

תוכנה 1 – אביב תשע"ה

תרגיל מספר 5

קלט-פלט, עיבוד קבצים מתקדם וחישוב מדעי (IO, Parsing and Scientific Computing)

הנחיות כלליות:

- קראו בעיון את קובץ נהלי הגשת התרגילים אשר נמצא באתר הקורס.
- הגשת התרגיל תיעשה במערכת ה-moodle בלבד (<http://moodle.tau.ac.il/>).
- יש להגיש קובץ zip יחיד הנושא את שם המשתמש ומספר התרגיל (לדוגמא, עבור המשתמש aviv יקרא הקובץ aviv_hw5.zip). קובץ ה-zip יכיל:
 1. קובץ פרטים אישיים בשם details.txt המכיל את שמכם ומספר ת.ז.
 2. קבצי ה-java של התוכניות אותם התבקשתם לממש.
- יש לממש את המתודות הנדרשות על פי החתימות וערכי ההחזרה המצוינים בתרגיל.
- מותר ואף רצוי להגדיר מתודות עזר.
- קבצי הפלט המופקים ע"י התוכנית צריכים להיות זהים לקבצי הפלט לדוגמא המצורפים לתרגיל.

Bioinformatics: Gene Expression Analysis

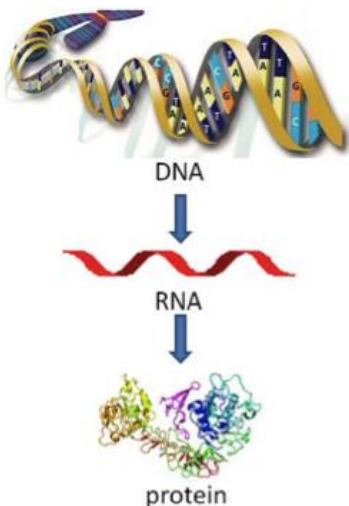
בתרגיל זה נכתוב תוכנית המנתחת מידע ביולוגי **אמיתי** וננסה לזהות גנים שרמת הביטוי שלהם מבדילה בין שני תתי-סוגים של סרטן השד. זיהוי גנים כאלו הוא חשוב ביותר, כי הגילוי יאפשר לנו ללמוד על המנגנונים הביולוגיים העומדים מאחורי תתי-הסוגים השונים, ואולי גם לאבחן ולטפל טוב יותר בחולות על בסיס רמת הביטוי של גנים אלו.

במהלך התרגיל נתאמן על קריאה וכתובה של קבצים בפורמט מוגדר, עיבוד מתקדם של טקסט, עבודה מתקדמת עם מערכים, וגם נלמד קצת בעקיפין על מהפיכת הרפואה המותאמת אישית שמוביל תחום הביואינפורמטיקה.

נתחיל עם רקע ביולוגי מקוצר ומקורב:

בגרעין של רוב התאים בגוף שלנו, שמור אותו רצף דנ"א המכיל כ-20,000 מקטעים הקרויים **גנים**. רוב הגנים הידועים מקודדים הוראות לייצור חלבון (החלבונים משמשים את התאים בגופינו לצרכים מבניים ותפקודיים).

כשתא צריך לייצר חלבון מסוים, הוא קודם ממיר את מקטע הדנ"א של הגן הרלבנטי למולקולת רנ"א-שליח (**mRNA**), וזו יוצאת מהגרעין אל חלל התא, שם היא נקראת ע"י מנגנוני בנית החלבונים של התא שיוצרים על פיה חלבון מתאים. ככל שיש יותר עותקים של mRNA כך יוצרו יותר מולקולות של החלבון הרלבנטי, ונוכל לומר ש-"הביטוי של הגן" הוא גבוה יותר (כלומר נוצרות ממנו יותר מולקולות mRNA ואח"כ גם יותר מולקולות חלבון).



תאים שונים יבטאו גנים ברמות שונות כתלות ב-

- **סוג התא** - תא שריר יבטא גנים אחרים (או ברמות ביטוי שונות) מתא כבד כי כל סוג תא זקוק לחלבונים שונים לצורך תפקודו.
- **מצב התא** - גם בתאים מאותו סוג יש שינויים ברמות ביטוי הגנים לאורך זמן - למשל: לאחר ארוחה יש צורך בהרבה מולקולות אינסולין ולכן תאי הבלבל יתחילו לבטא את הגן המקודד לחלבון זה.
- **מחלה** - במחלות שונות יש פגיעה ברמות הביטוי של גנים בהשוואה לתא נורמלי. בהרבה סוגים של סרטן משתנות רמות הביטוי של גנים רבים, מה שמאפשר לתאים לאמץ תכונות חדשות וביניהן כאלו שגורמות לחלוקה מואצת שלהם (Proliferation).



