פרויקט ביו אינפורמטיקה - חלק 2

מגישים: נדב לאונארונס, אלון קרומר, נועה ברק.

1. סעיף ניקוי הדאטא – הקובץ הנקי מצורף, **'new_cog_words_bac'**.

2. הצגת אלגוריתם ההשוואה שבנינו ב-**High Level** :

נרצה לבטא את הדימיון בין שני הפרופילים בצורה שמתחשבת ב-6 פרמטרים שונים, ונותנת משקל ייחודי לכל פרמטר כזה. המחשבה שלנו היתה שבכל הרצה מתקבלים ערכים אחרים, וההבדלים בין פרופילים יקבל דגש בפרמטרים אחרים בכל פעם. לכן, בנינו מערך "משקולות" עבור כל פרמטר שבאמצעותו מחליטים בצורה נוחה כמה "משקל" לתת לכל פרמטר בהרצה. כך שניתן לייצר אופטימיזציה לאלגוריתם כתלות בדאטא שאיתו עובדים, מ שיביא לידי ביטוי את הייחודיות של כל פרופיל.

ששת הפרמטרים שהגדרנו הם :

1. עד כמה כל פרופיל נפוץ בדאטאבייס שלו .
2. האם הפרופילים נמצאים בעיקר בייצורים מ-Phylum דומה?
3. האם שני הפרופילים מופיעים במגוון דומה של Phylum ?
4. האם הפרופילים מרוכז בעיקר באותם ייצורים זהים ?
5. האם הפרופילים נמצאים בעיקר בייצורים שחיים ב-habitats דומה?
6. האם הסדרים הנפוצים ביותר שבהם הקלסאטר מופיע זהים בין שני הפרופילים?

הרעיון החישובי

- לכל פרמטר מהרשימה, ניתן ציון, ועליו נבצע נירמול אל תוך הטווח של 0-100 (פונקציית normalize).
- בנוסף יהיה לנו מערך של משקולות שכולן בין 0 ל-1, וסכומם 1. המשקולות בעצם ייתנו משקל ייחודי לכל פרמטר מהרשימה (כדי שיהיה קל לבצע מודיפיקציות בשלב ההרצות).
- בסוף נבצע ממוצע משוקלל, כלומר נכפיל כל תוצאה של פרמטר במשקולת המתאימה לה, ונקבל מס' אחד שמבטא את הדימיון הכולל בין הפרופילים.

לדוגמא:

אם ציוני ההבדל לסעיפים הם לפי הסדר הבא : [80,40,30,10,90,100].

ונניח שהמשקולות הן : [0.2,0.2,0.1,0.1,0.1,0.3]

אז הציון הסופי יהיה : $80 * 0.2 + 40 * 0.2 + 30 * 0.1 + 10 * 0.1 + 90 * 0.1 + 100 * 0.3 = 67$

<u>פירוט צורת החישוב של כל פרמטר :</u>

1. עד כמה הוא נפוץ בתור הDB שלו :

- החישוב : מספר ה-Occurances של הקלאסטר, חלקי מספר השורות ב-DB שלו כפול 100.
- לציון שכל פרופיל מקבל, נעשה ערך מוחלט על ההפרש ביניהם, ואת זה נחסיר ממאה. לדוגמא אם אחד קיבל ציון 80 ואחד ציון 60, ההפרש יהיה 20, וציון הדימיון לסעיף זה יהיה 80=100-20 .

2. שלושת ה-phylum שבו הקלאסטר נמצא באופן הכי נפוץ :

- נתבונן ב-3 ה-phylum שבו הקלאסטר היה הכי נפוץ עבור כל פרופיל (או אם יש פחות אז מה שיש):
  - אם יש חפיפה בכל ה-3 הראשונים - ציון 100 לסעיף זה.
  - אם יש חפיפה ב-2 מתוך 3 הראשונים - ציון 70 לסעיף זה.
  - אם יש חפיפה ב-1 מתוך 3 הראשונים - ציון 40 לסעיף זה.
  - אם אין חפיפה כלל - ציון 0 לסעיף זה.

3. עד כמה הקלסאטר מפוזר על פני phylum-ים מגוונים :

- נסמן: X = כמות הphylum-ים השונים שמופיעים בפרופיל של הפלסמיד.

  Y = כמות הphylum-ים השונים שמופיעים בפרופיל של הכרומוזומים.

