סיכום פרוייקט בביואינפורמטיקה

סמסטר אביב 2005

בעית ALSH
עבור גרפים כלליים

מגישים:

עמית קרפ, 034822544

דנה זילברקלנג, 040521585

תוכן העניינים:

הגדרת הבעיה: 3

פתרונות לבעיה 3

1. אלגוריתם קירוב (שימוש ב-A* Like Algorithm) 3

א. הגדרת האלגוריתם 3

ב. מימוש האלגוריתם ובעיות 3

ג. מסקנות 6

2. אלגוריתם Decision Tree 7

א. הגדרת האלגוריתם 7

ב. מימוש האלגוריתם 7

ג. תוצאות 8

בעיות שגילינו בתוכנה הקודמת, שפעלה עבור עצים בלבד 10

עמידה בתוכנית העבודה 11

הגדרת הבעיה:

מטרה:

לפתור את בעיית ה- Subtree homeomorphism , כלומר:
בהנתן Pattern graph ו- Text graph, למצוא תת עץ של ה- Text graph, שהוא איזומורפי ל- pattern Graph. או להחליט שאין כזה.
בנוסף, מותרת מחיקה של צמתים מדרגה 2 מה- Text graph. תוך הורדת ציון מתאים.

פתרונות לבעיה

1. אלגוריתם קירוב (שימוש ב-A Like Algorithm)*

א. הגדרת האלגוריתם

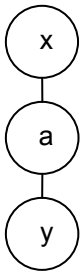
בהינתן שני גרפים G1(V1;E1) ו- G2(V2;E2), ניצור גרף חדש בשם Association Graph, AG(V,E) כאשר V = V1*V2 , E = V*V.
נוסיף קשת ל-AG בין הצמתים (a1,b1) ו-(a2,b2) אמ"מ:

- יש קשת בין a1 ו-a2 בגרף G1 וגם יש קשת בין b1 ו-b2 בגרף G2.

או

- אין קשת בין a1 ו-a2 בגרף G1 וגם אין קשת בין b1 ו-b2 בגרף G2.

הפתרון לבעיה יהיה מציאת הקליק בעל משקל מקסימלי (Maximum weighted clique מוגדר כ- max {W(S) : S is a clique in G}).



<u>התייחסות למחיקות:</u>

כדי לאפשר מחיקות של צמתים מדרגה 2 בגרף ה-text (G2), נוסיף לגרף ה-AG צומת מחיקה (phi,i) אמ"מ i הינה צומת מדרגה 2 ב-G2. כדי לשמור על שלמות הקליק, נוסיף את הקשתות הבאות:
לכל צמת a מדרגה 2 בגרף G2, המחובר לצמתים x ו-y, נוסיף קשת בין (phi,a) לבין (i,x), ו-(j,y) לכל i,j. בנוסף נחבר קשת בין (i,x) לבין (j,y) ואם יש צומת ב-AG המחובר הן (i,x) והן ל- (j,y), נחבר אותו גם כן ל- (phi,a).

ב. מימוש האלגוריתם ובעיות

מימשנו את האלגוריתם במלואו, בהתאם לדרישות שהוצגו לעיל.
בעת מימוש הקוד ובמהלך בדיקתו על קלטים שונים התגלו מספר בעיות:

<u>*בעיה 1*</u>

כאשר מחברים צומת מחיקה (phi, a) לכל הצמתים ב-AG המחוברת הן (i,x) והן ל- (j,y), יתכן שמחברים אותה גם לצומת (i,a), כלומר, יתכן שנתאים את צומת i עם a ובמקביל נמחק את a – מצב לא חוקי.

<u>*פתרון בעיה 1*</u>

הפתרון לבעיה זו הינו שינוי בתהליך הוספת הקשתות ל-AG:
לכל צמת a מדרגה 2 בגרף G2, המחובר לצמתים x ו-y, נוסיף קשת בין (phi,a) לבין (i,x), ו-(j,y) לכל i,j. בנוסף נחבר קשת בין (i,x) לבין (j,y) ואם יש צומת ב-AG המחובר הן (i,x) והן ל- (j,y), נחבר אותו גם כן ל- (phi,a), **אלא אם כן הוא מהסוג (k,a), עבור k כלשהו**.

