**סקירה כוללת**

התוכנית מורכבת משני קבצי Python עיקריים: template.py ו-part\_A.py. ה-template.py משמש כמודול בסיסי המספק פונקציונליות כללית לניתוח קבצי GenBank, חילוץ נתונים לפורמט מובנה, אימות תרגומים והצגה חזותית של נתונים. ה-part\_A.py מתבסס על template.py כדי לבצע ניתוחים ספציפיים על נתונים גנומיים של Bacillus clausii.

**template.py - היסוד**

קובץ זה מכיל פונקציות עזר ומשתנים קבועים לעיבוד נתונים גנומי למטרות כלליות.

* נתיבים ומשתנים קבועים של נתונים: הוגדרו נתיבים לתיקיות ורשימות של חומצות אמינו הידרופוביות ונוקלאוטידים של DNA.
* GenBank Parsing: פונקציות לקריאה וניתוח של קבצי GenBank, המרתה ל-panda DataFrames בשביל מניפולציה קלה (כולל מיצוי מפורט של תכונות ורצפים של גנים).
* ולידציה ופונקציות סטטיסטיות: מומשו פונקציות לאימות תרגומי גנים וחישוב סיכומים סטטיסטיים של תכונות גנומיות.
* כלי הדמיה: כולל פונקציות להפקת היסטוגרמות כדי להמחיש את התפלגות אורכי הגנים ותכונות גנומיות אחרות.

**part\_A.py - לוגיקה ספציפית ליישום**

בנוי על template.py, קובץ זה מגדיר מחלקה של BacillusClausiiGB כדי לעבד ולנתח את הנתונים הגנומיים של Bacillus clausii.

* אתחול ועיבוד נתונים מראש: המחלקה קוראת את קובץ ה-GenBank שצוין, מחלצת ומאחסנת את הרצף הגנומי, תכונות הגנים והמטא-נתונים בתכונות מובנות מראש לגישה קלה.
* שיטות ניתוח נתונים: מימוש פונקציות שונות לניתוח הנתונים הגנומיים, אשר כוללים ייצוא נתונים ל-CSV, מציאת רצפי קידוד חלבוניים (CDS), חישוב אורכי גנים והפרדת גנים לפי סוג הגדיל שלהם.
* ניתוח סטטיסטי והדמיה: מימוש שיטות לחישוב סטטיסטיקות על אורכי גנים, יצירת היסטוגרמות לניתוח חזותי, וחישוב האחוז של נוקלאוטידים 'C' ו-'T' בתוך הרצף הגנומי.
* ניתוח אחוזי CT: מימוש שיטות מיוחדות שמטרתן לנתח ולהמחיש את תוכן ה-CT על פני סוגי גנים שונים, תוך הדגשת גנים עם אחוזי ה-CT הגבוהים והנמוכים ביותר.

**עקרונות של הנדסת תוכנה**

העיצוב תומך במספר עקרונות מרכזיים של הנדסת תוכנה:

* מודולריות: ההפרדה ל-template.py ו-part\_A.py מאפשרת תיחום ברור בין פונקציונליות למטרות כלליות והיגיון ספציפי ליישום.
* שימוש חוזר: פונקציות ומחלקות מיועדות לשימוש חוזר עבור משימות דומות עם נתונים שונים, תוך מזעור שכפול קוד.
* הפשטה: פרטים של עיבוד וניתוח נתונים מובלעים בתוך פונקציות ומחלקות, וחושפים רק ממשקים נחוצים למשתמש הקצה. זה מקל על ההבנה והשימוש בתוכנית.
* גמישות: התוכנית בנויה כך שתוכל להכיל בקלות ניתוחים נוספים או שינויים קיימים ללא שינוי מבני משמעותי.
* טיפול בשגיאות: כולל בדיקות וטיפול בחריגים לניהול של קלט או בעיות בלתי צפויות במהלך ניתוח קבצים ועיבוד נתונים.

**זרימה לוגית**

1. אתחול: התוכנית מתחילה ביצירת מופע של BacillusClausiiGB, שמנתח מיד את קובץ GenBank שסופק.
2. הכנת נתונים: לאחר מכן הוא מעבד מראש את הנתונים כדי לזהות רצפים מקודדים ולא מקודדים, להפריד גנים לפי גדיל ולחשב אורכי גנים.
3. ניתוח: מתבצעים ניתוחים שונים, כולל סיכומים סטטיסטיים, הדמיה של התפלגות גנים וניתוח תוכן CT.
4. יצירת פלט: לבסוף, התוכנית מפיקה את הניתוחים דרך הדפסות, היסטוגרמות וקבצי CSV להמשך סקירה.