**תרגיל כיתה 10 – מרחק Hamming ורשתות**

**מרחק Hamming**

בשיעור הקודם דיברנו על דמיון ומרחק, ולמדנו על מרחק אוקלידי וקורלציות ספירמן ופירסון. מרחק נוסף עליו נדבר בקצרה היום נקרא מרחק המינג (Hamming) – המרחק בין שני רצפים. חישוב מרחק המינג נעשה ע"י ספירת מספר הפעמים בהם יש הבדל בין שני רצפים במיקום מסויים.

|  |  |
| --- | --- |
| **String** | **Hamming Distance** |
| 1 0 0 0 0 1 0 0 1 1 0 0 | 2 |
| 1 0 1 0 0 1 0 0 0 1 0 0 |
|  |  |
| Point | 1 |
| Paint |

אפשר להשתמש במרחק המינג על מנת לחשב את המרחקים בין רצפי נוקלאוטידים, את המרחק בין אנשים במולקולות HLA (עבור תרומת מח עצם), רצפי חומצות אמינו וכו'.

שאלה: על רוב התאים שבגופנו נמצאות מולקולות ממשפחת ה-Human Leukocyte Antigen או HLA. מערכת החיסון משתמשת במולקולות אלו על מנת לזהות אילו תאים שייכים לגוף ואילו הם זרים. על מנת לתרום מח עצם, דם טבורי או איברים, יש צורך להתאים את המולקולות האלו בין התורם לנתרם במידה רבה ככל האפשר על מנת שלא תהיה דחייה של השתל. לתהליך הזה קוראים HLA Typing או HLA Matching. הבעיה היא, שבניגוד לתרומת דם למשל, כאן יש צורך להתאים בין מספר גדול של מולקולות HLA. בדרך כלל רצוי שיהיו בין 8-10 מולקולות תואמות על מנת לבצע את ההשתלה.

בקובץ hladb.RData ישנם נתונים עבור 19 מולקולות HLA עבור 90 אנשים – כאשר הראשון מבינהם הוא הנתרם. אנו צריכים למצוא מי מבין 89 האנשים האחרים יכול לתרום לו בעזרת חישוב מרחק Hamming בינו לבין שאר האנשים.

load("hladb.RData")

samples = hladb$Sample

hamming = data.frame()

sample1 = hladb[which(hladb$Sample == "Recipient"),-1]

for(i in 2:length(samples)) {

sample2 = hladb[which(hladb$Sample == samples[i]),-1]

hammingDistance = sum(sample1 != sample2)

hamming = rbind(hamming, data.frame(samples[i], hammingDistance))

}

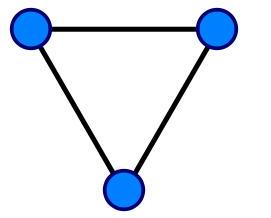
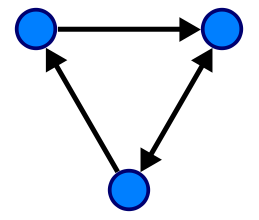
colnames(hamming) = c("Sample2", "Hamming")

hamming$Sample2[hamming$Hamming <= 9][1] "ERR188027"

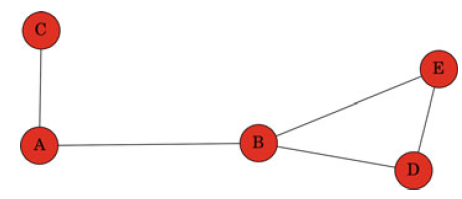
את חישוב מרחק Hamming אנחנו עושים בעזרת הביטוי sum(sample1 != sample2) שתפקידו להשוות בין הביטויים ב-sample1 ו-sample2, להחזיר TRUE או FALSE במקרה של אי התאמה או התאמה בהתאם עבור כל אחת מהעמודות (שהן מולקולות ה-HLA). לבסוף, הביטוי יסכום את מספר הפעמים שבהם אין התאמה.

