**חלק ב – אנליזת חלבונים בעזרת אתר ה UniProt**

**סעיף א**הסבר: בסעיף זה התבקשנו להצליב בין הנתונים שיש לנו בקובץ ה GeneBank לבין הקובץ שהורדנו מ UniProt. את ההצלבה ביצעתי ע"י המזהה 'GOA'. מזהה זה הינו משותף ל2 המאגרים ומזהה חלבונים באופן ייחודי.  
דרך: על מנת לבצע את ההצלבה באופן האידיאלי מבחינת זמני ריצה, הכנתי 2 מילונים. מיליון עבור הערכים מ UniProt, ועבור הערכים מ GeneBank. בחרתי במילון על מנת לבצע שליפה ב O(1) על מנת לבצע את פעולת ההצלבה המלאה ב O(n). המפתח במילונים אלו היה המזהה 'GOA' ו Name והערכים היו רצף הגנים. כלומר סהכ 4 מילונים.  
תוצאות:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| מקור | ערכים קיימים | ערכים עם מזהה GOA | ערכים עם מזהה Name | הצלבה ע"פ GOA | הצלבה ע"פ Name |
| UniProt | 8541 | 8541 | 6938 | 3381 | 3715 |
| GeneBank | 4237 | 3381 | 4171 | 3381 | 3715 |

הערות: נראה כי הרישומים הקיימים ב GeneBank עבור החלבונים - חסרים ולא מושלמים. כאשר קיימים חלבונים שלא כוללים את המזהה 'GOA'. להערכתי המזהים החסרים נובעים מכך שמאגר ה GeneBank עיקרו הינו גנים ולא חלבונים, לעומת UniProt שזה עיקר ייעודו. בנוסף אף על פי שיש מיפוי רחב יותר של GOA ב UniProt ההצלבה המוצלחת יותר בוצעה ע"י שם החלבון (כנדרש בתרגיל).  
צילומי מסך: ראה נספח **חלק ב(א)**

**סעיף ב**הסבר: בסעיף זה התבקשנו לזהות את החלבונים המכילים חלקים טרנסממברנלי,ולאפיין אותם.  
דרך: בעזרת מאגר UniProt הוספתי עמודה בשם Transmembrane, ע"פ עמודה זו ניתן לזהות את החלבונים המכילים את האיזורים האלו.  
תוצאות: עבור חלבונים אלו -> האורך הממוצע - 238.58, האורך המקסימלי- 493, האורך המינימלי- 134. צילום מסך של ההיסטוגרמה ניתן למצוא ב**נספח חלק ב(ב).**הערך הממוצע של חומצות האמינו ההידרופוביות בחלבונים אלו הינו : 0.496, זוהי כמות נכבדת מאוד של חומצות אמינו הידרופוביות, וזה אכן תואם את הציפיות שלנו מחלבונים המכילים מקטעים טרנסממברנלי. פירוט על כל חלבון ניתן לראות ב**נספח חלק ב(ב)**

**סעיף ג**הסבר: בסעיף זה נדרשנו לנתח את רצפי הגנים המקודדים (CDS), עם השוואה בין טרנסממברנלי לרגילים.  
דרך: סיוע בספריית Matplolib על מנת להכין את ההיסטוגרמות.  
תוצאות: את התוצאות ניתן לראות בנספח **חלק ב(ג)**

**נספח חלק ב(א)**

להלן פלט של התכנית, הערה חשובה את gn len הינו מספר החלבונים המכילים את המזהה 'GOA'

Text

Description automatically generated

**נספח חלק ב(ב)**היסטוגרמה עבור האורכים של החלבונים המכילים קטעים מסוג טרנסממברנלי

Chart, histogram

Description automatically generated

התפלגות חומצות האמינו על-פי החלבונים המכילים מקטעים טרנסממברנלי  
{'yrbG': 0.44954128440366975, 'ydbS': 0.5094339622641509, 'ydfS': 0.4723404255319149, 'ydbT': 0.44421906693711966, 'yshB': 0.5706214689265536, 'yxlG': 0.5077519379844961, 'ydfR': 0.4888888888888889, 'yetF': 0.42857142857142855, 'yshE': 0.4626865671641791, 'ykjA': 0.4732510288065844, 'yoaK': 0.5911111111111111, 'ecfT': 0.5622641509433962}

**נספח חלק ב(ג)**מניסוח השאלה לא היה ברור האם הנתונים הרצויים הינם עבור המופעים של GC או עבור האורך של החלבונים. ולכן עשיתי את שניהם:

עבור האורכים:Graphical user interface, chart, histogram

Description automatically generated

עבור ה GC:  
Chart, bar chart

Description automatically generated  
נתונים נוספים:

|  |  |
| --- | --- |
|  |  |