בצורה זו הקליק המקסימלי שיבחר, יכיל או צומת מחיקה (phi, a), או צומת רגילה (i,a), אך לא את שניהם.

בעיה 2

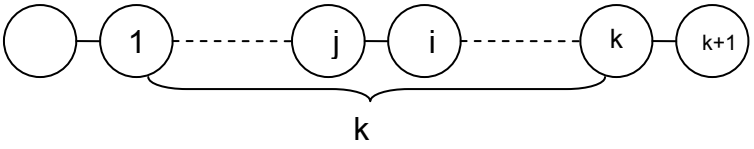
כיצד נאפשר רצף של MAX_DELETION_NUM מחיקות הגדול מאחד אך אינו חורג ממספר זה?

פתרון בעיה 2

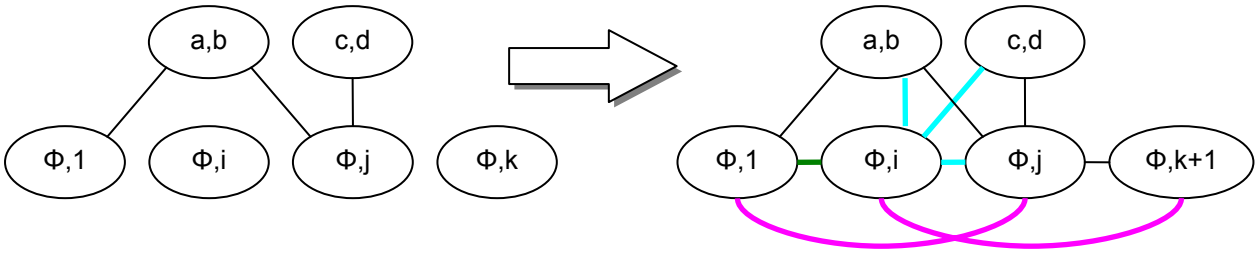
עבור טיפול ברצף של k צמתים מדרגה 2 (MAX_DELETION_NUM <= k), נחבר את צומת המחיקה המשוייך לצומת ה-i ל:

1. צומת המחיקה המשוייך לצומת j הסמוך ל-i.
2. כל הצמתים המחוברים לצומת המחיקה הסמוך לו.
3. מותר לחבר לצמתי מחיקה נוספים, בתנאי שסך הכל יצרנו רצף צמתי מחיקה שאינו גדול מ- k+1.

Text:



AG:

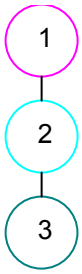


* אין קשת בין (Φ,1) לבין (Φ,k+1) כיוון שרצף המחיקות גדול מ-k.

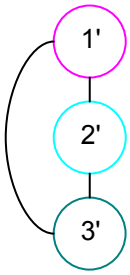
בעיה 3

מעגל הקיים ב-text מול שרוך ב-pattern – לא תהיה התאמה עפ"י האלגוריתם. (בגרף AG הצומת (1,1') לא תחובר ל-(3,3')).

Pattern



Text



פתרון בעיה 3

פתרון לבעיה זו בוצע ע"י הוספת קשתות נוספות ל-AG כדלקמן:
נוסיף קשת ל-AG בין הצמתים (p1,t1) ו-(p2,t2) גם אם אין קשת בין p1 לבין p2 וכן יש קשת בין t1 לבין t2. בצורה זו אנו מאפשרים לאלגוריתם ל"קבל" או "לא לקבל" קשתות ב-text .