**רשתות**

כאשר אנחנו עושים ניתוח נתונים רב-מימדיים למידע ביולוגי, נצפה שיהיו קשרים רבים בין שניים או יותר מהמשתנים. דוגמאות לכך יכולות להיות גנים מסויימים שמשתנים יחד, הורמונים המופרשים כתלות זה בזה, נוירונים היורים יחד ושרשראות הדבקה של מחלות מדבקות.

על מנת להבין את הקשרים האלו נרצה לבנות רשתות (networks) של קשרים, כשהמרחק בין שתי דוגמאות/מימדים הוא הבסיס לקביעה יש קשר בין שתי הדוגמאות/מימדים הללו. את הרשתות שנמצא נוכל לייצג באמצעות גרפים. גרף של רשת הוא אוסף של **נקודות** ('צמתים', nodes), ואוסף של **קשרים** ('קשתות', edges) בין הנקודות הללו.

הקשרים יכולים להיות עם כיוון (directed) או ללא כיוון (undirected) – כלומר, ייתכן ונקבל מידע על הכיוון שבו הקשר מתקיים (הורמון X גורם להפרשה של הורמון Y, אבל הורמון Y לא גורם להפרשה של הורמון X).



לדוגמא, באיור הזה יש דוגמא לרשת ללא כיוון (אין בה חיצים בקשתות שמצביעים על כיווניות), עם 5 צמתים ו-5 קשתות.

כדי להבין איך כל צומת קשורה לצומת אחרת, נשתמש במטריצת המרחקים עליה למדנו בשיעור הקודם. המרחק במטריצה מצביע לנו על הקשר בין כל שני צמתים. כעת נקבע סף מסויים שמגדיר לנו "קשר". הסף הזה תלוי בנתונים שלנו – ולפיו נחליט איזה ערכים במטריצה נחשבים "קשר" ואיזה לא. נבנה מחדש את המטריצה כך שתהיה בנויה מ-1 ו-0 – כאשר מציבים 0 כשאין קשר ו-1 כשיש. למטריצה כזו קוראים **מטריצת סמיכות** (Adjacency Matrix) ואיתה ניתן לבנות את הגרף.

**שאלה 1:**

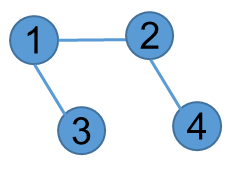
נתונה מטריצת דמיון המחושבת לפי קורלציית פירסון בין גנים. שרטט את רשת הגנים המתקבלת מסף 0.9. מה המשמעות של הרשת?

