**דוח מסכם- סדנה לפרויקטים**

**שמות המגישים- עידן קורנפלד (322525627), רועי שלוסברג ()**

**תקציר**

חלבונים קושרי DNA ממלאים תפקידים קריטיים בתהליכים ביולוגיים מרכזיים כגון בקרה על ביטוי גנים, שכפול גנומי, תיקון DNA ואריזת DNA. הבנה טובה יותר של מנגנוני הקישור מאפשרת חקר מסלולים תאיים, איתור מוטציות בעלות משמעות קלינית, ופיתוח תרופות ממוקדות למחלות הקשורות לבקרת ביטוי גנים.

יכולת הקישור ל- DNA נקבעת לרוב על פי מאפיינים מבניים ותפקודיים הנגזרים מרצף חומצות האמינו. לכן, ניתן לפתח מודלים חישוביים המזהים דפוסים ברצף החלבון לצורך חיזוי קישור ל-DNA. עם זאת, מדובר במשימה מאתגרת במיוחד - שכן הקשר בין הרצף הראשוני למבנה התלת־ממדי ולפונקציה הביולוגית מורכב, ודורש מודלים מתקדמים המאפשרים הכללה טובה על מגוון רחב של חלבונים ואורגניזמים.   
בשנים האחרונות, התפתחותם של מודליProtein Language Models מבוססי Transformersהובילה לפריצת דרך משמעותית ביכולת ללמוד ייצוגים עשירים של רצפי חלבונים, אך מודלים אלה נוטים להיות גדולים מאוד ודורשי משאבים חישוביים ניכרים, מה שמגביל את השימוש בהם בפועל.

מטרת הפרויקט הייתה לפתח את MinProtBert, גרסה קלה ויעילה של מודל ProteinBert, המאפשרת חיזוי חלבונים קושרי DNA תוך שמירה על ביצועים גבוהים, אך עם הפחתה משמעותית בזמן ריצה ובדרישות חישוביות. בנוסף, ביקשנו להעריך את יכולת ההכללה של המודל על משימות ביואינפורמטיות נוספות, כמו חיזוי מבנה שניוני וסיווג חלבונים ממברנליים, ולהשוות את ביצועיו למודלים קיימים בנושא זיהוי חלבוני קושרי דנ"א כגון מודל Deep-WET למגוון סטי-דאטה מגוונים של רצפי חלבונים.

בבסיס הפרויקט עומדת ארכיטקטורת Transformers ובפרט מודל BERT, שהותאם לרצפי חלבון. לצורך פיתוח MinProtBERT השתמשנו בגישת Knowledge Distillation, שבה מודל ProteinBert שימש כמודל "מורה" וממנו נגזר מודל "תלמיד" קומפקטי. תהליך האימון כלל שני שלבים עיקריים: Pre-training שאחריו קיבלנו מודל גנרי מוכלל לשפת חלבונים ו- Fine-tuning למשימות סיווג ייעודיות, בהן חיזוי חלבונים קושרי DNA, חיזוי מבנה שניוני של חלבונים, וסיווג חלבונים חוצי ממברנה.

MinProtBERT הציג ביצועים כמעט זהים ל־ ProteinBert בכל שלוש המשימות המרכזיות, תוך הפחתה משמעותית בזמני הריצה. במשימת חיזוי חלבונים קושרי DNA התקבלו תוצאות זהות כמעט לחלוטין עם חיסכון של כ־38% בזמן החישוב. במשימת חיזוי מבנה שניוני התקבלו ביצועים דומים בכל המדדים, אך MinProtBERT קיצר את זמן הריצה בכ־66%. בנוסף, במשימת סיווג חלבונים ממברנליים המודל הציג שיפור קל בכל המדדים, לצד קיצור של כ־16% בזמן הריצה.   
בבדיקות על אורגניזמים שונים, MinProtBERT שמר על ביצועים גבוהים גם על רצפים מאורגניזמים שונים, עם ירידה קלה בבקטריות וארכיאה. ניתוח נוסף שבחן השפעת סינון רצפים מסט-הדאטה הראה כי הסרת רצפים בעלי אחוז דימיון גבוה (SUB) שיפרה את הביצועים, בעוד שסינון אקראי (CUT) בגודל זהה כמעט לא השפיע, מה שמדגיש את יכולתו של המודל ללמוד ייצוגים כלליים יותר ולא להסתמך על כפילויות. לבסוף, בהשוואה למודל קיים מהשנתיים האחרונות בשם Deep-WET, בו MinProtBERT הציג ביצועים טובים יותר בכל המדדים (Accuracy, AUC, MCC).

הפרוייקט הראה ש- MinProtBERT מאפשר להשיג ביצועים גבוהים כמעט זהים לאלו של ProtBERT תוך הפחתה משמעותית בדרישות משאבים חישוביים ובזמני ריצה, מה שהופך אותו לפתרון יעיל ליישומים ביואינפורמטיים מגוונים. השימוש ב- Knowledge Distillationמאפשר לשמר ייצוגים איכותיים גם במודל קומפקטי, תוך יכולת הכללה טובה על מגוון רחב של חלבונים ואורגניזמים. המודל מהווה בסיס למחקר עתידי בזיהוי תכונות חלבונים, חקר קשר רצף-פונקציה, ופיתוח תרופות ממוקדות, ומדגים את הפוטנציאל של שילוב גישות NLP מתקדמות במחקר חלבונים.

