יצרנו את הפונקציות הבאות:

* פונקציה בשם getAAfreq שמחשבת תדירות הופעה של חומצה אמינית נתונה.
* פונקציה בשם getCodonFreq שמחשבת תדירות הופעה של קודון נתון.
* פונקציה בשם ORFlength שמחזירה את אורכי הORF.

כדי לבדוק את נכונות הפונקציות נשווה בין:

* תדיריות החומצה האמינית ארגנין (R) שחושבו על ידי הפונקציה לתדירויות הנתונות.
* תדיריות הקודון AGG שחושבו על ידי הפונקציה לתדיריות הנתונות.
* אורכי הORF שחושבו על ידי הפונקציה לאורכים הנתונים.

ההשוואה בוצעה באמצעות סקריפט בשם main\_A (מצורף כקובץ .m ובנספח א').

**פיצ'רים**

סקריפט ההרצה של הרגרסור מצורף בקובץ בשם train\_script.

בנוסף לפיצ'רים הנתונים, בחרנו להוסיף פיצ'רים נוספים:

* **(tRNA Adaptation Index) tAI** – מדד המשערך את האדפטציה של קודון למאגר הtRNA שבתא. קצב התרגום של הORF פרופורציונלי לרמות מולקולות הtRNA שמזהות אותו. פיצ'ר זה חושב באמצעות התוכנה stAIcalc. התוכנה מוציאה את ערכי הtAI בקובץ .txt: known\_tAI לKnown\_set\_Bacillus וunknown\_tAI לUnknown\_set\_Bacillus.
* **mRNAfold** – קיפול הmRNA משפיע על יעילות ביטוי הגן. עבור אנרגיית הקיפול הנתונה ב100 החלונות שמכילים 30 נוקלאוטידים, חישבנו ממוצע שיהווה מדד לקיפול הmRNA כולו באמצעות הפונקציה mRNAfoldAVG.
* **GCcontent** – תדירות הופעת נוקלאוטידים C או G. מחושב על פי הנוסחה: . בין נוקלאוטידים C וG יש 3 קשרי מימן, לעומת A וT בניהם יש 2 קשרי מימן. לכן ORF שבו GCcontent יחסית גבוה מעיד על ORF יותר יציב, מה שמשפיע על ביטוי גנים. פיצ'ר זה חושב באמצעות הפונקציה GCcontent.
* **AAfreq** – 20 פיצ'רים, כל אחד מהם מתאר תדירות הופעת חומצה אמינית שונה. חומצות אמינו מהוות את אבני הבניין של החלבונים, תדירות גבוהה של חומצה אמינית מסוימת יכולה להעיד על חשיבות האזור ברצף. החישוב של פיצ'ר זה בוצע בtrain\_script.
* **promotor freq** – 3 פיצ'רים המתארים את תדירות הופעת 3 פרומוטורים: CTAAA, CCGATAT, TATAAT. אלה פרומוטורים מוכרים בחיידקים ובייחוד בBacillus subtilis. הפרומוטורים מסמנים את תחילת האיזור הפונקציונלי ברצף. הפרומוטור מאפשר את התחלת תהליך השעתוק. ככל שהוא מופיע יותר, ניתן להניח שהORF משועתק יותר, ורמות החלבון יהיו גבוהות יותר. החישוב של פיצ'ר זה בוצע באמצעות הפונקציה promoter\_freq.

**בחירת פיצ'רים**

סקריפט ההרצה של הרגרסור מצורף בקובץ בשם train\_script, ובנספח ב'.

לאחר יצירת הפיצ'רים, היה עלינו לבחור מתוכם את הפיצ'רים שנתנו את התוצאה הטובה ביותר (קורולציה גבוהה בין רמות החלבון החזויות לרמות החלבון הנתונות). תחילה, בדקנו את הקורולציה בין רמות החלבון עבור כל אחד מהפיצ'רים בנפרד ובחרנו את הפיצ'ר בעל הקורולציה הגבוהה ביותר. לאחר בחירתו הוספנו בכל פעם פיצ'ר אחד ובדקנו האם הוא מעלה את ערך הקורולציה. הפיצ'רים שנבחרו הם:

* tAI
* תדירות הופעת החומצות האמיניות: Q, H, I, L, M, S, W, Y, N.
* תדירות הופעת הקודון AGG
* ממוצע קיפול הmRNA
* תדירות הופעת הפרומוטור CCGATAT

הdata חולק באופן רנדומלי לסט train (המכיל 90% מהגנים) וסט validation (המכיל 10% מהגנים).

האימון של הרגרסור בוצע על סט הtrain, וכדי לבדוק את טיב חיזוי רמות החלבון בדקנו קורולציה בין רמות החלבון החזויות לסט הvalidation ורמות החלבון הנתונות לסט הvalidation. כדי לקבל רגרסור אופטימלי, הרצנו אותו בתוך לולאה עם 100 איטרציות ושמרנו את התוצאה עם הקורולציה הטובה ביותר.

הרגרסור שנבחר:

הקורולציה שהתקבלה: עם

**רמות חלבון חזויות לunknown set**

סקריפט לחלק מצורף בקובץ בשם unknown\_set\_prediction

תחילה חישבנו את הפיצ'רים הנבחרים עבור הUnknown set באופן זהה לחישוב עבור הKnown set.

כדי למצוא את רמות החלבון הצפויות, נכפול את X (מטריצת הפיצ'רים) בוקטור b (הרגרסור שנבחר).

לאחר מכן סידרנו את הגנים לפי רמות החלבון החזויות להם כאשר 1 – רמת החלבון הנמוכה ביותר ו700 – רמת החלבון הגבוהה ביותר.

דירוג הגנים מצורף בקובץ GeneRank.m.