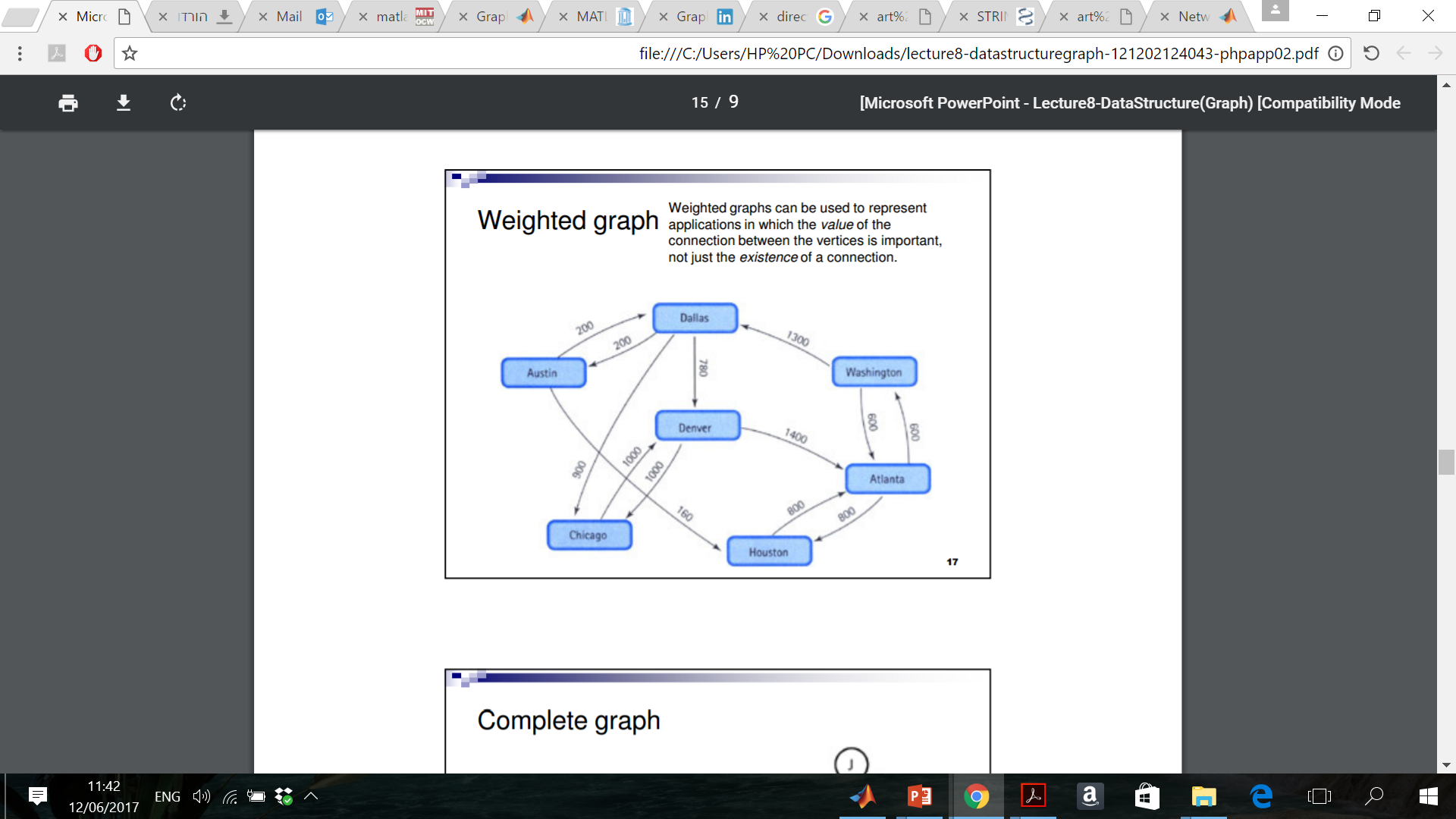
**תשובה:**

מכיוון שמה שמעניין אותנו הוא אם קיימת קורלציה ברמת הביטוי של הגנים בין אם הקורלציה חיובית או שלילית, נסתכל על הערך המוחלט של קורלציית פירסון בהשוואה לסף. הסף שלנו הוא 0.9 ולכן מטריצת הסמיכות המתקבלת:

(למה איפסנו את האלכסון? מכיוון שאין משמעות לקורלציה של גן עם עצמו).

על מנת לשרטט את הגרף, כל גן הוא צומת וכל קשת בין שני צמתים מעיד על קיום קורלציה. נקבל את הגרף הבא:



הגדרות נוספות:

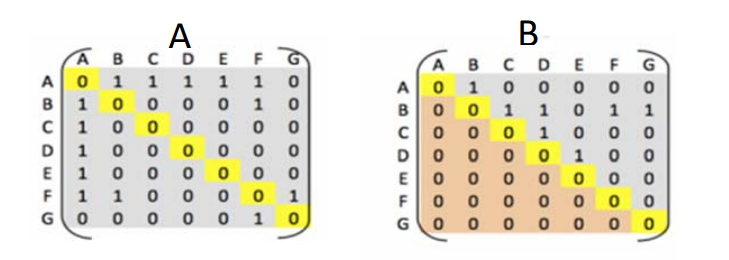
לפעמים לכל קשת יש **משקל** שיכול להיות מרחק בין הצמתים או כל ערך שמצביע על עוצמת הקשר בין שני הצמתים. למשל, בדוגמא הזו, הקשר (או המרחק) בין Austin ו-Dallas הוא 200, והוא קצר פי 3 מהמרחק בין Washington ל-Atlanta, שהוא 600.