בעיה 4

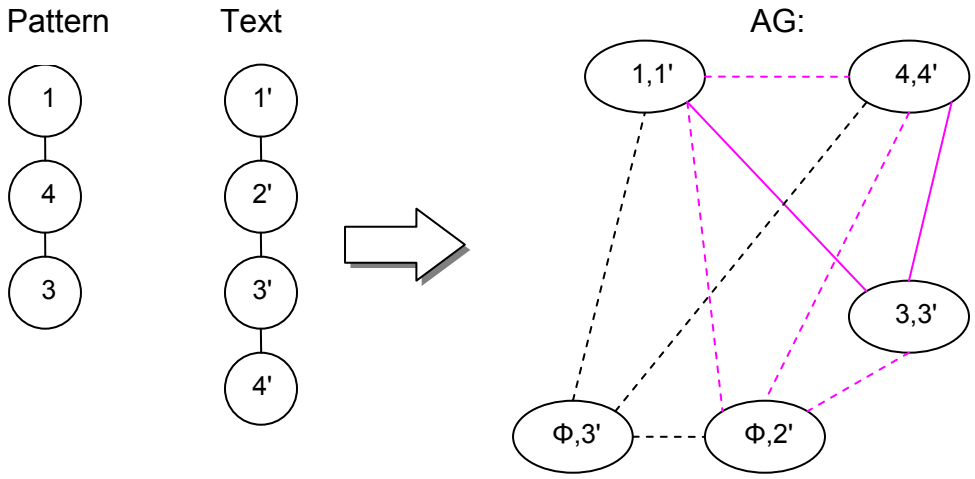
בעקבות בעיה 1 והפתרון שלה, נוצרה בעיה חדשה שנובעת משני המהלכים הבאים:

- לא חיברנו את הצמתים (phi,a) לצמתים מסוג (k,a) (כדי שהקליק המקסימלי שיבחר, לא יכיל גם מחיקה של צומת וגם התאמה שלה).
- לכל צמת מדרגה 2 בגרף G2 המחובר לצמתים 'x ו-'y, נחבר קשת בין ('i,x) לבין ('j,y) לכל הצמתים i,j ב-Pattern.

הבעיה החדשה הינה שיתכן והוספנו קליקים חדשים שלא היו צריכים להיות ב-AG:
הוספנו קשת בין (i,x) לבין (j,y) שמטרתה הינה מחיקה של הצומת a (חיבור (phi,a) לקליק), אך בו זמנית לא חיברנו את (phi,a) לצמתים מסוג (k,a) שהיו מחוברים ל-(i,x) ול-(j,y). לכן יתכן שהוספנו קליק חדש שכולל את הצמתים (i,x), (j,y) ו- (k,a), ואינו כולל כמובן, את (phi,a). הקליק הנ"ל שאינו צריך להיות ב-AG המקורי, יכול להתקבל כקליק האופטימלי.

לפיכך , כאשר "לוקחים" לקליק המקסימלי קשת שהוספנו בין (i,x) לבין (j,y), עלינו "להכריח" את האלגוריתם "לקחת" לקליק את הצומת (phi,a), כיון שללא צומת זה המסמל את מחיקת a, אין לקשת זו מהות.

דוגמא לבעיה זו:
עבור מחיקה של שני הצמתים ,2' ו-3' ברצף, חיברנו את (1,1') עם (4,4') בקשת מחיקה, דבר שיצר קליק חדש שאינו חוקי. ונקבל את ההתאמה: ([1,1'], [3,3'], [4,4']ו- [phi,2'])



פתרון בעיה 4

פתרון לבעיה 3 יכול להתבצע ע"י הצגת "קשת המחיקה" בין (i,x) לבין (j,y) ע"י צומת.
נשתמש במהלכים הבאים:

- נוסיף לגרף AG צומת חדש שיסומן ע"י [(i,x) , (j,y)].
- נחבר צומת זה ל- (phi,a) ולכל הצמתים שהיו בקליק שלה.
- ניקוד הצומת החדש יהיה סכום הניקודים של (i,x) ו- (j,y).
- לא נוסיף את הקשת בין (i,x) לבין (j,y)!