שבבי דנ"א. כל שבב מודד את רמות הביטוי של אלפי גנים בדגימה ביולוגית אחת. ערכי הביטוי המופקים מכל שבב יתורגמו לעמודה אחת במטריצת הביטוי.

טכנולוגיות מהפכניות חדשות (כגון **שבבי דנ"א** או RNA-Seq) מאפשרות לנו למדוד את רמת מולקולות ה-mRNA בדגימה ביולוגית מסוימת, וכך הן מאפשרות לנו לדעת מה רמת הביטוי הספציפית של כל גן באותה הדגימה (באיזה עוצמה כל גן מופעל בדגימה מסוימת).

ניתוח רמות ביטוי גנים (Gene Expression Analysis), אחד מענפי הביואינפורמטיקה), לרוב יתחיל עם קובץ ענק המכיל מטריצת ביטוי גנים. במטריצה זו כל **שורה מייצגת גן (Gene)**, וכל עמודה מייצגת דגימה (Sample). כלומר - כל תא במטריצה מייצג את רמת הביטוי של גן מסוים בדגימה מסוימת.

ניתן להציג את מטריצת הביטוי כ-Heatmap כפי שמופיע בתרשים משמאל.

19 דגימות (Samples)

בתרגיל זה, ננתח אוסף נתונים המכיל 1259 גלאי גנים (שורות המטריצה) ו-19 דגימות (עמודות המטריצה) שנלקחו משני תת-סוגים של גידולים של סרטן השד. הדגימות מתחלקות לשתי קבוצות: ER+ ו-ER-*. כלומר, לכל דגימה יש תווית שממפה אותה לאחת משתי המחלקות.

נחפש גנים שנבדלים בביטוי שלהם בין 2 המחלקות (Differentially expressed genes). כלומר, עבור כל גן, נבדוק האם ממוצע הביטוי שלו על הדגימות מסוג ER+ שונה בצורה משמעותית מממוצע הביטוי שלו על דגימות מסוג ER-. לשם כך נשתמש במבחן סטטיסטי בשם two sample t-test.

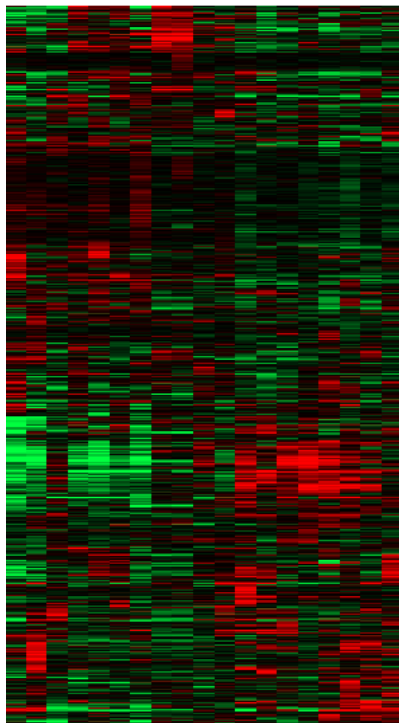
למציאת גנים כאלו יש משמעות ביולוגית ורפואית - הם מאפשרים לנו ללמוד על המנגנונים הביולוגיים שמאפיינים כל תת סוג, לאבחן בצורה מדויקת יותר גידולים חדשים על בסיס גנים אלו, ובהמשך להציע טיפול מותאם אישית לחולה לפי האפיון הגנטי שלה.

למעוניינים לקרוא דוגמא למאמר שפורסם על סמך אוסף נתונים זה:

<http://www.ncbi.nlm.nih.gov/pubmed/22025563>

* מה משמעות ER+? בהיעדר דרך טובה יותר, כיום נהוג לאפיין גידולים בשד לפי קיומו של קולטן האסטרוגן (ER) על דופן התאים הסרטניים, דבר שניתן לבדוק בבדיקת מעבדה פשוטה יחסית. נמצא שיש הבדל גדול בפוטנציאל הגרורתי ובתגובה לטיפול בין גידולים מסוג ER+ לבין גידולים מסוג ER-, אבל פחות ידוע על ההבדל הביולוגי בין שני תתי הסוגים. התקווה היא שעם מציאת גנים מפרדים בין שני תתי הסוגים נוכל ללמוד עוד על דרך היווצרות תתי-הסוגים ועל אפשרויות לטפל בהם בצורה טובה יותר.