כדי לעבור בין נקודה לנקודה ישנם **מסלולים** שונים שניתן לקחת. מסלול מוגדר כדרך המחברת בין שתי נקודות. לדוגמא, בין אטלנטה לאוסטין יש רק מסלול אפשרי אחד:

בין אוסטין לאטלנטה יש כמה מסלולים,למשל:

|  |  |  |
| --- | --- | --- |
| **מרחק או מסלול גיאודזי** | המסלול הקצר ביותר בין שני צמתים |  |
| **מסלול ממוצע** | ממוצע המרחק הגיאודזי בין כל הצמתים ברשת (למעט singletons) |  |
| **Degree** | מספר הקשרים שיש לצומת   * Indegree – מספר הקשתות בכיוון הצומת * Outdegree – מספר הקשתות היוצאות מהצומת |  |
| **Hub** | צומת בעל קישוריות גבוהה |  |
| **Singleton** | צומת ללא קישוריות |  |
| **Betweenness Centrality** | מספר הפעמים שצומת מופיע במסלול הגיאודזי בין שני צמתים | במקרה הזה הצומת הצהוב הוא הצומת עם הדרגה הכי גבוהה וגם הצומת עם המרכזיות הכי גבוהה.    לצומת ב דרגה יותר גבוהה, אבל לצומת א מרכזיות יותר גבוהה. |
| **Clique** | אזור בו כל הצמתים מקושרים |  |
| **קוטר הגרף** | המסלול הכי ארוך מכל המסלולים הגיאודזים |  |

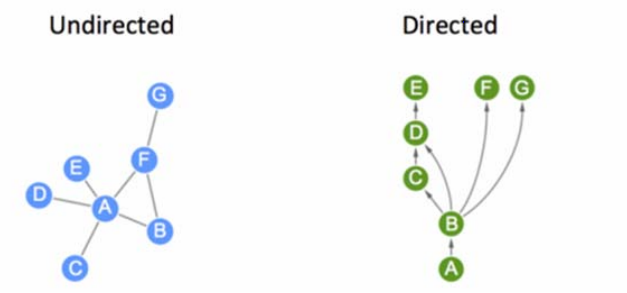
**שאלה 2:**

נתונות שתי מטריצות שכנות/דמיון:

1. שרטט את הרשתות המוגדרות מכל מטריצה.
2. זהה בכל רשת את הצומת בעל ה- קישוריות (degree) הגדולה ביותר.
3. מה הקוטר של כל רשת?
4. מה ההשלכות של להוציא את צומת B מכל אחד משתי הרשתות ?
5. מה המשמעות של להוציא את צומת A מכל אחד משתי הרשתות ?

**תשובה:**

1. ניתן לראות כי מטריצה A מציגה קשרים דו-כיווניים (למשל A קשור ל-B, ו-B קשור ל-A) בעוד שמטריצה B מציגה קשרים חד-כיווניים (A קשור ל-B, אבל B לא קשור ל-A). ניתן אם כך להסיק שמטריצה A היא ללא כיווניות (undirected) ומטריצה B היא עם כיוון (directed). נשרטט:



2. בגרף של מטריצה A (undirected) – ניתן לראות כי לצומת A יש את הקישוריות (degree) הגבוהה ביותר. במטריצה B (directed) הגרף הוא directed אז ה-indegree הגבוה ביותר הוא של צומת D (2 קשתות נכנסות) וה-outdegree הגבוה ביותר הוא של צומת B (4 קשתות יוצאות).

