GRUČENJE

Dano imamo naslednjo matriko izražanja:

	T_1	T_2	T_3	T_4	T_5
g_1	0	9	7	3	12
g_2	10	8	1	0	4
g_3	3	0	10	1	0
g_4	1	12	8	2	10
g_5	1	8	8	1	13
g_6	12	10	0	2	2

1. Izračunajte gruče z uporabo metode voditeljev za $k = 3$. št. voditeljev

Naključno izberemo začetne voditelje: $X = \{g_1, g_3, g_5\}$

1. Poračunamo razdalje genov do voditeljev (evklidska metrika):

X	g_1	g_3	g_5
$D(g_1, X)$	0	247	8
$D(g_2, X)$	210	211	212
$D(g_3, X)$	247	0	241
$D(g_4, X)$	16	253	25
$D(g_5, X)$	8	241	0
$D(g_6, X)$	295	286	311

Primer: $D(g_1, g_2) = \sqrt{(0-10)^2 + (9-8)^2 + (7-1)^2 + (3-0)^2 + (12-4)^2} = 210$

2. V vsaki vrstici izberemo min vrednost.
Iz vsakega stolpca odčitamo izbrane gene ⇒ nove gruče.

$C_1 = \{g_1, g_2, g_4\}$
$C_2 = \{g_3, g_6\}$ ⇒ $X = \{C_1, C_2, C_3\}$
$C_3 = \{g_5\}$

3. Najdemo vrednosti za nove voditelje (aritmetična sredina vrednosti genov iz matrike izražanja):

$$C_1 = \left(\frac{0+10+1}{3}, \frac{9+8+12}{3}, \frac{7+1+8}{3}, \frac{3+0+2}{3}, \frac{12+4+10}{3}\right) = (3.67,\ 9.67,\ 5.33,\ 1.67,\ 8.67)$$

$C_2 = (7.5,\ 5,\ 5,\ 1.5,\ 1)$

$C_3 = g_5$

Postopek ponavljamo dokler ne dosežemo konvergence:
- gruče se nehajo spreminjati (npr. naslednjo iteracijo dobimo enake gruče)
- razlike med starimi in novimi voditelji postanejo zanemarljive

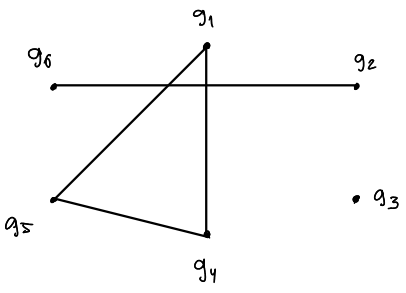
2. Izračunajte gruče z uporabo metode "poškodovanih" klik za
   (a) $\theta = 10$
   (b) $\theta = 15,5$

Osnova za metodo je matrika razdalj.
Uporabimo evklidsko razdaljo.
Razdalje moraš sam izračunat iz matrike izražanja.

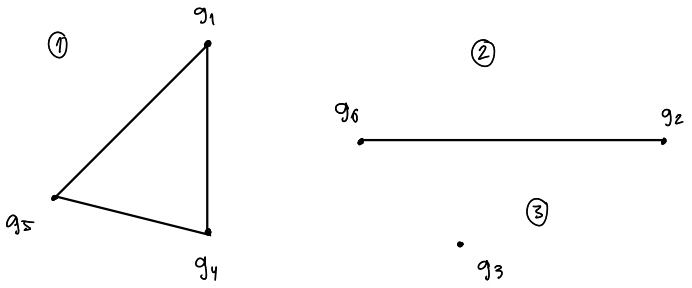
	g_1	g_2	g_3	g_4	g_5	g_6
g_1	0	14.49	15.72	4	2.83	17.18
g_2		0	14.53	13.64	14.56	4.12
g_3			0	15.91	15.52	16.91
g_4				0	5.10	15.91
g_5					0	17.64
g_6						0

a) $\theta = 10$ ⇒ med genoma imamo povezavo, če je razdalja med njima manjša od θ.

1. Imamo začetni graf:

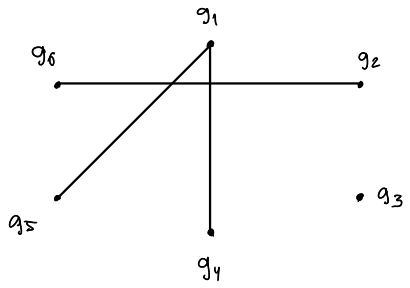


2. Iz grafa odčitamo klične grafe (polne podgrafe):



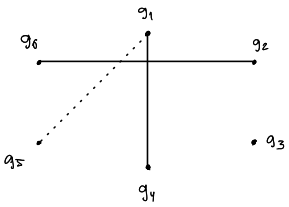
⇒ To so naše gruče: $C_1 = \{g_1, g_4, g_5\}$
$C_2 = \{g_2, g_6\}$
$C_3 = \{g_3\}$

Če bi imeli npr. graf:

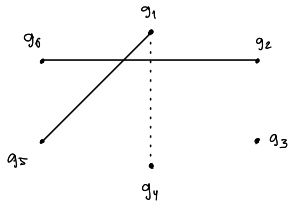


⇒

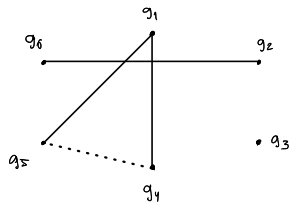
ne bi morali razvrstiti vseh v gruče. V tem primeru lahko dodamo ali odstranimo eno povezavo, npr.:



$C_1 = \{g_2, g_6\}$
$C_2 = \{g_1, g_4\}$
$C_3 = \{g_3\}$
$C_4 = \{g_5\}$



$C_1 = \{g_1, g_5\}$
$C_2 = \{g_2, g_6\}$
$C_3 = \{g_3\}$
$C_4 = \{g_4\}$

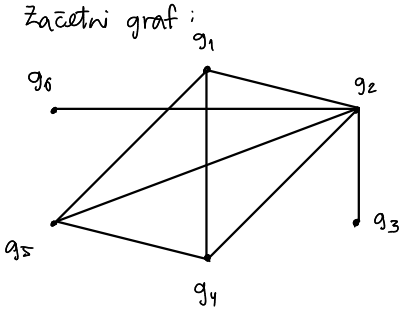


$C_1 = \{g_1, g_4, g_5\}$
$C_2 = \{g_2, g_6\}$
$C_3 = \{g_3\}$

Vsi ti različni nabori gruč so enakovredni.

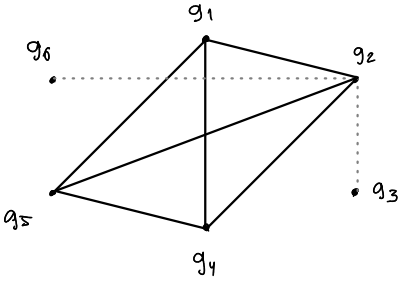
b) $\Theta = 15.5$

Začetni graf:



⇒

Tega ne moremo rešit samo z eno povezavo, ker bo graf še vedno povezan. Zato izbrišemo DVE:



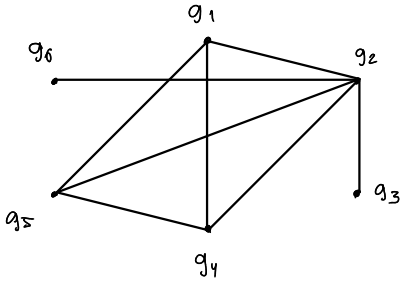
⇒

$C_1 = \{g_1, g_2, g_4, g_5\}$
$C_2 = \{g_3\}$
$C_3 = \{g_6\}$

V najslabšem primeru moramo izbrisati vse povezave.

3. Izračunajte gruče z uporabo hevristike CAST za $\theta = 15$.

Osnova je matrika razdalj in isti graf kot prej v (b):



1. Izberemo točko z najvišjo stopnjo (ta z največ povezavami): g_2 (5)
Ona gre prva v našo gručo: $C_1 = \{g_2\}$.
Gradimo gručo okoli točke.

2. Najdemo ji najbližjo točko: g_6
Dodamo jo v gručo: $C_1 = \{g_2, g_6\}$

3. Izberemo točke najbližje gruči:

$D(g_1, C_1) = (d(g_1, g_2) + d(g_1, g_6))/2 = 15.84$
$D(g_3, C_1) = 15.72$
$D(g_4, C_1) = 14.78$
$D(g_5, C_1) = 16.1$

⇒ s pomočjo Θ določimo ali je točka bližnja: ali velja $d(g_i, C) < \Theta$
če jih je več, dodamo najmanjšo vrednost:

$C_1 = \{g_2, g_6, g_4\}$

Ponavljamo, dokler nimamo nič več za dodat.
Dobimo:

$C_1 = \{g_2, g_6, g_4, g_5, g_1\}$

4. Ali imamo kaj za odstranit iz gruče? Gledamo razdalje gerov znotraj gruče do gruče:

$D(g_1, C_1) = 7.7$
$D(g_2, C_1) = 9.36$
$D(g_5, C_1) = 7.75$
$D(g_4, C_1) = 8.02$
$D(g_6, C_1) = 10.97$

⇒ Nič za odstranit ⇒ našli smo prvo gručo, ki je particija.

5. Iz grafa odstranimo točke v prvi gruči, ponovimo vse. Ostane samo g_3, zato $C_2 = \{g_3\}$.

Končni gruči:
$C_1 = \{g_1, g_2, g_4, g_5, g_6\}$
$C_2 = \{g_3\}$

4. Izračunajte drevo, ki ponazarja hierarhično gručenje (dendrogram) z uporabo UPGMA.

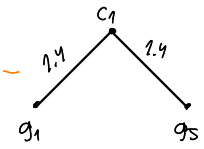
Osnova je matrika razdalj.
1. Najdemo min razdaljo
2. Združimo gena
3. Odstranimo gena iz matrike
4. Dodamo gručo v matriko
5. Ponavljamo dokler ne zmanjka genov v matriki.