1259 גלאי גנים (Gene probes)



Expression level:



High

Low

מה עליכם לעשות ?

בין קבצי התרגיל נתון לכם שלד של המחלקה GeneExpressionAnalyzer.

עליכם להשלים את מימוש המתודות הבאות בתוך מחלקה זו על פי ההוראות המופיעות בהמשך:

- parseGeneExpressionFile
- applyVarianceFilterOnDataset
- getDataEntriesForLabel
- writeTopDifferentiallyExpressedGenesToFile

בתרגיל זה עליכם להגיש את המחלקה GeneExpressionAnalyzer לאחר שתשלימו אותה, ואת קובץ הפלט GDS5088-DiffGenes.txt שתיצרו בסעיף 8.

תיאור אופן הרצת התוכנית

לאחר שתשלימו את מימוש המחלקה, אם נריץ אותה עם הארגומנטים הבאים:

```
GDS4085 "genotype/variation" "estrogen receptor-negative" "estrogen
receptor-positive" 0.01
```

כך ש-

- הארגומנט הראשון מציין את שם אוסף הנתונים אותו יש לנתח (בהנחה שקובץ בשם `GDS4085_soft` נמצא בתיקיה הנוכחית).
- הארגומנט השני מציין את סוג התוויות שלפיהן נחלק את הדגימות למחלקות (למשל – מין, קבוצת גיל, תת סוג של מחלה).
- הארגומנט השני והשלישי מציינים שני ערכי תוויות השייכים לסוג התוויות שציון בארגומנט השני. נרצה למצוא גנים שנבדלים ברמות הביטוי שלהם בין דגימות בעלות ערך התוויות הראשון לבין דגימות בעלות ערך התוויות השני (למשל – זכר\נקבה עבור סוג תוויות של מין).
- הארגומנט הרביעי מציין סף מובהקות סטטיסטית בין 0 ל-1.

המחלקה תקרא ותעבד את תוכן קובץ הקלט לפי פורמט SOFT שיוגדר בהמשך, תסנן החוצה גנים לא אינפורמטיביים בעלי שונות נמוכה, ותכתוב לקובץ פלט את רשימת הגנים שמתבטאים באופן שונה על שתי מחלקות הדגימות.

פלט התוכנית:

כתגובה לארגומנטים הנ"ל, תיצור התוכנית את קובץ הפלט `GDS4085-DiffGenes.txt` בספרייה הנוכחית (קובץ זה נתון לכם כדוגמא), ותדפיס למסך את הפלט הבא:

```
Gene expression dataset loaded from file GDS4085.soft.
Dataset contains 19 samples and 1259 gene probes.
Dataset contains 2 label subsets of type [genotype/variation]: [estrogen receptor-negative,
estrogen receptor-positive]
```

```
Applying variance filter on dataset: Kept 629 top variable genes having variance of at least 0.13.
```

```
68 differentially expressed genes identified using alpha of 0.01000 when comparing the two sample
groups [estrogen receptor-negative] and [estrogen receptor-positive].
Results saved to file GDS4085-DiffGenes.txt.
```

המחלקה GeneExpressionDataset

המחלקה GeneExpressionDataset מייצגת אוסף מידע, והיא מאחסנת מטריצת ביטוי גנים ונתונים נלווים.

```
public class GeneExpressionDataset {

    public String datasetTitle = "NA";

    public int samplesNumber; // מספר הדגימות
    public int genesNumber; // מספר גלאי הגנים

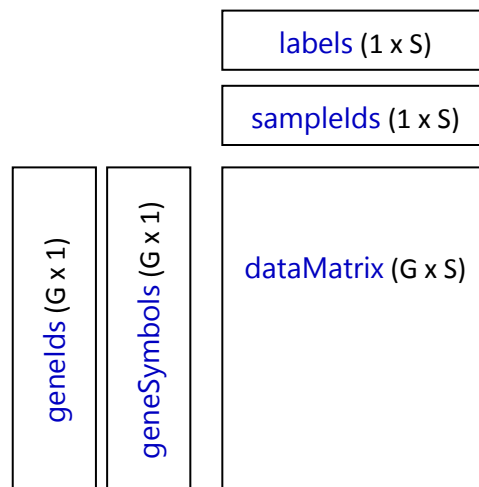
    public String[] sampleIds; // מזהה ייחודי עבור כל דגימה
    public String[] geneIds; // מזהה ייחודי עבור כל גלאי גנים
    public String[] geneSymbols; // שם הגן הנמדד ע"י כל גלאי גנים
    public float[][] dataMatrix; // מטריצת נתוני הביטוי של כל הגלאים (שורות) על כל
    // הדגימות (עמודות)

    public int labelSubsetNumber; // מציין את מספר מחלקות הדגימות באוסף הנתונים
    public String[] labels; // תווית לכל דגימה המשייכת אותה למחלקה מסוימת
}