3. עבור מטריצה A (undirected) –(C🡪A🡪F🡪G)

ועבור מטריצה B (directed) – (A🡪B🡪D🡪E)

4. בגרף undirected – הרשת לא תיפגע, ה-degree של נקודות A ו-F יקטנו ב-1.

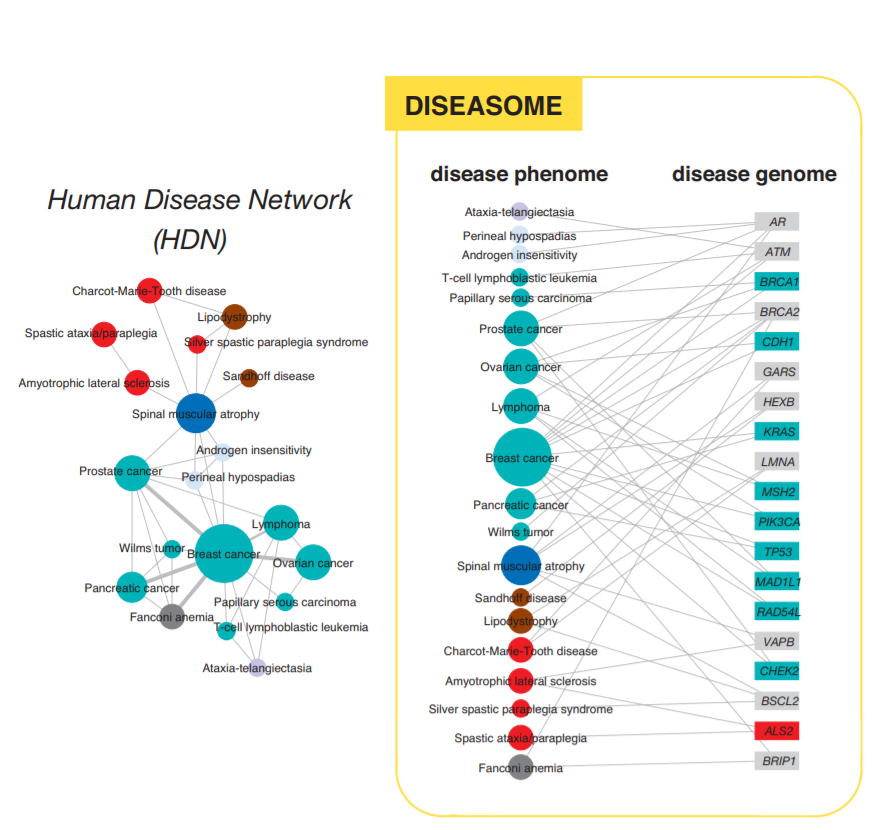
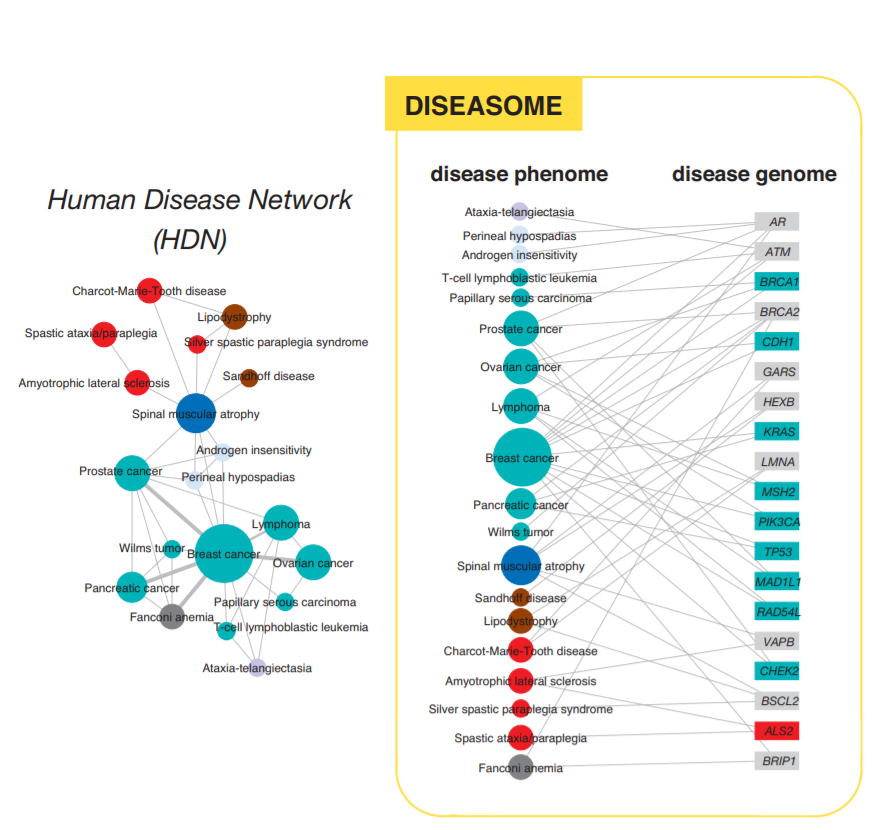
בגרף directed – הרשת תיפגע מכיוון שצומת B היא hub – נקודות A,F,G יהפכו ל-singleton, ה-degree של נקודות C ו-D יקטנו ב-1.