**מבוא**

**רקע ביולוגי**

חלבונים קושרי DNA (DNA-Binding Proteins) מהווים קבוצה רחבה ומגוונת של חלבונים המעורבים כמעט בכל תהליך ביולוגי מרכזי בתא. הם מזהים רצפי DNA ספציפיים או מבנים מרחביים מסוימים ב־DNA, ונקשרים אליהם כדי לווסת, להפעיל או לשנות את פעילותם של גנים ושל מערכות תאיות שלמות. אחד התפקידים העיקריים של חלבונים אלו הוא בקרה על ביטוי גנים: חלבוני בקרה (כגון פקטורי שעתוק) נקשרים לאזורים פרומוטוריים או רגולטוריים בגנום, ומווסתים את קצב השעתוק בהתאם לצרכים הפיזיולוגיים של התא.   
מעבר לכך, חלבונים קושרי DNA ממלאים תפקידים מרכזיים גם בשכפול הגנום - כאשר הם פותחים, מייצבים ומארגנים את סליל ה־DNA במהלך סינתזת הגדילים החדשים. בנוסף, הם משתתפים בתהליכי תיקון DNA, שם הם מזהים נזקים גנטיים ומובילים לגיוס מנגנוני תיקון מתאימים, ובכך משמרים את יציבות הגנום. חלבונים מסוימים משמשים גם בתפקידי אריזה וארגון כרומטין, כדוגמת ההיסטונים - אשר יוצרים מבנים קומפקטיים המאפשרים אחסון יעיל של המידע הגנטי בגרעין התא.

מעבר לתפקידים הפיזיולוגיים, חלבונים קושרי DNA קשורים ישירות גם למצבים פתולוגיים. מוטציות המשפיעות על פעילותם או על דפוסי הקישור שלהם עלולות לגרום לשיבושים בבקרת ביטוי גנים, לפגיעה במנגנוני תיקון נזקי DNA ולהתפתחות מחלות שונות, כולל סרטן ומחלות גנטיות מורכבות.

למעשה, בגנומים של יונקים וחיידקים קיימים עדיין עשרות אלפי חלבונים שלא אופיינו מבחינה פונקציונלית, והיכולת לזהות אילו מהם קושרי DNA היא מפתח לקידום האנוטציה הגנומית ולחשיפת תפקידים ביולוגיים חדשים. מסיבה זו, הבנת המנגנונים המולקולריים של קישור DNA היא יעד מחקרי מרכזי, והיכולת לזהות חלבונים קושרי DNA באופן חישובי תורמת להאצת מחקרים בתחומי האנוטציה הגנומית, פיתוח תרופות ממוקדות וחקר מחלות. נוסף לכך, הבנה מדויקת של דפוסי הקישור מאפשרת גם תכנון מושכל של חלבונים מלאכותיים עם יכולות קישור מותאמות מטרה, תחום הצובר עניין הולך וגובר בביו-הנדסה סינתטית.

עם זאת, חיזוי חלבונים קושרי DNA על בסיס רצף החלבון בלבד מהווה אתגר משמעותי. למרות שרצף חומצות האמינו מכיל מידע סמנטי וביוכימי עשיר, הקשר בינו לבין יכולת הקישור ל־DNA אינו ישיר.   
תכונות מבניות ותפקודיות, כגון: קיפול חלבון, מטענים חשמליים, דינמיקת אינטראקציות ותכונות הידרופוביות, משפיעות במידה רבה על יכולת הקישור, אך אינן תמיד ניתנות להסקה באופן מיידי מרצף החלבון. במקרים רבים, גם שינויים מזעריים ברצף-כמו החלפת חומצת אמינו אחת - עלולים לשנות משמעותית את יכולת הקישור, מה שמדגיש את הצורך במודלים מתקדמים המסוגלים ללמוד דפוסים מורכבים ולא ליניאריים ישירות מהנתונים.

השיטות הניסיוניות הקיימות, כגון EMSA או ChIP-seq , מספקות מידע אמין על חלבונים קושרי DNA, אך הן יקרות, איטיות ותלויות בתנאים ספציפיים של דגימה. כאן נכנס לתמונה הפתרון החישובי: שימוש במודלי למידה עמוקה מאפשר לזהות דפוסים חבויים ברצפי חלבונים ולחזות יכולת קישור ל- DNAבקנה מידה רחב, בדיוק גבוה ובעלויות נמוכות משמעותית. עם זאת, האתגר נותר מורכב במיוחד עבור מודלים חישוביים, משום שנדרש מהם להתמודד עם כמות עצומה של דאטה, רמות שונות של דמיון בין רצפים, ושונות גבוהה בין אורגניזמים וכל זאת מבלי להסתמך על מידע מבני חיצוני.   
אתגר נוסף הוא המחסור בדאטה מתויג: רק חלק קטן מהחלבונים זוהה באופן אמין וחד משמעי כקושר או לא קושר DNA. עובדה זו מקשה על אימון ישיר של מודלים מונחי־תיוג, ומדגישה את יתרונם של מודלים המבוססים על למידה מראש מתוך מאגרי רצפים לא מתויגים - המאפשרים לרכוש ידע כללי על דפוסי החלבונים, ולאחר מכן ליישמו גם במשימות שבהן כמות הדאטה המתויג מוגבלת.

**רקע חישובי**

התקדמות דרמטית בתחום עיבוד השפה הטבעית (NLP) בשנים האחרונות הושגה בזכות פיתוחם של מודלי Transformers, שהוצגו לראשונה ב־2017 במאמר *“Attention is All You Need*” [1]. הייחודיות של מודלי Transformers נובעת ממנגנון ה-Self-Attention, המאפשר למודל לשקלל את הקשרים בין כל זוג עמדות ברצף. בכל שלב, כל טוקן (רכיב בקלט) ברצף "מתבונן" בכל שאר הטוקנים ומחשב כמה כל אחד מהם חשוב להבנת ההקשר שלו. המודל משקלל את התשובות לייצוג כולל. בכך מתאפשרת תפיסה יעילה של יחסי תלות ארוכי־טווח ברצף, אשר שיטות קלאסיות המבוססות על חלונות מקומיים לא הצליחו ללכוד בנוסף, Transformers מאפשרים עיבוד מקבילי של כל הרצף בבת אחת, בניגוד למודלים כמו RNN, תכונה שמקצרת את זמן האימון ומאפשרת שימוש בסטים גדולים מאוד של דאטה. ביישום על רצפי חלבון, כל חומצת אמינו ברצף מומרת לטוקן (רכיב נפרד בקלט) ומקבלת ייצוג חדש שמכיל מידע על שאר החומצות הרלוונטיות, גם אם הן מרוחקות מאוד זו מזו ברצף הלינארי. יתרון זה מאפשר למודל ללמוד דפוסים מבניים ותפקודיים מורכבים שמבוססים על אינטראקציות גלובליות, ולא רק על סמיכות מקומית.