① 1.

	g_1	g_2	g_3	g_4	g_5	g_6
g_1	0	14.49	15.72	4	2.83	17.18
g_2		0	14.53	13.64	14.56	4.12
g_3			0	15.91	15.52	16.91
g_4				0	5.10	15.91
g_5					0	17.64
g_6						0

2. $C_1 = \{g_1, g_5\}$

razdalja med genoma / 2



3, 4.

	C_1	g_2	g_3	g_4	g_6
C_1	0	14.52	15.62	4.55	17.41
g_2		0	14.53	13.64	4.12
g_3			0	15.91	16.91
g_4				0	15.91
g_6					0

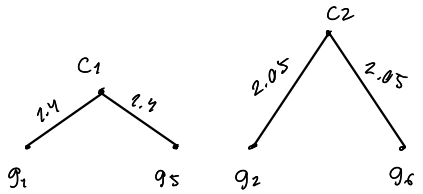
② 1.

	C_1	g_2	g_3	g_4	g_6
C_1	0	14.52	15.62	4.55	17.41
g_2		0	14.53	13.64	4.12
g_3			0	15.91	16.91
g_4				0	15.91
g_6					0

3, 4.

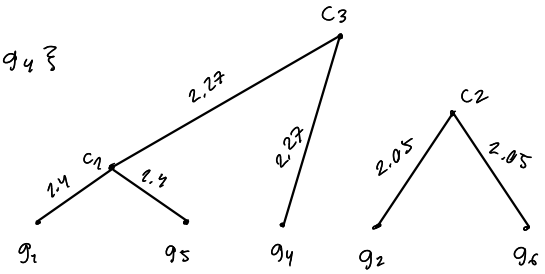
	C_1	C_2	g_3	g_4
C_1	0	15.97	15.62	4.55
C_2		0	15.72	14.78
g_3			0	15.91
g_4				0

2. $C_2 = \{g_2, g_6\}$

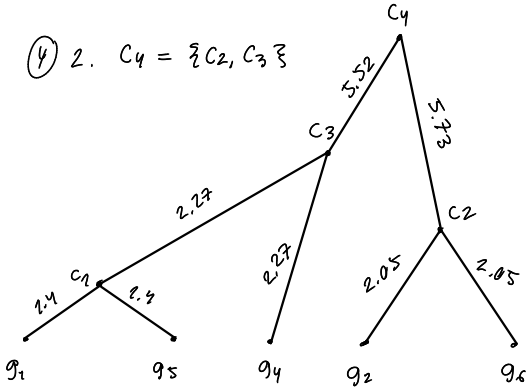


③ 1, 2. $C_3 = \{C_1, g_4\}$

Razdalja med C_3 in C_1 je enaka razdalji med C_3 in g_4, torej je $d(C_3, g_4) = d(C_1, g_4)/2$



④ 2. $C_4 = \{C_2, C_3\}$



3, 4.

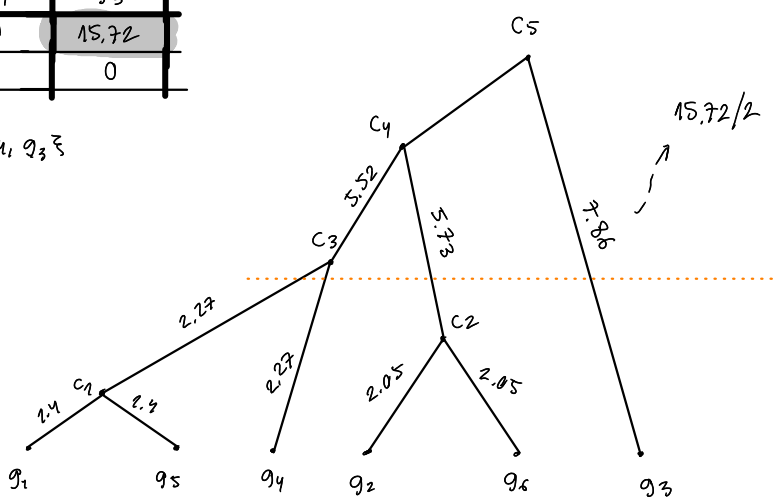
	C_3	C_2	g_3
C_3	0	15.57	15.72
C_2		0	15.72
g_3			0

Ampak I guess odštejem še 2.27 na levi strani in 2.05 na desni?...
Prej sem imel $d(C_1, g_4)$, zdaj pa $d(C_3, C_2)$ ⇒ dve gruči ⇒ popravit (?)

⑤

	C_4	g_3
C_4	0	15,72
g_3		0

1, 2, 3; $C_5 = \{C_4, g_3\}$



Odrisno kje prerežemo drevo, tam odčitamo gruče. Npr. tu dobimo:

$\{g_1, g_5\}$
$\{g_4\}$
$\{g_2, g_6\}$
$\{g_3\}$