- נחשב את הציון של סעיף זה כך : $(1 - (|X - Y| \,/\, max(X, Y)) * 100$

4. שלושת ה-Hosts שבו הקלאסטר נמצא באופן הכי נפוץ :

- נתבונן ב-3 ה-Hosts - שמות הייצורים, שבהם הקלאסטר היה הכי נפוץ עבור כל פרופיל (או אם יש פחות אז מה שיש):
  - אם יש חפיפה בכל ה-3 הראשונים - ציון 100 לסעיף זה.
  - אם יש חפיפה ב-2 מתוך 3 הראשונים - ציון 70 לסעיף זה.
  - אם יש חפיפה ב-1 מתוך 3 הראשונים - ציון 40 לסעיף זה.
  - אם אין חפיפה כלל - ציון 0 לסעיף זה.

5. שלושת ה-Habitats שבו הקלאסטר נמצא באופן הכי נפוץ :

- נתבונן ב-3 ה-Habitats שבו הקלאסטר היה הכי נפוץ עבור כל פרופיל (או אם יש פחות אז מה שיש):
  - אם יש חפיפה בכל ה-3 הראשונים - ציון 100 לסעיף זה.
  - אם יש חפיפה ב-2 מתוך 3 הראשונים - ציון 70 לסעיף זה.
  - אם יש חפיפה ב-1 מתוך 3 הראשונים - ציון 40 לסעיף זה.
  - אם אין חפיפה כלל - ציון 0 לסעיף זה.

6. שלושת ה-Orders שבו הקלאסטר נמצא באופן הכי נפוץ :

- נתבונן ב-3 ה-Orders שבו הקלאסטר היה הכי נפוץ עבור כל פרופיל (או אם יש פחות אז מה שיש):
  - אם יש חפיפה בכל ה-3 הראשונים - ציון 100 לסעיף זה.
  - אם יש חפיפה ב-2 מתוך 3 הראשונים - ציון 70 לסעיף זה.
  - אם יש חפיפה ב-1 מתוך 3 הראשונים - ציון 40 לסעיף זה.
  - אם אין חפיפה כלל - ציון 0 לסעיף זה.

3. סעיף מימוש האלגוריתם – קובץ הקוד המלא מצורף כקובץ **IPYNB**, **'mini_project_part_2'**.

4. <u>סעיף תוצאות ההרצה והמסקנות :</u>

הערכים של q1,q2,d אשר בחרנו הם:

D=3 ; **Q2=40** ; **Q1=360**

אם נסתכל על שני הקבצים שיש לנו נוכל לראות את הנתונים הבאים:

בקובץ של הפלסמידים יש 922 גנומים שונים, הכוללים 308,940 קודונים של 4 ספרות (כלומר אותיות COG).

לעומת זאת, בקובץ של ה-BACTERIAL GENOMES, קיימים 520 גנומים שונים ויש בו 5,290,664 קודונים של 4 ספרות.

ניעזר במספרים אלו על מנת לחשב את היחס בין Q1 לQ2 שלנו, מכיוון שמספרים אלו משפיעים על ההסתברות לתדירות של רצף מסויים לחזור על עצמו במספר גנומים שונים. נשים לב שמספר גנומים גבוה יותר יעלה את ההסתברות, וגם מספר קודונים גבוה יעלה את ההסתברות למציאת רצף מסויים באקראי.

$$\frac{Q1}{Q2} = \frac{5290664}{308940} \times \frac{520}{920} \approx 9$$

לכן היחס שנבחר בין ה-Q-ים הוא 9. בנוסף, נעדיף Q כמה שיותר גבוה , כי זה יחזיר תוצאות שהופיעו יותר פעמים מה שכנראה יותר משמעותי, ולכן אחריי ניסוי וטעייה קיבלנו שאלו ה-Qים הגבוהים ביותר עבור d=3, שמשמרים את היחס ומחזירים תוצאות רלוונטיות.

לאחר הרצה נקבל שהקלאסטר שקיבלנו הוא : '0444', '0601', '1173'

מהקלאסטר שמנו לב שקיבלנו "abc transporter" שהיא משפחה של חלבונים המהווים משאבות תלויות ATP. משפחה זו משתתפת בתהליכי הובלה של מגוון רחב של חומרים ביניהם מינרלים, סוכרים, חומצות אמינו.

ניתן לחלק את המשפחה לטרנספורטר EFFLUX ו-INFLUX כך שהראשון משמש להוצאת חומרים מהתא והשני להכנסה לתא, ובנוסף ישנם גם טרנספורטרים דרך ממברנות פנים-תאיות. הטרנספורטר משתמש ב-ATP כמקור אנרגיה ובדרך כלל מורכב מארבעה חלבונים: 2 אשר משמשים אותו ליצירת "ערוץ" בממברנה, בעוד השניים האחרים משמשים אותו לקשירת ATP. חשוב לציין כי משפחת החלבונים הזאת נפוצה מאוד במגוון יצורים : חיות, צמחים, שמרים ופטריות.