פתרון זה מונע הוספת קליקים חדשים, מכיון שלא הוספנו קשתות ל-AG.
אך הפתרון יוצר שתי בעיות חדשות:

- סיבוכיות האלגוריתם עולה מאוד: ב-AG יש כעת (m*n)^2+n צמתים במקום m*n+n צמתים.
- בעת צורך במחיקת 2 צמתים לא רצופים שיש בינהם צומת, נוצרת בעית ניקוד: ניקוד הצומת שבין שני צמתי המחיקה נחשב פעמיים.

ג. מסקנות

מפאת הקשיים בפתרון בעית ALSH תוך שימוש ב-A* Like Algorith, קשיים הנובעים בעיקר ממימוש מחיקות, הוחלט שבמסגרת הפרוייקט שלנו יוקפא האלגוריתם הנ"ל והפתרון יתרכז באלגוריתם השני (ראה בהמשך).

2. אלגוריתם *Decision Tree*

א. הגדרת האלגוריתם

ניצור "עץ החלטות", כך שכל צומת בעץ תייצג התאמה בין צומת מה-PATTERN לצומת מה-TEXT ותכיל את ציון ההתאמה . מסלול מהשורש לעלה בעץ ייצג התאמה מלאה של כל צמתי ה-PATTERN לצמתים ב-TEXT.
ציון ההתאמה הכולל יהיה סכום ציוני ההתאמה שעל המסלול מהשורש לעלה בעץ.
בכל פעם שנרצה להרחיב את העץ (להתאים בין צמתים מה- PATTERN וה-TEXT), נבדוק שההתאמות שביצענו עד שלב זה מאפשרות את ההתאמה הזו, אחרת לא נבצע אותה.
בנוסף, לא נרחיב את העץ למסלולים שלא יכולים לשפר את ההתאמה הטובה ביותר עד כה.
מחיקות של צמתים מדרגה 2 ב- TEXT אפשריות, עד לאורך שנקבע מראש.

ב. מימוש האלגוריתם

- עץ ההחלטות מתבצע בשיטת DFS, כך שסיבוכיות המקום של האלגוריתם הינה – O(m). בכל שלב שומרים רק מסלול אחד בעץ, כלומר התאמה אחת בין צמתי ה-TEXT וה-PATTERN.
- כאשר לא ניתן לשפר את ההתאמה הטובה ביותר עד כה, לא נמשיך את "המסללול" הנוכחי בעץ.
- הכל שלב, כאשר בודקים התאמה של ן צומת מה-TEXT וצומת מה-PATTERN , עוברים על כל ההתאמות שבוצעו כד כה במסלול בעץ, ובודקים שניתן לבצע את ההתאמה הנ"ל.
- ניתן "להתעלם" מקשתות ב-Text graph , כלומר אם בין שני צמתים ב-PATTERN אין קשת ובין זוג צמתים ב-TEXT יש קשת, התאמה של זוגות הצמתים האלו אחד לשני מותרת. פתרון זה פותר את בעיה 3 שהוצגה ב-AG Algorithm .
- עבור כל צומת מדרגה 2, מאפשרים מחיקה שלה בתנאי שהמסלול בעץ ההתאמות מאפשר זאת.
- כמו כן, ניתן לדלג על עד מספר קבוע מראש של צמתים מדרגה 2 ברצף.
- יעילות הזמן של האלגוריתם בפועל טובה מאוד, גם עבור קלטם גדולים. וזאת מסיבה שרוב המסלולים בעץ ההתאמות אינם מבוצעים כלל. (מפורט בהמשך).
- ניסינו ליעל את האלגוריתם ככל שהצלחנו, תוך המנעות ,ככל שניתן, מהקצאת זיכרון דינמי שמתבצעת פעמים רבות, "וחיתוך" של מסלולים בעץ מוקדם ככל שניתן.

ג. תוצאות

- הרצנו את האלגוריתם על קלטים רבים. הקלטים שבדקנו היו בגדלים שונים, חסרי או בעלי מעגלים, וכאלה שהפתרון מכיל מחיקות. בכל המקרים התוצאות שהתקבלו אכן ייצגו את ההתאמה המקסימלית הנכונה בין ה-text ל-pattern.