```

- שימו לב ש- $dataMatrix[i][j]$ מייצג את ערך הביטוי של גלאי גן i על דגימה j (עבור סעיף 6 נדרש שכל שורה המייצגת את ערכי הביטוי של גן מסוים על כל הדגימות תיוצג בתור המימד הראשון של המערך הדו מימדי).

תיאור סכמתי של המערכים המתארים אוסף נתונים (S מייצג את מספר הדגימות ו- G את מספר גלאי הגנים):



השדה `labelSubsetNumber` מציין כמה קבוצות שונות של דגימות קיימות לפי הקריטריון שניתן בתור הארגומנט השני של התוכנית (`labelType`). ערך השדה שווה למספר הערכים הייחודיים הקיימים במערך המאוחסן בשדה `labels`.

הערות אחרונות לפני תחילת המימוש:

- אין צורך לבצע בדיקות על הקלט בתרגיל זה – ניתן להניח שהמשתמש מפעיל את התוכנית עם ארגומנטים נכונים, ושקובץ הקלט הוא בפורמט SOFT תקין העונה על המבנה שמוגדר בהמשך.
- יש לממש את המחלקה בצורה כללית כך שתוכל לפעול על כל קובץ ביטוי גנים בפורמט SOFT (המקיים את המבנה ואת ההנחות שיובאו בהמשך). בסעיף 8 נריך את התוכנית על קובץ נוסף. כלומר – אין לקודד ערכים ספציפיים לאוסף נתונים מסוים (כמו מספר הדגימות או שם של מחלקת דגימות מסוימת).

קדימה לעבודה...

חלק א: קריאה ועיבוד של קובץ בפורמט SOFT

1. שימרו את קובץ הקלט GDS4085.soft (שניתן לכם בקבצי העזר) בספריית הפרויקט.

הציצו בקישור הבא המכיל תיאור של אוסף נתונים זה:

<http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4085>

2. פיתחו את הקובץ בתוכנת גיליון נתונים והתרשמו מהפורמט שלו.

תיאור מלא של פורמט ה-SOFT המשמש לאחסון נתוני ביטוי גנים במאגר ה-GEO נמצא בקישור:

<http://www.ncbi.nlm.nih.gov/geo/info/soft2.html#SOFTformat>

- ✓ שימו לב: שורה המתחילה עם ^ מייצגת הגדרת ישות חדשה. שורה המתחיל עם ! מייצגת מאפיין של הישות.
- ✓ טבלת הנתונים שמופיעה בסוף קובץ הקלט מופרדת ע"י טאב ("\t").

הנה סקירה של שדות המחלקה GeneExpressionDataset שהוצגה לעיל, וכיצד למלא אותם בנתונים מקובץ הקלט:

אלמנט בקובץ הקלט שמכיל את הנתונים	שם השדה במחלקה GeneExpressionDataset
!dataset_title	String datasetTitle
!dataset_sample_count	int samplesNumber
!dataset_feature_count	int genesNumber
!dataset_table_begin !dataset_table_end טבלת הנתונים מופיעה בין השורה: לבין השורה:	String[] sampleIds String[] geneIds String[] geneSymbols float[][] dataMatrix
✓ השורה הראשונה בטבלת הנתונים מכילה את מזהי הדגימות (ה- GSM801843 ערך לדוגמא:	
✓ העמודה הראשונה משמאל בטבלת הנתונים מכילה את מזהי גלאי הגנים (ה- NM_016242.2_psr1 ערך לדוגמא:	
✓ העמודה השנייה משמאל בטבלת הנתונים מכילה את שמות הגנים	