לכן, לאחר שמצאנו את הקלאסטר הזה נסיק כי בנוסף לשלושת הקודונים שמצאנו (שתיים פרמיאז ואחת אנטיפיאז), כנראה **<u>קיים עוד קודון</u>** במבנה הזה. לאחר חיפוש נוסף, קיבלנו את הקלאסטר הבא:

'0444', '0601', '1173','0747'.

מצורף פה צילום של תוצאת ההרצה המלאה :

```
                              Cluster found:
                             ('0444', '0601', '0747', '1173')
Data from new_cog_words_bac.txt:
Numbers of appearances:
316
Top 3 phylums:
{'Firmicutes', 'Proteobacteria', 'Actinobacteria'}
number of phylums:
12


Top 3 hosts:
{'Agrobacterium_radiobacter_K84_uid58269', 'Agrobacterium_H13_3_uid63403', 'Agrobacterium_tumefaciens_C58_uid57865'}


Top 3 habitat:
{'Host', 'Soil and Sediment', 'Not Annotated'}


Top 3 orders:
{"['0747', '1173', '0601', '0444']", "['0747', '0601', '1173', '0444']", "['0601', '1173', '0444', '0747']"}



Data from cog_words_plasmid.txt:
Numbers of appearances:
54
Top 3 phylums:
{'Euryarchaeota', 'Proteobacteria', 'Cyanobacteria'}
number of phylums:
5


Top 3 hosts:
{'Rhizobium_etli_CFN_42_uid58377', 'Rhizobium_leguminosarum_bv__viciae_3841_uid57955', 'Rhizobium_leguminosarum_bv__trifolii_WSM1325_uid58991'}


Top 3 habitat:
{'Plant', 'Soil and Sediment', 'Not Annotated'}


Top 3 orders:
{"['0747', '1173', '0601', '0444']", "['0747', '0601', '1173', '0444']", "['0601', '1173', '0747', '0444']"}
```

מצורף פה המידע הרלוונטי על כל COG מהקובץ :

COG0747 ; Amino acid transport and metabolism;ABC-type transport system, **periplasmic component**;

COG1173 ;Inorganic ion transport and metabolism;ABC-type dipeptide/oligopeptide/nickel transport system, **permease component**;

COG0601 ;Amino acid transport and metabolism;METABOLISM;Inorganic ion transport and metabolism;ABC-type dipeptide/oligopeptide/nickel transport system, **permease component**;

COG0444 ;Amino acid transport and metabolism;METABOLISM;Inorganic ion transport and metabolism;ABC-type dipeptide/oligopeptide/nickel transport system, **ATPase component**;

ניתוח ההבדלים :

תחילה ברור כי המשאבה שמצאנו חיונית ושכיחה מאוד בדאטאבייס.

בפרופילים שהוחזרו, נראה שיש שוני בטופ 3 PHYLA של כל DATABASE. דבר זה יכול להעיד על ספציפיות של התפקוד, למשל ה-'Euryarchaeota' phylum שנמצא בטופ 3 של קובץ הפלסמיד שבדרך כלל נמצא בסביבות עם תנאים קיצוניים כמו ב-SMALL INTESTINE. בנוסף 'Cyanobacteria' גם מופיע שם והוא בעל יכולות פוטוסינטטיים.

הדומיננטיות של PHYLA אלו יכולה להעיד על התאמות שצריך בשביל סביבות קיצוניות ובעייתיות (במקרה שלנו גם יכולות מטבוליות) שאולי לא יהיו נחוצות בכרומוזומים (ש"בונים" על היכולות האלו שקיימות אצל הפלסמיד). בסביבה קיצונית, ניתן להניח כי היצור יהיה צורך בשימוש מוגבר בטרנספורטר, ולכן אם קיימים עותקים רבים בפלסמידים, הדבר מהווה יתרון ומאפשר תגובה מהירה ויעילה יותר בעת הצורך.

בנוסף נוכל לראות שמבחינת ה-HABITAT בקובץ הפלסמיד הקלאסטר שלנו מופיע הרבה בצמחים, בעוד שבכרומוזומים הוא מופיע הרבה בבני אדם. דבר זה יכול להעיד בנוסף ליכולות הפוטוסינטטיות שראינו קודם שבאמת הקלאסטר בפלסמידים מיוחס לצמחים ולפעילות הצמחית שלא נמצאת אצל בעלי חיים כמו בני אדם.