- הרצנו את האגוריתם על קלטים ממאגר e.coli שהינם עצים, והשוונו את התוצאות לאלו שהתקבלו מהתוכנה הקיימת לעצים (של אלעד) . ניתן לראות בדוגמת ההרצה הבאה שהתוצאות זהות, למעט:
  1. הבדלים הנובעים מבאגים בתוכנה הקיימת לעצים (ראה בהמשך).
  2. האלגוריתם שלנו לא ישלים התאמה שאינה יכולה לשפר את ההתאמה המירבית עד כה, ולכן לעיתים מופיעות באלגוריתם הקיים לעצים תוצאות שאינן מופיעות באלגוריתם ה-Decision Tree. בכל מקרה ההתאמה המקסימלית עבור כל זוג גרפים תמיד זהה בשני האלגוריתמים.

האלגוריתם הקיים לעצים

Text File Name	Match Score	P-Value
ecoliK_3phenylpropionate_degrad...	-24.6515	0.94
ecoliK_3phenylpropionate_degrad...	-24.6515	0.94
ecoliK_3phenylpropionate_degrad...	-24.6515	0.94
ecoliK_3phenylpropionate_degrad...	-24.6515	0.94
ecoliK_3phenylpropionate_degrad...	-24.6515	0.94
ecoliK_D_galactarate_degradation....	-24.6515	0.94
ecoliK_D_galactarate_degradation....	-23.4135	0.65
ecoliK_D_galactarate_degradation....	-25.4135	0.98
ecoliK_D_galactarate_degradation....	-23.4135	0.65
ecoliK_D_galactarate_degradation....	-24.6515	0.94
ecoliK_D_galacturonate_degradati...	-24.6515	0.94
ecoliK_D_galacturonate_degradati...	-23.4135	0.65
ecoliK_D_galacturonate_degradati...	-25.4135	0.98
ecoliK_D_galacturonate_degradati...	-23.4135	0.65
ecoliK_D_galacturonate_degradati...	-24.6515	0.94
ecoliK_D_galucarate_degradation....	-24.6515	0.94
ecoliK_D_galucarate_degradation....	-23.4135	0.65
ecoliK_D_galucarate_degradation....	-25.4135	0.98
ecoliK_D_galucarate_degradation....	-23.4135	0.65
ecoliK_D_galucarate_degradation....	-24.6515	0.94
ecoliK_GDP_mannose_matabolis...	-23.4135	0.65
ecoliK_GDP_mannose_matabolis...	-24.1755	0.94
ecoliK_GDP_mannose_matabolis...	-23.4135	0.65
ecoliK_GDP_mannose_matabolis...	-23.4135	0.65
ecoliK_GDP_mannose_matabolis...	-23.4135	0.65
ecoliK_KDO_biosynth.grp	-23.4135	0.65
ecoliK_KDO_biosynth.grp	-22.1755	0.5
ecoliK_KDO_biosynth.grp	-22.1755	0.5
ecoliK_KDO_biosynth.grp	-22.1755	0.5
ecoliK_KDO_biosynth.grp	-22.1755	0.5
ecoliK_L_arabinose_degradation.g...	-24.6515	0.94
ecoliK_L_arabinose_degradation.g...	-24.6515	0.94
ecoliK_L_idionate_degradation.grp	-23.4135	0.65
ecoliK_L_idionate_degradation.grp	-23.4135	0.65
ecoliK_NAD_dephosphorylation.grp	-24.6515	0.94
ecoliK_NAD_dephosphorylation.grp	-20.711	0.48
ecoliK_NAD_dephosphorylation.grp	-22.711	0.65