על בסיס ארכיטקטורה זו פותח מודל BERT (Bidirectional Encoder Representations from Transformers) [2], אחד ממודלי השפה המשפיעים ביותר, אשר הציג קפיצת מדרגה ביכולת לנתח רצפים מורכבים. ייחודו של BERT הוא בכך שהוא דו-כיווני - הוא קורא את הרצף כולו במקביל ומבין את ההקשר של כל טוקן על פי מה שבא לפניו וגם על פי מה שבא אחריו וכך לומד קשרי תלות מורכבים בין עמדות שונות ברצף. מודלים מסוג BERT מאומנים בשני שלבים: בשלב ה- Pre-training המודל לומד באופן לא מונחה (unsupervised) על כמויות גדולות מאוד של דאטה. כך הוא רוכש הבנה סמנטית עמוקה של הדפוסים בתוך הרצף כולו. לאחר שלב זה המודל עובר שלב של Fine-tuning, שבו הוא מותאם למשימות ספציפיות באמצעות דאטה מתויג, משימת חיזוי סיווג.

בעקבות ההצלחה של מודל BERT בעיבוד שפה טבעית, נעשו התאמות של הארכיטקטורה גם לתחום הביולוגי, ובפרט לניתוח רצפי חלבון. אחת הדוגמאות הבולטות היא ProteinBert [3], מודל Transformer רחב היקף שפותח במיוחד לעיבוד מידע חלבוני. המודל שומר על הארכיטקטורה הדו-כיוונית של BERT, אך מותאם באופן ייעודי לרצפי חומצות אמינו. כך שהמודל מייצר במקביל גם ייצוג פרטני לכל חומצת אמינו וגם ייצוג גלובלי לרצף החלבון כולו, מה שמאפשר לו ללמוד קשרים מורכבים ברמות שונות של הארגון המבני והתפקודי גם ברמת החלבון וגם ברמת החומצה אמינית.

ProteinBert (נקרא גם ProtBert) אומן בשלב הקדם אימון (pre-training) על יותר מ־206 מיליון רצפים ממסד הנתונים UniRef100, יחד עם אנוטציות ביולוגיות מ-Gene Ontology במטרה ללמוד ייצוגים המשלבים גם מידע על הרצף וגם מידע פונקציונלי. המודל כולל 30 שכבות Transformer ומכיל כ-420 מיליון פרמטרים, מה שמאפשר לו לקלוט הקשרים גלובליים מורכבים ולבצע משימות מגוונות, כגון: חיזוי מבנים שניוניים, סיווג פונקציות חלבון, זיהוי דומיינים ותכונות ביופיזיקליות, והסקת תכונות מורכבות מתוך הרצף בלבד - גם במקרים של חלבונים ארוכים ומורכבים. בזכות ביצועיו הגבוהים ורב־השימושיות שלו, ProtBERT הפך בפועל ל־benchmark עבור מגוון רחב של משימות חלבוניות, והוא משמש נקודת ייחוס מרכזית למחקרים חדשים בתחום.

עם זאת, העלות החישובית של ProteinBert גבוהה: המודל דורש משאבי GPU משמעותיים, נפח זיכרון גדול וזמני ריצה ארוכים, ולכן שימוש בו עשוי להיות מאתגר בפרויקטים רחבי היקף או בסביבות עם משאבים מוגבלים.

בעקבות האתגרים החישוביים של מודלים גדולים כמו ProteinBert, שבהם מספר הפרמטרים ועומק השכבות גבוהים מאוד, פותחה גישה יעילה בשםKnowledge Distillation . הרעיון המרכזי בתהליך זה הוא שימוש במודל "מורה" גדול, בעל ביצועים גבוהים, כדי להדריך מודל "תלמיד" קטן יותר. המודל התלמיד מאומן לא רק לשחזר את התוויות הסופיות של המורה, אלא גם לחקות את התפלגות ההסתברויות של המורה ואת הייצוגים הפנימיים שנלמדו בשכבותיו. כך, המודל הקטן רוכש ידע עמוק ועשיר מבלי להזדקק לאותה כמות משאבים חישוביים.

גישה זו הודגמה לראשונה בצורה בולטת במודל DistilBERT [4] בתחום עיבוד השפה הטבעית, שהצליח לשמר כ־97% מביצועי BERT תוך שימוש בכ־40% פחות פרמטרים והרבה פחות משאבי GPU. ההצלחה של DistilBERT סללה את הדרך ליישום טכניקות דומות גם בתחום הביואינפורמטיקה, במטרה ליצור גרסאות קומפקטיות ויעילות יותר של מודלים חלבוניים מבוססי Transformers, תוך שמירה על רוב הביצועים של המודלים המקוריים יחד עם הפחתה משמעותית במשאבים החישוביים ובזמן הריצה.  
מעבר לחיסכון בפרמטרים, היתרון המעשי של Knowledge Distillation ("זיקוק ידע") מתבטא בכך שהוא מאפשר להריץ מודלים מתקדמים גם בסביבות מוגבלות בחומרה (כמו מחשבים אישיים או שרתים עם GPU יחיד), ובכך הופך אותם לכלים שימושיים באמת בקנה מידה רחב.