נספחים והיכרות נוספת עם המשאבה :

Figure 1. Exemplar tree showing phylogenetic relationships of Arabidopsis and 2 rice proteins in ABC subfamily B. We Brought this figure as an example of the vast variety found in the ABC-Transporter Family.

Figure 2. The various parts of the ABC-Transporter protein (in INFLUX transporter).

The TMD domain is where the membrane is crossed (usually hydrophobic). In our case, COG0747 is the periplasmatic domain and is in charge of moving the substrate through the membrane.

The NBD domain, where the ATP binds in the cytoplasme – in our case, COG 0444 is the ATPASE – the site in the NBD where the ATP->ADP happens.

COG1173 and COG0601 are both the permease components. The red site is the bound substrate.

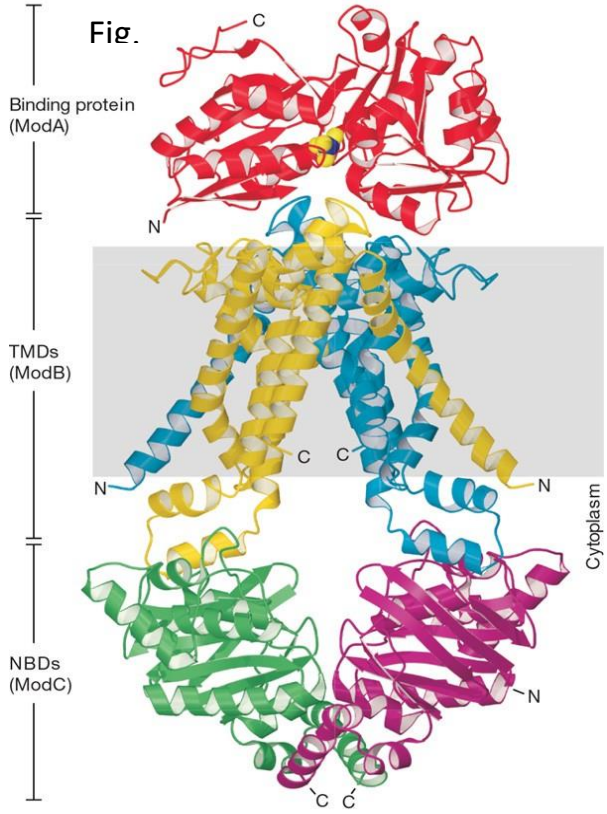
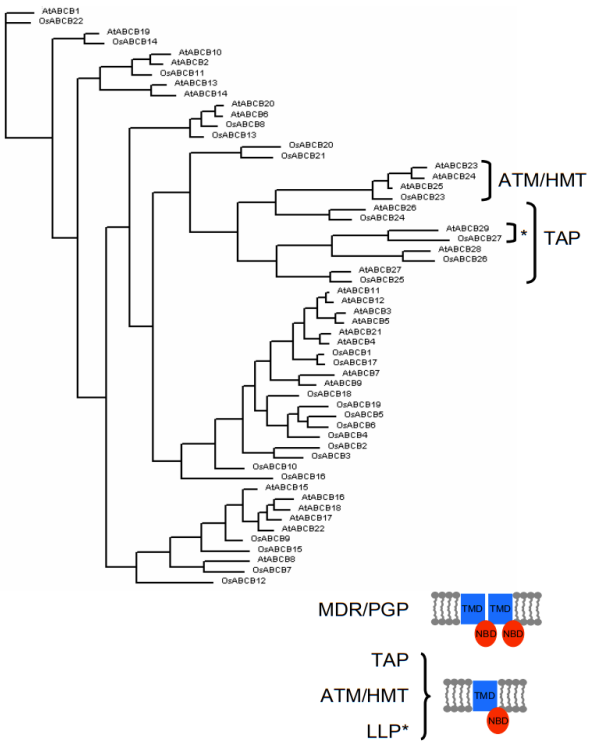


Fig. 1



ביבליוגרפיה של התמונות :

1. Dassa E, Bouige P. The ABC of ABCs: ***A phylogenetic and functional classification of ABC systems in living organisms***. *Res Microbiol.* 2001;**152**:211–229. doi: 10.1016/S0923-2508(01)01194-9. [PubMed] [CrossRef] [Google Scholar]

2. Hollenstein, K., Frei, D. & Locher, K. ***Structure of an ABC transporter in complex with its binding protein***. *Nature* **446**, 213–216 (2007). https://doi.org/10.1038/nature05626