<p>EMCN (geneSymbols). ערך לדוגמא: ✓ הערכים המספריים בטבלת הנתונים יאוחסנו בשדה dataMatrix.</p> <p>הערה טכנית: יתכן שעבור מספר גלאי גנים שונים יופיע אותו שם גן.</p>	
<p>מעריך ה-labels יגדיר עבור כל דגימה לאיזה תת-קבוצה (מחלקה) היא שייכת. כדי למלא מעריך זה יש לעבד את המקטעים בקובץ שמגדירים תתי קבוצה של דגימות, ומתחילים ב-"^SUBSET". מקטע SUBSET מכיל את המאפיינים הבאים:</p> <ul style="list-style-type: none"> • התווית של תת הקבוצה (שורה המתחילה ב-!subset_description) • רשימת הדגימות המשתייכות לתת הקבוצה כשהן מופרדות בפסיקים (בשורה המתחילה ב-!subset_sample_id). • סוג התווית (שורה המתחילה ב-!subset_type). <p><u>יש להתייחס רק למקטעי SUBSET אשר שייכים לסוג התווית שניתן בתור הארגומנט השני של התוכנית ואח"כ גם צוין בקריאה לפונקציה (labelType).</u></p> <p>עבור קובץ הנתונים הספציפי שלנו, סוג התווית היחיד שקיים הוא "genotype/variation" וכל תא במעריך labels יכיל את הערך "estrogen receptor-positive" או "estrogen receptor-negative" בלבד.</p> <p>הצעה: בזמן קריאת הקובץ, בכל פעם שנתקלים במקטע מסוג SUBSET המתאר תת-קבוצה (מחלקה) מסוג labelType, כדאי לאחסן במעריך אחד את שם תת-הקבוצה, ובמעריך שני את רשימת הדגימות ששייכות אליה. בגמר קריאת הקובץ ניתן לבנות ממערכי עזר אלו את מעריך ה-labels הסופי.</p> <p>שימו לב שפורמט SOFT מאפשר לשייך דגימה מסוימת לכמה תוויות שונות (למשל דגימה מסוימת יכולה להיות שייכת לתווית MALE לפי סוג תווית GENDER ולתווית YOUNG לפי סוג תווית AGE. בתוכנית שלנו, נתייחס בכל הרצה רק לסוג תוויות שניתן בתור הארגומנט השני לתוכנית.</p>	String[] labels

3. כתבו מתודה המקבלת שם קובץ המציין קובץ קלט בפורמט SOFT, ומחזרת המציינת סוג תווית, ומחזירה אובייקט מסוג GeneExpressionDataset הכולל בין היתר עבור כל דגימה את התווית שמתאימה לה עבור סוג התווית שצוין.

חתימת המתודה:

public static GeneExpressionDataset parseGeneExpressionFile (String filename, String labelType)
throws IOException

חלק ב': סינון של גלאי גנים בעלי שונות נמוכה

4. באוסף ביטוי גנים טיפוסי, מרבית הגנים כמעט ואינם משתנים על פני כלל הדגימות ועל כן מומלץ כבר בתחילת תהליך ניתוח הנתונים להקטין את גודל אוסף הנתונים שלנו ע"י סילוקם של גלאי גנים לא אינפורמטיביים אלו.

כתבו מתודה המקבלת אובייקט מסוג `GeneExpressionDataset`, ומספר שלם `N` המציין את מספר גלאי הגנים שיש להשאיר באוסף הנתונים לאחר הסינון. המתודה תשנה את אובייקט אוסף הנתונים שקיבלה כך שיכיל רק את `N` גלאי הגנים בעלי השונות הגבוהה ביותר.

כלומר, לאחר פעולתה של המתודה על אוסף הנתונים, מטריצת הביטוי שבו תכיל רק `N` שורות, המייצגות את `N` גלאי הגנים בעלי השונות הגבוהה ביותר באוסף הנתונים המקורי.

המתודה תחזיר את ערך השונות שנקבע כסף (יהיה שווה לשונות של גלאי הגן המדורג במיקום `N` לפי שונות).

חתימת המתודה:

```
private static float applyVarianceFilterOnDataset(
    GeneExpressionDataset geneExpressionDataset,
    int N)
```

שימו לב:

- ✓ יש לעדכן את כל השדות באובייקט אוסף הנתונים המושפעים מעדכון מספר גלאי הגנים.
- ✓ נוסחא לחישוב שונות של וקטור X המכיל n ערכים:

$$Var(X) = \frac{1}{n} \sum_{i=1}^n (X_i - Mean(X))^2$$

$$Mean(X) = \frac{1}{n} \sum_{i=1}^n X_i$$

- ✓ הצעה: חשבו את השונות עבור כל אחד מגלאי הגנים לפי סדר הופעתם באוסף הנתונים ושימרו ערך זה במערך. מיינו עותק של מערך זה, ובעזרתו מיצאו ערך סף אשר מעליו נמצאים `N` גלאי גנים עם שונות גבוהה או שווה לו. עתה עיברו על מערך השונויות המקורי ובחרו `N` גנים עם שונות מעל לסף.
- ✓ על המטריצה המסוננת להכיל תמיד `N` שורות, גם אם יותר מגן אחד הינו בעל שונות ששווה לשונות שנקבעה כסף.

