# תרגיל 2 – אלגוריתמים בביולוגיה חישובית

תיאור המודל:

בחרנו לממש מודל מרקוב חבוי (HMM – Hidden Markov Model) למטרת התרגיל – זיהוי אזורי CpG ב-DNA, בדומה למודל שלמדנו בכיתה.

בחרנו במודל זה מכיוון שהמודל תואם למבנה הנתונים שלנו – במודל HMM המצבים החבויים מייצגים מאפיינים מבניים בגנום שאינם נצפים ישירות. קיומם של איי CpG הוא תכונה סמויה, אך ניתן להסיק עליהם דרך רצפי הDNA. לכן, מודל מרקוב חבוי הוא כלי נוח ומתאים לזיהוי איי CpG – בזכות היכולת שלו לדמות מצבים חבויים, לשמר הסתברויות מעבר ופליטה, ולהתמודד עם רעש ונתונים לא שלמים.

המודל מבצע אימון על בסיס זוגות של נתוני DNA והתיוגים שלהם, באמצעות מודל ה - HMM שבו נלמדים הסתברויות המעבר והסתברויות הפליטה. לאחר מכן מתויג קובץ פלט חדש באמצעות הפרמטרים שוערכו בשלב האימון – בעזרת אלגוריתם ויטרבי שמטרתו למצוא את רצף המצבים (C/N) הסביר ביותר עבור הנתונים החדשים. אלגוריתם ויטרבי מיישם חישוב דינאמי (Dynamic Programming) כדי למקסם את ההסתברות לרצף המצבים שנבחר עבור התצפית (רצף ה-DNA).

המודל מורכב מהרכיבים הבאים:

1. שני מצבים חבויים (States):

* C – מצב המייצג אזור אי CpG.
* N – מצב המייצג אזור שאינו אי CpG.

נניח שכל מצב מקושר להסתברות פליטה של בסיסי ה-.DNA

1. תצפיות (Observations):

* התצפיות הן רצפים של האותיות {A,T,G,C,N} כך שארבע האותיות הראשונות מייצגות את בסיסי הדנ"א ואות N מייצגת פער בתצפית.

1. הסתברויות התחלה (Start Probabilities):

* הסתברות התחלתית עבור כל מצב:

1. הסתברויות מעבר (Transition Probabilities):

* הסתברויות המעבר בין המצבים: , עבור כל צמד מצבים.

1. הסתברויות פליטה (Emission Probabilities):

* הסתברויות הפליטה עבור כל מצב: , כאשר הוא הבסיס ה-t ברצף הדנ"א.

**הגדרת הפרמטרים במודל:**

פרמטרים:

* – וקטור הסתברויות התחלה.
* A – מטריצת הסתברויות מעבר.
* B – מטריצת הסתברויות פליטה.

חישוב הפרמטרים:

* וקטור הסתברויות התחלה – :

נחשב את שכיחות המצבים הראשוניים ברצף המוערכים (מספר הפעמים שכל מצב מופיע כמצב ראשון חלקי מספר הרצפים הכולל).

* מטריצת הסתברויות מעבר – A:

נחשב את היחס בין המעברים בין כל מצב למצב , לחלק למספר הכולל של מעברים מהמצב .

לדוגמה:

* מטריצת הסתברויות פליטה – B:

נחשב את היחס בין שכיחות כל תצפית במצב S, לחלק למספר הכולל של תצפיות במצב S.

לדוגמה:

**לימוד הפרמטרים:**

1. איסוף הנתונים:

* נשתמש בקובץ ה-fasta שניתן לנו במסגרת התרגיל ובקובץ התיוגים המתאים.
* על מנת להגדיל את סט האימון שלנו הפכנו את הרצפים המופיעים בקובץ ה-fasta והתאמנו לכך את התיוגים המופיעים בקובץ התיוגים וכך קיבלנו סט אימון בגודל כפול.

1. שיערוך פרמטרים:

* כאמור, נעשה שימוש באומדן נראות מירבית (MLE) באופן הבא:

**הנחות המודל:**

1. תכונת המרקוביות:

במודל אנחנו מניחים כי הסתברות של מצב תלויה רק במצב הקודם, כלומר:

1. תכונת עצמאות הפליטות:

ההנחה היא שהפליטה תלויה רק במצב הנוכחי ולא במצבים או תצפיות אחרים, כלומר:

1. ההנחה כי הנתונים המתויגים מספיק גדולים לייצג את ההתפלגות האמיתית של המעברים והפליטות.

**אופטימליות של הפרמטרים**

מטריצות ההסתברויות הנלמדות לפי אומדן הנראות המירבית הם האופטימליים בהנחה שאכן המדגם שלנו מייצג את ההתפלגויות בעולם האמיתי.

יתרון של השיטה שממומשת בתרגיל היא בפשטות החישובית, אך היא תלויה בגודל ואיכות הנתונים. ולכן, במדגם קטן או במדגם מוטה נצפה שיהיו שגיאות באומדנים.

**תהליך בחירת התיוגים בהינתן המודל:**

כדי לבחור את התיוגים עבור התצפיות, נשתמש באלגוריתם ויטרבי (Viterbi Algorithm), שהוא אלגוריתם דינאמי למציאת רצף המצבים הסביר ביותר במודל מרקוב חבוי.

שימוש באלגוריתם ויטרבי:

1. אתחול:

* חישוב ההסתברות להתחיל במצב ולהפיק את :

ונבחר את המצב שמקבל את ההסתברות הגבוה יותר בחישוב זה.