**שיטות עבודה**

**החלק של רועי על DeepWet:**

כחלק מתהליך ההערכה ביקשנו להשוות את MinProtBERT למודל עדכני, שפותח במיוחד עבור משימת חיזוי חלבונים קושרי DNA. לצורך כך חיפשנו עבודות חדשות מהשנים האחרונות המתמקדות במשימה זו, ומצאנו את מודל DeepWET, שפורסם בשנת 2024 [6].

במאמרם, המחברים הציגו מודל ייעודי לזיהוי חלבונים קושרי DNA. הרעיון המרכזי היה לקחת ייצוגים מספריים של רצפי חלבונים – שנבנו בעזרת שיטות מקובלות ובסיסיות בתחום של עיבוד שפה רדודה (כגון Word2Vec ו-fastText) – ולשפר אותם באמצעות אלגוריתם אופטימיזציה ולאחר מכן למקד את המידע החשוב ביותר. הייצוגים המעובדים הוזנו לרשת עצבית מסוג CNN, אשר למדה להבחין בין חלבונים הקושרים DNA לבין כאלה שאינם קושרים.

חשוב לציין כי המחברים סיפקו מאגר נתונים ייחודי, שכלל כ-2,000 חלבונים קושרי DNA ולא קושרי DNA עבור אימון המודלים, לצד סט מבחן עצמאי של כ-300 רצפים נוספים.

בהתבסס על כך, אימנו את MinProtBERT על אותם נתוני אימון שהוגדרו במאמר, ובחנו אותו על סט המבחן שסיפקו המחברים. ההרצה בוצעה באמצעות סביבת Hugging Face Transformers, תוך הוספת שכבת סיווג פשוטה לייצוג [CLS] של כל רצף – כפי שעשינו קודם. ביצועי המודל נבחנו לפי אותם מדדים ששימשו במאמר המקורי: דיוק (Accuracy), שטח מתחת לעקומה (AUC) ומקדם מתאם (MCC).

באופן זה, יצרנו בסיס השוואה שקוף וישיר מול DeepWET, הן מבחינת הדאטה והן מבחינת מדדי הביצועים, כדי לאפשר בחינה אובייקטיבית של היתרונות והחסרונות של MinProtBERT במשימת חיזוי קישור ל-DNA.

**תוצאות**

**תוצאות כלליות – השוואה בין ProtBert ו-MinProtBert**

בשלב הראשון של הניסויים בחנו את ביצועי המודל שפותח (MinProtBERT) לעומת מודל הבסיס ProtBERT בשלוש משימות מרכזיות:

1. חיזוי מבנה שניוני של חלבונים (Secondary Structure Prediction, Q3).
2. חיזוי חלבון ממברנלי (Membrane Protein Classification, Q2).
3. חיזוי קישור DNA לחלבון (DNA Binding Classification).

**מדדי ביצועים:**

* Accuracy: אחוז החיזויים הנכונים הכולל.
* Precision: דיוק החיזויים החיוביים (מתייחס ל-false positives).
* F1 score: מאזנת בין Precision ל-Recall, שימושי במיוחד במצבים של קלאסים לא מאוזנים.

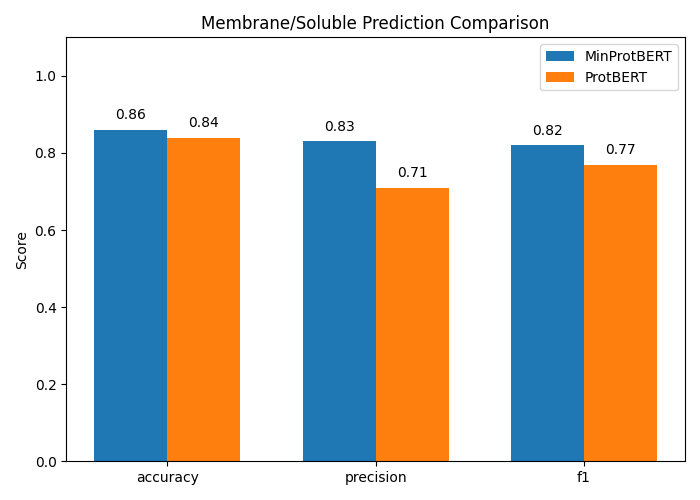
התוצאות עבור שלושת מבחני fine-tune השונים מובאות בטבלה א ובהרחבה בגרפים א-ד:

**טבלה א – השוואת ביצועי MinProtBERT מול ProtBERT**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Runtime (דקות)** | **F1** | **Accuracy** | **Precision** | **מודל** | **משימה** |
| 387 | 0.76 | 0.76 | 0.76 | ProtBERT | מבנה שניוני (Q3) |
| 132 | 0.75 | 0.76 | 0.75 | MinProtBERT |
| 697 | 0.77 | 0.84 | 0.71 | ProtBERT | ממברנלי (Q2) |
| 586 | 0.82 | 0.86 | 0.83 | MinProtBERT |
| 443 | 0.97 | 0.971 | 0.97 | ProtBERT | קישור DNA |
| 273 | 0.97 | 0.969 | 0.96 | MinProtBERT |

**גרף א – השוואה בין המודלים במשימת חיזוי מבנה שניוני**תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

תוכן בינה מלאכותית גנרטיבית עשוי להיות שגוי.

**גרף ב – השוואה בין המודלים במשימת חיזוי חלבון ממברנלי/מסיס**

**גרף ג – השוואה בין המודלים במשימת חיזוי קשירת DNA**

**תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

תוכן בינה מלאכותית גנרטיבית עשוי להיות שגוי.**

**גרף ד – השוואה בין המודלים בזמני ריצה**

**תמונה שמכילה טקסט, צילום מסך, תרשים, עלילה

תוכן בינה מלאכותית גנרטיבית עשוי להיות שגוי.**

**שמירה על ביצועים במקביל לחיסכון במשאבים:**

שמירת ביצועים:

כפי שניתן לראות בטבלה א - MinProtBERT מצליח לשמור על ביצועים כמעט זהים לאלו של ProtBERT בכל שלוש המשימות, על אף היותו מודל קטן ויעיל יותר חישובית. בפרט, הוא שומר על Accuracy בטווח של כ-0.02 מתוצאות המקור.