Decision Tree

Match Score	P-Value
-24.6515	0.94
-24.6515	0.94
-24.6515	0.94
-24.6515	0.94
-24.6515	0.94
-24.6515	0.94
-23.4135	0.65
-25.4135	0.98
-23.4135	0.65
-24.6515	0.94
-24.6515	0.94
-23.4135	0.65
-25.4135	0.98
-23.4135	0.65
-24.6515	0.94
-24.6515	0.94
-23.4135	0.65
-25.4135	0.98
-23.4135	0.65
-24.6515	0.94
-23.4135	0.65
-24.1755	0.94
-23.4135	0.65
-23.4135	0.65
-23.4135	0.65
-23.4135	0.65
-22.1755	0.5
-22.1755	0.5
-22.1755	0.5
-22.1755	0.5
-24.6515	0.94
-24.6515	0.94
-23.4135	0.65
-23.4135	0.65
-24.6515	0.94
-20.711	0.48
-22.711	0.65

- זמני ריצת האלגורתים מהירים למדי, ואף מהירים יותר מזמן ריצת אלגוריתם העצים. אנו מדדנו זמנים לקלטים עד גודל של 20 צמתים.

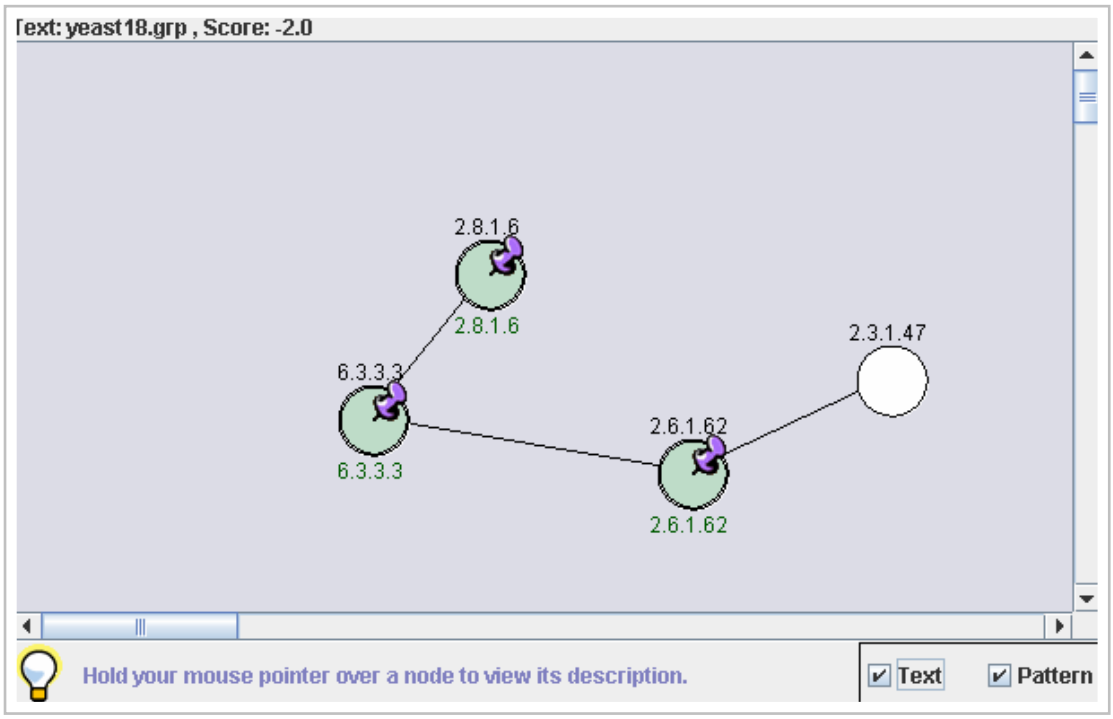
- הרצנו את האלגוריתם על 6400 קלטים (שהינם גרפים) שונים שנלקחו ממאגר השמרים (80 גרפים שונים שהרצנו זה עם זה). התקבלו זמני ריצה נמוכים ביותר. להלן פירוט חלק מהתוצאות, המכיל את זמני הריצה, גדול הקלטים, ואת חמשת ההתאמות המקסימליות עבור כל ריצה . ל**הלן התוצאות עם זמני הריצה הגרועים ביותר שהתקבלו**. הטבלה המלאה מצויה בקובץ המצורף DTreeAlgResults.xcl.