1. שלב החישוב:

עבור כל תצפית ולכל מצב s, נחשב את ההסתברות המצטברת המקסימלית שהמצב s ב-t מגיע מרצף מצבים אפשרי קודם:

ז"א, הוא המסלול המקסימלי שהוביל למצבs במיקום t ברצף.

1. מעקב לאחור – backtracking:

לאחר סיום חישוב כל ההסתברויות, מוצאים את המצב הסביר ביותר במיקום האחרון (T) ברצף:

וחוזרים אחורה לאורך המסלול שחושב על מנת למצוא את .

אלגוריתם ויטרבי מבטיח את מציאת הרצף הסביר ביותר של מצבים במודל HMM ועובד בצורה דינאמית כך שהוא יעיל גם עבור רצפים ארוכים.

A diagram of a network

Description automatically generatedאילוסטרציה למודל **HMM**

בגרף מתואר באופן ויזואלי המבנה של מודל מרקוב חבוי עבור משימת זיהוי איי CpG .

בתרשים ניתן לראות את שני המצבים החבויים בעיגולים הכחולים ו- שמייצג את הבסיס שחזינו בו .

בהמשך למה שמתואר במסמך בפירוט, מודל המרקוב החבוי משערך את ההסתברויות של כלל המעברים ובעזרת אלגוריתם ויטרבי מוצא את המסלול בעל ההסתברות המצטברת הגבוהה ביותר.

הערכת ביצועי המודל:

ביצענו אימון על בסיס הנתונים שקיבלנו בתיקיית data, על מנת להכפיל את גודל סט האימון ביצענו reverse-complement לרצפים בקובץ שסופק, והערכנו את ביצוע המודל על ידי שימוש במידע שסופק על כרומוזום 2.

על מנת להריץ את הטסט יש לכתוב את המילה "evaluate" במקום נתיב output בשורת הפקודה ולהכניס את קובץ האינפוט CpG-islands.2K.chr2.lbl.fa.gz כך שהמודל יציג את ניתוח ביצועיו על נתוני הטסט שקיבלנו.

כדי לתייג את הנתונים נריץ בדרך הרגילה את קבצי הטסט.

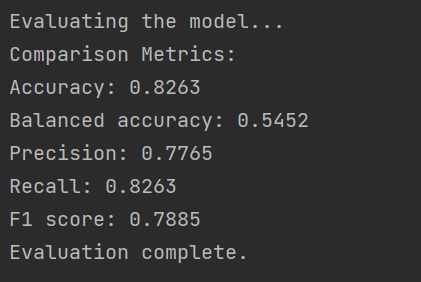
להלן התוצאות שהתקבלו בהרצה:

A black background with white text

Description automatically generated

A black background with white numbers

Description automatically generated

Evaluation Metrics

ניתן לראות שהתוצאות שמתקבלות על ידי שימוש במודל הן תוצאות טובות סה"כ למשימה מורכבת כמו זיהוי איי CpG ובעבור מודל יחסית פשוט עם הנחות משמעותיות:

**Accuracy = 82.6%**

משמעות – היחס הכולל של ניבויים נכונים מתוך כל הניבויים שבוצעו.

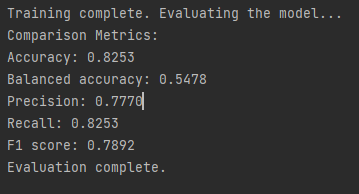
פירוש – דיוק של כ-83% מצביע על כך שהמודל מסווג נכון אזורי CpG ואזורי non-CpG ב-83% מהמקרים, שזו תוצאה יחסית טובה.

**Balanced Accuracy = 0.5452**

המודל מצליח לזהות היטב תיוג אחד אך פחות טוב תיוג אחר למשל את C שמופיע פחות פעמים.

ניסינו לשפר את המודל באמצעות איזון הנתונים כך שיכלול משקלים מותאמים לשיפור זיהוי התיוג הנדיר יותר על ידי הגדלת המשקל של כיתה "C" בעת חישוב הסתברויות המעבר והפליטה. כמו כן ניסינו להוסיף smoothing.

ניתן לראות כי השינוי שיפר את הבלנס בצורה זניחה והזיק מעט לדיוק כך שבסופו של דבר בחרנו להישאר עם המודל המקורי. אלו התוצאות של המודל עם תוספת זאת:



**Precision = 77.6%**

משמעות – המדד בוחן את היחס בין הניבויים החיוביים שנמצאו נכונים לבין סך כל הניבויים החיוביים.

פירוש – כאשר המודל מנבא אזור כ- CpG הוא צודק ב-77.6% מהמקרים. זה מצביע על כך שקיים שיעור מסוים של false positives.

**Recall = 82.6%**

משמעות – המדד בוחן את היחס בין הניבויים החיוביים הנכונים לבין סך כל המקרים החיוביים בפועל.

פירוש – רגישות של כ-82.6% מציינת שהמודל מצליח לזהות כ-82% מכלל אזורי ה-CpG בפועל. זה מצביע על יכולת טובה של המודל לזהות מקרים חיוביים.

**F1 Score = 78.8%**

משמעות - מדד F1 הוא הממוצע ההרמוני בין הדיוק החיובי לרגישות, ומאזן בין שני המדדים:

פירוש – מדד F1 של כ-78.8% מצביע על איזון טוב בין הדיוק החיובי לרגישות. המדד חשוב כאשר יש צורך לאזן בין שיעור החיוביים הכוזבים לשיעור השליליים הכוזבים.