יתרון במשימת הממברנלי:

כפי שניתן לשים לב בגרף ב - מעניין לראות כי במשימת סיווג חלבונים ממברנליים MinProtBERT אף מתעלה על ProtBERT בכל המדדים (Accuracy, Precision, F1) ובפרט שיפור של 0.05 בF1. ייתכן שהפשטות היחסית של המודל אפשרה לו הכללה טובה יותר בדאטה זה, שלא דורש מודל מורכב ומספר רב של שכבות.

יעילות חישובית:

לפי גרף ד, זמן הריצה של MinProtBERT קצר משמעותית בכל המשימות – חיסכון של עשרות אחוזים לעומת ProtBERT עם שונות רבה. כך למשל יש חיסכון של 16% במשימת חיזוי חלבון ממברנלי ו66% ריצה במשימת חיזוי מבנה שניוני, כך שמשימה ארוכה יותר מביאה לשיפור זניח יותר. הדבר ממחיש את הערך המעשי של תהליך ה-distillation: מודל קטן יותר שמאפשר אימון והרצה מהירים בהרבה.

משימת קישור DNA:

לפי גרף ג, במשימה המרכזית – חיזוי חלבונים קושרי DNA – שני המודלים הגיעו לתוצאות כמעט זהות (F1 = 0.97), כלומר MinProtBERT מצליח לעמוד בסטנדרט הגבוה של ProtBERT גם כאן, ובכך מממש את מטרת הפרויקט. ובנוסף, עמידה במשימת זיהוי קשירת דנ"א בשימוש במודלי שפה.

**תוצאות השוואה – קשירת דנ"א**

**השוואה בין גדלי הסטים**

כדי לבחון האם ביצועי המודל נשענים על נוכחותם של רצפים דומים מאוד בדאטה-סט, ביצענו ניסוי סינון על פי אחוז דמיון בין רצפים. מטרת הניסוי הייתה לוודא כי המודל אינו מסתמך על זיהוי טריוויאלי של חלבונים כמעט זהים, וכי הוא מסוגל להכליל (generalize) גם במצבים בהם קיימת שונות גבוהה יותר בין הרצפים.

בוצעו שני סוגי סינון:

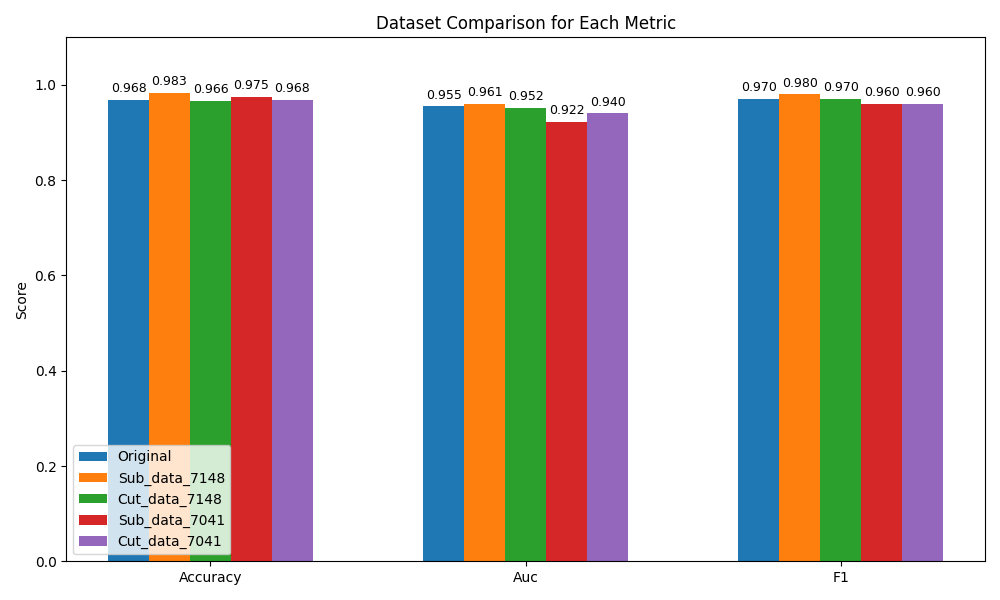
1. סינון חכם (sub) – איחוד רצפים בעלי דמיון גבוה (מעל 0.85 ומעל 0.75) כך שיישאר ייצוג אחד מכל קבוצה דומה.
2. סינון אקראי (cut) – בחירת תת-קבוצה אקראית בגודל זהה לתת-הקבוצה שנוצרה מהסינון החכם, לצורך השוואה.

תוצאות הניסוי מובאות בהרחבה בטבלה ב':

**טבלה ב – תוצאות MinProtBert במבחן קשירת DNA עבור סטים בגדלים שונים**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **סט נתונים** | **גודל** | **Accuracy** | **AUC** | **F1** |
| Full dataset | 12600 | 0.9682 | 0.9552 | 0.97 |
| Sub (SIM ≤ 0.15) | 7148 | **0.9832** | **0.961** | 0.98 |
| Cut (Random 7148) | 7148 | 0.966 | 0.9521 | 0.97 |
| Sub (SIM ≤ 0.25) | 7041 | 0.9754 | 0.9218 | 0.96 |
| Cut (Random 7041) | 7041 | 0.9679 | 0.9404 | 0.96 |

**גרף ה – השוואה סטים בגדלים שונים במבחן קשירת DNA**



כפי שניתן לראות בהדגשה בטבלה ב, הסינון החכם של רצפים דומים – SUB (במיוחד ברמת דמיון של 0.85) – לא הוביל לירידה בביצועי המודל, אלא אף לשיפור ב-Accuracy ו-F1 בהשוואה לדאטה-סט המלא. לעומת זאת, סינון אקראי (CUT) בגודל מקביל לא הניב שיפור משמעותי, והמדדים שמרו על רמתם כמעט זהה לדאטה המלא.