חלק ג': זיהוי גנים מפרידים בין שתי קבוצות של דגימות

5. כתבו מתודת עזר אשר תשמש אותנו לשלוף את ערכי הביטוי של גן מסוים אבל רק על דגימות המשתייכות למחלקת תוויות מסוימת.

חתימת המתודה:

```
public static double[] getDataEntriesForLabel(float[] data, String[] labels, String label)
```

המתודה תחזיר את כל הערכים מהמערך `data`, אשר בתא המקביל להם במערך `labels` מצוי הערך `label`.

לדוגמא, בהפעלת המתודה באופן הבא אנחנו שולחים שורה מהמטריצה המייצגת את ערכי הביטוי של גן `i` על כל הדגימות. המתודה תחזיר מערך המכיל את ערכי הביטוי של הגן `i` רק עבור דגימות שהערך שלהן במערך ה-`labels` שווה ל-`"estrogen receptor-positive"`:

```
getDataEntriesForLabel(geneExpressionDataset.dataMatrix[i], geneExpressionDataset.labels, "estrogen receptor-positive")
```

6. עתה ננסה לזהות גנים אשר ערכי הביטוי שלהם שונים באופן מובהק סטטיסטית בין שתי תת-קבוצות (מחלקות) של דגימות (כלומר, עבור כל גן נבדוק האם ממוצע ערכי הביטוי שלו על דגימות המחלקה הראשונה שונה ממוצע ערכי הביטוי שלו על דגימות המחלקה השניה). לשם כך נשתמש במבחן הידוע `t-test`, אשר מימוש שלו כבר נתון לכם ע"י המתודה:

```
calcTtest(geneExpressionDataset, geneIndex, label1, label2);
```

המתודה `calcTtest` מחזירה ערך בטווח 0-1 (נקרא `p-Value`) אשר יהיה נמוך יותר ככל שהערכים של הגן שונים בין דגימות של שתי הקבוצות (משקף את ההסתברות להיתקל במצב הנצפה באקראי, נהוג להתייחס אל גנים כאל מפרידים בין 2 קבוצות אם ערך ההתאמה שלהם נמוך מ-0.05 או 0.01).

המתודה `calcTtest` משתמשת במימוש חיצוני למבחן `t-test` ועל כן יש להוסיף את קובץ ה-`jar` הבא לפרוייקט: `commons-math3-3.2.jar`

לאחר שהורדנו אותו מהקישור הבא:

<http://apache.spd.co.il/commons/math/binaries/commons-math3-3.2-bin.zip>

תזכורת: כדי להוסיף קובץ `JAR` לפרוייקט באקליפס יש לבחור את הקובץ תחת התפריטים: Project -> Project properties -> Java Build Path -> Libraries -> Add external Jars

7. לסיום, כתבו מתודה אשר תכתוב לקובץ את רשימת הגנים המפרידים בין 2 קבוצות של דגימות.

עבור כל גן באוסף הנתונים שלנו, חשבו את ערך ההתאמה בעזרת המתודה `calcTtest` שתוארה בסעיף הקודם.

הדפיסו לקובץ את רשימת הגנים שהערך שהתקבל ב-`t-test` שלהם נמוך מפרמטר הסף `alpha`. הגנים צריכים להיות ממוינים לפי ערך ההתאמה שלהם, כאשר בראש הרשימה יופיע הגן המובהק ביותר שהינו בעל ערך ההתאמה המינימלי.

עבור כל גן, יש להדפיס במבנה טבלאי את ערך ההתאמה, את מזהה גלאי הגן, ואת שם הגן (יש לעקוב אחר הפורמט של קובץ ההדגמה GDS4085-DiffGenes.txt). המתודה תחזיר את מספר הגנים שזוהו כמפרידים ונכתבו לקובץ.