Text File:	Pattern File:	Text Nodes Num	Pattern Nodes Num	Time [sec]	score1	score2	score3	score4	score5
yeast64.grp	yeast57.grp	20	14	**0.344**					
yeast64.grp	yeast65.grp	20	17	**0.344**					
yeast64.grp	yeast64.grp	20	20	**0.297**	0				
yeast64.grp	yeast17.grp	20	15	**0.281**					
yeast64.grp	yeast79.grp	20	10	**0.25**	-4.143	0	-6.917	-6.917	-6.917
yeast65.grp	yeast79.grp	17	10	**0.25**	-63.017	-61.779	-63.779	-63.017	-65.017
yeast64.grp	yeast77.grp	20	9	**0.219**					
yeast64.grp	yeast31.grp	20	10	**0.203**	-74.517	-74.116	-74.517	-74.116	-74.517
yeast64.grp	yeast80.grp	20	7	**0.203**					
yeast65.grp	yeast31.grp	17	10	**0.203**	-75.66	-75.66	-75.66	-75.66	-75.66
yeast65.grp	yeast65.grp	17	17	**0.187**	0				
yeast57.grp	yeast79.grp	14	10	**0.172**					
yeast64.grp	yeast9.grp	20	9	**0.172**	-65.714	-65.714	-65.714	-65.714	-66.952
yeast64.grp	yeast56.grp	20	6	**0.172**					
yeast65.grp	yeast17.grp	17	15	**0.172**					
yeast65.grp	yeast64.grp	17	20	**0.172**					
yeast65.grp	yeast80.grp	17	7	**0.172**					
yeast65.grp	yeast77.grp	17	9	**0.171**					
yeast65.grp	yeast2.grp	17	9	**0.141**	-62.1	-62.1	-62.1	-62.1	-62.1
yeast65.grp	yeast56.grp	17	6	**0.141**	-46.11	-46.11	-46.11	-46.11	-46.11
yeast17.grp	yeast79.grp	15	10	**0.14**					
yeast64.grp	yeast1.grp	20	9	**0.14**	-24.908	-67.613	-30.417	-29.176	-29.176
yeast17.grp	yeast65.grp	15	17	**0.125**					
yeast42.grp	yeast52.grp	6	3	**0.125**	-24.652	-24.652	-24.652	-24.652	-24.652
yeast64.grp	yeast2.grp	20	9	**0.125**					
yeast64.grp	yeast30.grp	20	13	**0.125**					
yeast64.grp	yeast39.grp	20	9	**0.125**	-63.227	-67.199	-68.421	-68.421	-67.199
yeast65.grp	yeast9.grp	17	9	**0.125**	-71.503	-71.503	-72.202	-71.503	-71.503
yeast65.grp	yeast30.grp	17	13	**0.125**					
yeast65.grp	yeast39.grp	17	9	**0.125**	-66.421	-65.776	-67.014	-65.776	-66.421
yeast14.grp	yeast11.grp	1	1	**0.11**	-8.217				
yeast17.grp	yeast64.grp	15	20	**0.11**					
yeast21.grp	yeast59.grp	6	1	**0.11**	-8.217	-8.217	-8.217	-8.217	-8.217
yeast28.grp	yeast75.grp	3	1	**0.11**	-8.217	-8.217	-8.217		
yeast30.grp	yeast1.grp	13	9	**0.11**					
yeast32.grp	yeast10.grp	2	2	**0.11**	-15.196	-15.196			
yeast44.grp	yeast72.grp	1	2	**0.11**					
yeast46.grp	yeast32.grp	5	2	**0.11**	-13.317	-15.196	-16.434	-15.317	-16.434