תוצאה זו מעידה כי בחלק מהדאטה-סט המלא, ביצועי המודל נבעו מאימון על רצפים דומים מדי, מה שגרם ל-overfitting לרצפים חוזרים. הסרת הדמיון הגבוה אפשרה למודל ללמוד ייצוגים כלליים יותר, ולשפר את ביצועי המדדים.

כדי להבין את התופעה טוב יותר, נבחן את טיב הדאטה עצמו. חישוב אחוז הדמיון המקסימלי בין רצפים עם תיוגים שונים הראה מספר זוגות עם דמיון גבוה מאוד, עד כ־97%, כמו למשל בין RPS27 ל-RPS27L. שני החלבונים הם חלק מתת-היחידה הקטנה של הריבוזום (40S) והם כמעט זהים ברצף החומצות שלהם, אך הם פאראלוגים – כלומר חלבונים שהתפתחו מנקודת מוצא משותפת עם שינויים פונקציונליים קלים. ההבדלים הקטנים ברצף מכתיבים הבדלים בביטוי ברקמות שונות או בתגובה למצבי עקה תאיים. כתוצאה מכך, RPS27L לעיתים קשור למנגנוני תיקון DNA ומסומן כקושר-DNA, בעוד ש-RPS27 אינו משתתף במנגנונים הללו, ולכן קיבל תיוג שונה בבסיס הנתונים. לכן, נוכחות של שני החלבונים האלה במהלך האימון לדוגמה, מקשה על המודל להבחין בהבדל הדק שמשפיע על התיוג שלהן.

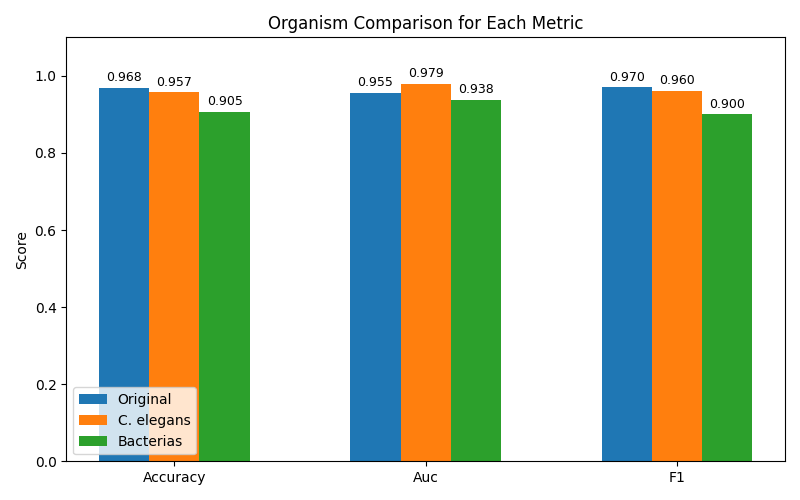
**תוצאות – השוואת ביצועים בין אורגניזמים שונים**

כחלק מבדיקת איכות הריצה בשאלת קשירת דנ"א, בחנו את ביצועי המודל על סט הנתונים המקורי של Homo sapiens לעומת אורגניזמים אחרים, תוך שימוש באותם מדדים: Accuracy, AUC ו-F1.   
טבלה ג' להלן מתארת את התוצאות:

**טבלה ג – תוצאות MinProtBert במבחן קשירת DNA עבור אורגניזמים שונים**

|  |  |  |  |
| --- | --- | --- | --- |
| **סט נתונים** | **Accuracy** | **AUC** | **F1** |
| Original data (Human) | 0.9682 | 0.9552 | 0.97 |
| C. elegnas | 0.957 | **0.979** | 0.96 |
| Bacteria & Archea | 0.905 | 0.938 | 0.9 |

**גרף ו – השוואה בין סטי אורגניזמים שונים במבחן קשירת DNA**



מהטבלה והגרף ניתן לראות מספר ממצאים מעניינים:

* עבור C. elegans, אין ירידה חדה ב-Accuracy או ב-F1 ביחס לסט האנושי, ואפילו נרשמה עלייה ב-AUC, מה שמעיד על יכולת כללית טובה של המודל ללמוד ייצוגים כלליים של רצפים גם באורגניזם זה.
* עבור בקטריות וארכאה, נצפתה ירידה בביצועים בכל המדדים. סביר כי ירידה זו נובעת מהעובדה שבשלב ה-pretraining על UniRef50, מודל ProtBERT ומורשו MinProtBERT נחשפו בעיקר לרצפים אוקריוטיים, ופחות לרצפים של בקטריות וארכאה, ולכן הייצוגים של רצפים אלו פחות מוכרים למודל.
* מסקנה עיקרית היא כי המודל מצליח לשמור על ביצועים טובים יחסית גם על אורגניזמים רחוקים מאדם, אם כי עבור אורגניזמים פחות ייצוגיים ב-pretraining ניכרת ירידה, כפי שמודגם בביצועים עבור בקטריות וארכאה.