חתימת המתודה:

```
public static int writeTopDifferentiallyExpressedGenesToFile(
    String outputFilename,
    GeneExpressionDataset geneExpressionDataset,
    double alpha,
    String label1,
    String label2) throws IOException
```

פורמט קובץ הפלט:

1	0.000013	NM_002996.3_psr1_at	CX3CL1
2	0.000069	NM_001005915.1_psr1_at	ERBB3
3	0.000106	NM_178155.1_psr1_a_at	FUT8
4	0.000149	NM_005623.2_psr1_at	CCL8
5	0.000197	NM_012243.1_psr1_s_at	SLC35A3
6	0.000247	NM_001826.1_psr1_s_at	CKS1B
7	0.000452	NM_004776.3_psr1_at	B4GALT5
8	0.000515	NM_014216.3_psr1_at	ITPK1
9	0.000595	NM_000127.2_psr1_at	EXT1
10	0.000613	NM_002300.4_psr1_x_at	LDHB

בעמוד האחרון תוכלו לראות שמספר גנים מאלו שדורגו בראש הרשימה שהפקנו, תועדו בעבר בספרות המדעית כבעלי תפקיד חשוב בסרטן השד ...

8. בדקו שהתוכנית שלכם עובדת על אוסף נתונים אחר השמור אף הוא בפורמט SOFT. שימרו בספריית הפרויקט את הקובץ GDS5088.soft (הכולל נתונים לגבי ביטוי גנים בשלבים שונים במהלך היריון) והריצו את התוכנית GeneExpressionAnalyzer שכתבתם תוך שימוש בארגומנטים הבאים:

```
GDS5088 "development stage" "pregnancy_trimester 1"
"pregnancy_trimester 3" 0.01
```

הגישו את הקובץ GDS5088-DiffGenes.txt שהתוכנית יצרה

שימו לב שבקובץ זה כל דגימה משויכת ליותר מתונית אחת ולכן חובה לממש נכון את סינון מקטעי ה-SUBSET על פי סוג התונית.

למתעניינים - תיאור של אוסף הנתונים ניתן למצוא בקישור הבא:

<http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5088>

פלט ההרצה:

```
Gene expression dataset loaded from file GDS5088.soft.
Dataset contains 48 samples and 33297 gene probes.
Dataset contains 5 label subsets of type [development stage]: [pregnancy_trimester 2,
pregnancy_trimester 1, pregnancy_trimester 3, Post-Partum, Non-Pregnant]

Applying variance filter on dataset: Kept 16648 top variable genes having variance of at least 0.16.

76 differentially expressed genes identified using alpha of 0.01000 when comparing the two sample
groups [pregnancy_trimester 1] and [pregnancy_trimester 3].
Results saved to file GDS5088-DiffGenes.txt.
```

9. [לא להגשה] הרצת רשות נוספת לצורך בדיקות שלכם:

הריצו את התוכנית GeneExpressionAnalyzer על הקובץ GDS4858.soft (שניתן גם הוא במסגרת קבצי התרגיל) תוך שימוש בארגומנטים הבאים:

```
GDS4858 gender "men" "women" 0.01
```

פלט ההרצה:

```
Gene expression dataset loaded from file GDS4858.soft.
Dataset contains 22 samples and 54675 gene probes.
Dataset contains 2 label subsets of type [gender]: [women, men]

Applying variance filter on dataset: Kept 27337 top variable genes having variance of at least 0.04.

367 differentially expressed genes identified using alpha of 0.01000 when comparing the two sample
groups [men] and [women].
Results saved to file GDS4858-DiffGenes.txt.
```

הגלים שיופיעו בראש הקובץ GDS4858-DiffGenes.txt:

1	0.000000	_224589at	XIST
2	0.000000	_224936at	EIF2S3
3	0.000000	_206042x_at	SNURF
4	0.000001	_227671at	XIST
5	0.000001	_224588at	XIST
6	0.000001	_201909at	RPS4Y1
7	0.000002	_201522x_at	SNURF
8	0.000002	_206700s_at	KDM5D
9	0.000002	_204409s_at	EIF1AY
10	0.000002	_205000at	DDX3Y
11	0.000002	_201016at	EIF1AX
12	0.000002	_221728x_at	XIST
13	0.000003	_209771x_at	CD24
14	0.000003	_214218s_at	XIST
15	0.000003	_204410at	EIF1AY

שאלה למחשבה: גגלו ומצאו מה מאפיין את הגנים בראש הרשימה...

בהצלחה!

