריתם נראה כך –



- איך נבחר את הפרמטר λ שלנו? נפעיל אסטרטגיית line-search כדי למצוא את גודל הצעד הטוב ביותר כדי להקטין את E למינימום.
- למה אנחנו ממשיכים כל עוד הגרדיאנט גדול מ ϵ? הגרדיאנט בנקודת מינימום/מקסימום הוא אפס. בסיום האלגוריתם מובטח לנו שהגרדיאנט אפסי(לא מובטח שמדובר במינימה, יכולה להיות גם נקודת פיתול).

<u>מה הבעיה בעצם ב*Gradient-Descent*?</u>

לא רק שאנחנו יכולים להיתקע בנקודת מינימום לוקלי ולא גלובלית, האלגוריתם אמור להיות מאוד איטי. לדוגמא – פונקציית רוזנברוק. הפונקציה תעשה "דיגזג" מאוד ארוך וייקח זמן רב להגיע למינימום.



פתרון לבעיה – "conjugate gradient descent". אדפטציה לאלגוריתם שתשמור מידע על הצעדים הקודמים. כאשר המשטח ריבועי תהיה התכנסות טובה מאוד של האדפטציה כנ"ל.

הרצאה 6 – **מציאת שורשים ושיטת מונטה קרלו**