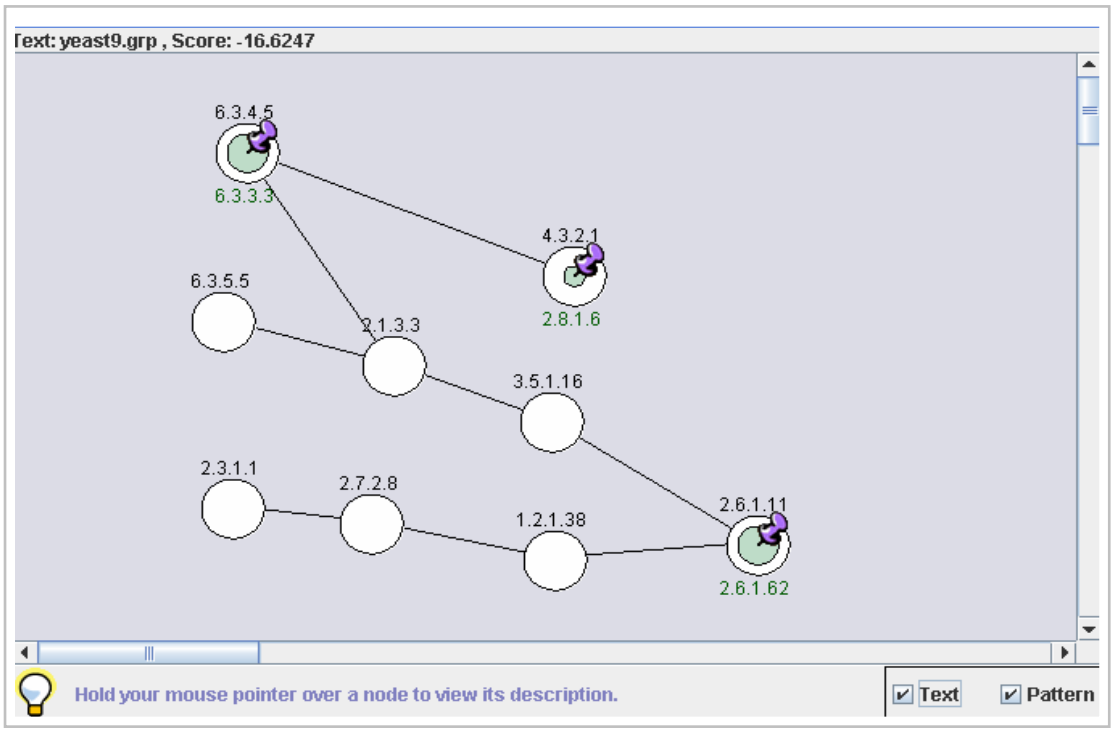
בעיות שגילינו בתוכנה הקודמת, שפעלה עבור עצים בלבד

במהלך השוואת האלגוריתם לאלגוריתם הקיים עבור עצים, התגלו שני באגים בתוכנה הקיימת:

1. בהרצת שני הקלטים הבאים: Pattern: temp_p.grp Text: Yeast18.grp (מאגר השמרים)
התקבל deletion score של (-2) למרות שלא התבצעה כלל מחיקה:



2. בהרצת שני הקלטים הבאים: Pattern: temp_p.grp Text: Yeast9.grp (מאגר השמרים)
יש מחיקה של צומת (2.1.3.3) למרות שדרגתה גדולה מ-2.



עמידה בתוכנית העבודה

Subject	completed
Learning the existing user interface and the existing code	√
Implementing Modified AG Algorithm	X/ √ *
Implementing Maximum weighted Clique Search Algorithm	√
Combining the two above algorithms	√
Implementing Decision Tree Algorithm	√
Linking the algorithms to the user interface, and making change in the interface if needed	√
Testing the two a algorithms on the same data base and comparing results	√
Testing the two a algorithms against the existing tree algorithm and comparing results	√
First draft submission and oral presentation	√
Algorithms improvement	√
Documentation	√
Final submission	√

*A reasonable solution for all the problems which acquired during the implementation couldn't be found

<u>פעילות מעבר להיקף הפרוייקט:</u>

- מציאת פתרונות לקשיים הרבים שנוצרו במסגרת מימוש אלגוריתם הקירוב.
- שינוים בתוכנת ה-GUI לשם שיפור הנוחיות.