The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation

Thomas Holbro*, Roger R. Beerli[†], Francisca Maurer*, Magdalena Koziczak*, Carlos F. Barbas III[†], and Nancy E. Hynes*[‡]

*Friedrich Miescher Institute, P.O. Box 2543, 4002 Basel, Switzerland; and [†]The Skaggs Institute for Chemical Biology and Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037

Edited by Owen N. Witte, University of California, Los Angeles, CA, and approved May 28, 2003 (received for review December 17, 2002)

ErbB2 is a receptor tyrosine kinase whose activity in normal cells depends on dimerization with another ligand-binding ErbB receptor. In contrast, amplification of *c-erbB2* in tumors results in dramatic overexpression and constitutive activation of the receptor. Breast cancer cells overexpressing ErbB2 depend on its activity for proliferation, because treatment of these cells with ErbB2-specific antagonistic antibodies or kinase inhibitors blocks tumor cells in the G₁ phase of the cell cycle. Intriguingly, loss of ErbB2 signaling is accompanied by a decrease in the phosphotyrosine content of ErbB3. On the basis of these results, it has been proposed that ErbB3 might be a partner for ErbB2 in promoting cellular transformation. To test this hypothesis and directly examine the role of the "kinase dead" ErbB3, we specifically ablated its expression with a designer transcription factor (E3). By infection of ErbB2-overexpressing breast cancer cells with E3, we show that ErbB3 is essential for the proliferation process. Loss of ErbB3 has dramatic effects on cell proliferation and on the expression of constitutively active ErbB2. Proliferative block induced by E3 is dependent on ErbB3 signaling. These results show that ErbB3 signaling and activity alone are insufficient for cell division. Furthermore, we identify a novel pathway by which active ErbB2 is coupled to the phosphatidylinositol 3-kinase pathway. Thus, the ErbB2/ErbB3 heterodimer functions as an oncogenic unit to drive breast tumor cell proliferation, which proceeds via inhibition of intracellular signaling pathways and directly targets various members of the cell cycle machinery (17–20).

eration, which proceeds via inhibition of intracellular signaling pathways and directly targets various members of the cell cycle machinery (17–20).

Interestingly, expression of ErbB3 is seen in many tumors that express ErbB2, including breast (21), bladder (22), and others. Furthermore, in many ErbB2-overexpressing breast tumors, ErbB3 has elevated levels of phosphotyrosine (15). ErbB3 itself has impaired tyrosine kinase activity (23) and needs a dimerization partner to become phosphorylated and acquire signaling potential (24). Indeed, we and others have shown that inactivation of ErbB2 leads to decreased ErbB3 tyrosine phosphorylation (17, 18, 25, 26). ErbB3, which contains six docking sites for the p85 adaptor subunit of phosphatidylinositol 3-kinase (PI3K), efficiently couples to this pathway (27, 28). Interestingly, it has

The family of ErbB receptors includes members: epidermal growth factor receptor (ErbB1), ErbB2, ErbB3, and ErbB4. Overexpression of ErbB2-related growth factor family receptors results in the formation of heterodimers. Ligand binding induces the dimerization of ErbB receptors, ultimately leading to stimulation of intracellular signaling cascades (1, 2). The physiological function of ErbB ligand signaling, is to promote cell proliferation. ErbB2 appears to be the predominant receptor in the ErbB family (5, 6). The intracellular signaling in normal development is regulated by genetically modified mice. Overexpression of ErbB3 and ErbB2/ErbB4 heterodimers in mice. ErbB3 has a similar impact on cell proliferation, whereas loss of ErbB2 or ErbB3 results in impaired development (8, 9).

A wealth of clinical data has shown that overexpression of tyrosine kinases, in particular ErbB2, is a key factor in human cancer development, and is a target for cancer therapies (10–13). Overexpression of ErbB2 is attributable to gene amplification and is associated with cancer and correlates with overall survival (14). Overexpression of ErbB2 is phosphorylated in breast cancer cells (15, 16). It has been observed that overexpression of ErbB2 results in efficient in-

www.pnas.org/cgi/doi/10.1073/pnas.0305111100

Essential function for ErbB3 in breast cancer proliferation

<http://breast-cancer-research.com/content/6/3/137>

The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation

<http://www.pnas.org/content/100/15/8933.short>

CX3CL1 expression is associated with poor outcome in breast cancer patients.

<http://www.ncbi.nlm.nih.gov/pubmed/23912959>

Role of Cks1 amplification and overexpression in breast cancer.

<http://www.ncbi.nlm.nih.gov/pubmed/19161979>