5. בגרף undirected – הרשת תיפגע מכיוון שצומת A היא hub – נקודות E,D,C יהפכו ל-singleton, ה-degree של נקודות B,F יקטנו ב-1.

בגרף directed – נראה כאילו הרשת כמעט ולא תיפגע, אך קודקוד A הוא מקור הסיגנל ברשת והעלמה שלו תעביר את התפקיד לקודקוד B.

**שאלה 3:**

הרשתות הבאות נלקחו מהמאמר The human disease network (Goh et al. PNAS 2007) ומתארות את הקשר בין מחלות לבין גנים שמוטציה בהם נמצאה כקשורה למחלה. גודל כל צומת מתארת את מספר הגנים שנמצאו קשורים למחלה והצבע מצביע על סוג המחלה.



ניתן לראות כי מצד ימין הרשת מופיעה כקשרים בין כל מחלה לגנים שלה, ומצד שמאל הרשת היא הקשרים בין המחלות עצמן.

1. האם ניתן להציג את המידע הזה בצורת רשת אחרת המבוססת גנים? אם כן תאר מה תהיה משמעות הצמתים ומה תהיה משמעות הקשתות?
2. מה מייחד את מחלת Spinal Muscular Atrophy?
3. ביכולתנו לרצף גנים על מנת למצוא מוטציות. האם יש גן אחד אותו נוכל לרצף כדי לזהות סרטן? אם כן – מיהו הגן? אם לא, האם יש סט מינימלי של גנים שיוכלו להבדיל בין סוגי הסרטנים הנתונים ברשת לבין כל מחלה אחרת?

**תשובה:**

1. כן, ניתן לייצר רשת הדומה לרשת השמאלית, אך שהצמתים יהיו הגנים והקשתות יחברו את הגנים אם הם מתבטאים יחד באותה המחלה:

2. ניתן לראות בגרף השמאלי שהמחלה היא hub בעלת between centrality המשתפת גנים הקשורים למחלות סרטן כמו גם למחלות אחרות.

3. לפי הגרף ששרטטנו לשאלה 1, ניתן לראות כי אין גן אחד האחראי על כל סוגי הסרטן. אין סט מינימלי שמבדיל בין כל סוגי הסרטן לשאר המחלות כי מכיוון ש T-cell lymphoblastic leukemia מקושר אך ורק לגן ATM, והגן הזה משותף למחלות שהן לא סרטן.