**השוואת ביצועים למודלים נוספים בתחום זיהוי קשירת DNA**

כחלק מהערכת היכולת הכללית של MinProtBERT לזהות חלבונים קושרי DNA, השווינו את ביצועיו למודלים נוספים שפורסמו לאחרונה בתחום. מצאנו מאמר עדכני, ,Deep-WETשמתאר שיטה מבוססת למידה עמוקה לזיהוי חלבונים קושרי DNA על בסיס רצפי חומצות אמינו בלבד. Deep-WET משתמש בשילוב של שלושה סוגי קידוד רצפים (Global Vectors, Word2Vec ו-fastText), משקלל את התכונות באמצעות אלגוריתם Differential Evolution, ומסנן תכונות לא רלוונטיות לפני הזנתן לרשת קונבולוציונלית. בניגוד למודל שלנו, Deep-WET אומן על דאטה-סט קטן יחסית, הכולל כ־2,000 רצפים לאימון וכ־300 למבחן.

כדי לבצע השוואה ישירה, הרצנו את MinProtBERT על אותו דאטה-סט והערכת הביצועים ניתנה לפי Accuracy, AUC ו-MCC. להלן טבלה ד' המציגה את התוצאות:

**טבלה ד – תוצאות MinProtBert וDeepWet במבחן קשירת DNA**

|  |  |  |  |
| --- | --- | --- | --- |
| **מודל** | **Accuracy** | **AUC** | **MCC** |
| MinProtBert | 0.7939 | **0.8361** | 0.5885 |
| DeepWet | 0.7808 | 0.805 | 0.559 |

מטבלה ד, ניתן לראות כי MinProtBERT מציג ביצועים טובים יותר בכל המדדים בהשוואה ל-DeepWet בכל המדדים. בפרט, במדד AUC המודגש – אנו רואים כי יש שיפור מ0.805 ל0.8361, שיפור משמעותי במדד זה.

מסקנה עיקרית היא כי גם על סט קטן וסטנדרטי כזה, MinProtBERT מצליח לשמור על ביצועים גבוהים יותר מהמודלים האחרים, ומראה כי הידע שנצבר ב-pretraining וב-distillation מאפשר לו לבצע זיהוי מדויק ואמין של חלבונים קושרי DNA, גם כאשר הנתונים מוגבלים בגודלם. התוצאה מחזקת את היכולת של המודל להתמודד עם משימות זיהוי חלבונים קריטיות ולהציע שיפור אמיתי לעומת גישות קיימות.

**סיכום ודיון**

התוצאות שהוצגו מעל מדגימות את היכולת של MinProtBERT לשמר ביצועים גבוהים, כמעט ברמת ProtBERT, תוך הפחתת משאבים חישוביים בצורה משמעותית. ניתן להצביע על מספר נקודות עיקריות המשקפות את היכולות והיתרונות של מודלי שפה גדולים וקטנים בתחום הביואינפורמטיקה.

ראשית, השימוש בתהליך ה-distillation ליצירת MinProtBERT ממחיש כיצד ניתן ליצור מודל קומפקטי ויעיל חישובית מבלי לפגוע בביצועים. Distillation מאפשר העברת ידע ממודל גדול ומורכב (ProtBERT) למודל קטן יותר, כך שהמודל הקטן "לומד" את הדפוסים המרכזיים ללא הצורך בכל השכבות והפרמטרים הרבים של המודל המקורי. זה בא לידי ביטוי בזמני הריצה הקצרים יותר של MinProtBERT, לצד שמירה על ביצועים כמעט זהים, ואפילו שיפור במדדים במשימת חלבונים ממברנליים, שבה מודל פשוט יחסית מצליח להכליל טוב יותר.  
אמנם, יש לציין כי החיסכון יכול להגיע לעשרות אחוזים, אך עם שונות ניכרת בין משימות שונות. שונות זו יכולה להיות מוסברת, למשל, בעומס משתנה על השרתים בהם הופעלו המודלים: בעת ביצוע ניסויים במקביל למטלות אחרות או תחת עומס מערכת גבוה, זמני הריצה נוטים להיות ארוכים יותר. לכן, אף על פי שהשוואה כזו אינה מדויקת לחלוטין, היא ממחישה בבירור כי המודל הקומפקטי מציע יתרון מעשי משמעותי בזמן ריצה, מה שמדגיש את הערך של תהליך ה-distillation במיוחד בסביבות שבהן משאבים מוגבלים.

שנית, הממצאים ממחישים את יעילותם של מודלי שפה חלבוניים (Protein Language Models) במטלות שונות של זיהוי תכונות חלבונים. השימוש ב-pretraining על מאגרי רצפים רחבים כמו UniRef50 מאפשר למודל ללמוד ייצוגים עשירים של רצפים, המאפשרים הכללה על מטלות שונות, גם במצבים בהם הדאטה-סט קטן או חלקי. תכונה זו בולטת במיוחד בהשוואות בין אורגניזמים שונים: בעוד שסט הנתונים של Homo sapiens משמר ביצועים גבוהים, סטים מאורגניזמים רחוקים יותר (כגון בקטריות וארכאה) מצביעים על ירידה במדדים, כנראה בשל ייצוג נמוך של רצפים אלו ב-pretraining. עם זאת, אפילו עבור C. elegans לא נצפתה ירידה חדה ב-Accuracy או ב-F1, ואף נרשמה עלייה ב-AUC, מה שמעיד על יכולת הכללה מרשימה של המודל על רצפים קרובים ביולוגית.

בהקשר של זיהוי חלבונים קושרי DNA, המודל מראה יתרון ברור בהשוואה לגישות אחרות שפורסמו לאחרונה, כגון Deep-WET. MinProtBERT מצליח לשפר את כל המדדים – Accuracy, AUC ו-MCC – על סט קטן ומוגבל בגודלו, המראה כי הידע הנצבר ב-pretraining וב-distillation מאפשר זיהוי אמין גם עם דאטה מוגבל.

כמו כן, ממצא זה מסביר מדוע הסינון החכם (SUB) משפר ביצועים: הסרת רצפים כמעט זהים אך עם תיוגים סותרים הפחיתה רעש תיוג והקטינה את התלות בכפילויות, כך שהמודל נאלץ ללמוד ייצוגים כלליים יותר ולזהות דפוסים אמיתיים הקשורים לתיוג, במקום להסתמך על שוני נקודתי בין רצפים כמעט זהים. נקודה זו היא גם חולשה במודל – הוא אינו מצליח להבחין בהבדלים דקים שבין רצפים עם תיוג שונה.

סינון אקראי (CUT) לא פוגע בביצועים, מה שמעיד על חוסן המודל בפני צמצום אקראי של הדאטה – שמירה על גיוון ורמת דמיון כללית מספיקה כדי לשמר את היכולת ללמוד ייצוגים שימושיים.

לסיכום, התוצאות מדגישות את היתרון המשולב של מודלי שפה חלבוניים עם distillation: יכולת הכללה גבוהה, חיסכון חישובית משמעותי, והתמודדות טובה עם משימות מורכבות כמו זיהוי חלבונים קושרי DNA. יתרון זה חשוב במיוחד ליישומים ביואינפורמטיים שבהם משאבים מוגבלים, ומחזק את המגמה להשתמש במודלים קומפקטיים אך עוצמתיים בתחום חיזוי תכונות חלבונים.

**כיווני מחקר לעתיד**

למרות ההישגים המרשימים של MinProtBERT, מספר כיווני מחקר מעניינים עדיין פתוחים לחקירה והרחבה:

1. השוואה למודלים נוספים

ניתן להרחיב את ההשוואה למודלים נוספים שפותחו לאחרונה לזיהוי חלבונים קושרי DNA או חיזוי תכונות חלבונים אחרות. מודלים מבוססי Transformers שונים, רשתות קונבולוציונליות, או מודלים היברידיים עם קידוד תכונות ביולוגיות (כגון Deep-WET) מציעים נקודות השוואה נוספות. מחקר מעמיק יוכל לבחון:

* כיצד MinProtBERT מתמודד מול מודלים עם ארכיטקטורה שונה או מאגרי דאטה שונים.
* אילו מאפיינים ביולוגיים ומבניים המודל הקטן מצליח ללמוד טוב יותר ממודלים אחרים.
* היכן המודל לא מצליח להגיע לתוצאות די טובות, כלומר היכן החולשות של מודל מבוסס NLP.

1. חקר רצפים דומים מאוד עם תיוג שונה

הממצא של זוגות רצפים כמעט זהים אך בעלי תיוגים שונים (כגון RPS27 ו-RPS27L) מצביע על אתגר ביולוגי ומחקרי מעניין:

* ניתן לבדוק האם המודל מסוגל ללמוד הבחנה מבוססת על תת-רצף קטן, שבו ההבדלים הפונקציונליים מתבטאים.
* חקירה מעמיקה של תתי-האזורים השונים ברצף יכולה לחשוף אילו פיסות מידע ביולוגי קריטיות לזיהוי פונקציה שונה, ומה מייחד את תת-החלק הזה מבחינת מבנה או קישור.
* פיתוח שיטות לניתוח ויזואלי של תשומת הלב (attention maps) לפי טוקנים עשוי לסייע לזהות את תתי-האזורים בהם המודל מתמקד להבחנה בין פאראלוגים דומים.

1. חיזוי קשירת DNA ברמת בטוקן

כיוון נוסף הוא הרחבת חיזוי קשירת DNA לרמת החומצה האמינית (token-level):

* המטרה היא לזהות באילו חומצות אמינו ברצף נוצר הקישור ל-DNA, ולא רק להעריך האם חלבון מסוים קושר DNA.
* חקירה כזו תאפשר זיהוי דפוסי רצף ספציפיים, קווים משותפים בין רצפים שונים, ואולי גם הבנה של תפקיד ההבדלים הקטנים בין רצפים כמעט זהים.
* שילוב של מודלי שפה פר-טוקן עם ידע מבוסס מבנה חלבוני עשוי להעצים את הדיוק ולספק יכולת ניתוח ביולוגי עשיר יותר, והבנה יותר טובה של אותן תתי-רצף שמאפשרים הבדל פונקציונאלי.

הכיוונים שהוצגו מאפשרים לא רק שיפור ביצועים, אלא גם הרחבת ההבנה הביולוגית של רצפים דומים עם תיוגים שונים, חקר קשר רצף-פונקציה ברמת הטוקן, והרחבת שימושי המודל למשימות שונות בתחום הביואינפורמטיקה. גישה זו מחזקת את הפוטנציאל של מודלים קומפקטיים אך רבי-עוצמה בשימור דיוק תוך חיסכון חישובית משמעותי.

**ביבליוגרפיה**

**מאמרים:**

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.

[3] Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. Bioinformatics.

[4] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[5] Geffen, Y., Ofran, Y., & Unger, R. (2022). DistilProtBert: Distillation of Protein Language Models. bioRxiv.

[6] Mahmud, S. M. H., Goh, K. O. M., Hosen, M. F., Nandi, D., & Shoombuatong, W. (2024). Deep-WET: Deep learning for protein–DNA binding prediction. Briefings in Bioinformatics.

**מאגרי דאטה:**

[1] UniRef50 (Uniprot Consortium, 2019) - מאגר רצפי חלבונים מקובצים לפי 50% דמיון רצפים, שימש עבור Pretraining של מודל השפה החלבוני.

[2] UniProtKB/Swiss-Prot (Uniprot Consortium, 2024) - רצפי חלבונים ממוינים עם אנוטציה איכותית, שימשו למשימות זיהוי חלבונים קושרי DNA.

[3] DeepLoc (Almagro Armenteros et al., 2017) - מאגר עבור סיווג חלבונים ממברנליים ומומסיים.

[4] CASP12, CB513, TS115 (מקורות שונים) - סטי Benchmark לחיזוי מבנה שניוני של חלבונים (משימת Q3).