Algoritmi v bioinformatiki - 2. Domača naloga

Jan Panjan

May 9, 2025

1. *Dano imamo naslednje zaporedje izidov metov kovanca*

$$V = CCCGCGGCCGC$$

pri čemer C označuje, da je bil izid meta cifra, G pa da je bil izid meta grb. Za mete imamo na voljo 3 kovance, A, B in C, veljajo naslednje verjetnosti:

Prehod:

%	A	B	C
A	40	30	30
B	30	40	30
C	30	30	40

Izpis:

%	C	G
A	75	25
B	80	20
C	20	80

Katera od možnosti je najbolj verjetna?

(a) *za vse mete smo uporabili kovanec A*

(b) *za vse mete smo uporabili kovanec C*

(c) *za vse mete smo uporabili kovanec B*

(d) $\Pi = AAACBCCBBCA$

Odgovor ustrezno utemeljite.

Za vse mete smo uporabili kovanec A

$(0.75)^7$	7-krat vržemo C z verjetnostjo 0.75
$(0.25)^2$	7-krat vržemo G z verjetnostjo 0.75
$(0.4)^{10}$	10-krat ne zamenjamo kovanca A z verjetnostjo 0.4

$$p(A) = (0.75)^7 \cdot (0.25)^4 \cdot (0.4)^{10} = 0.00209$$

Za vse mete smo uporabili kovanec B

$(0.8)^7$	7-krat vržemo C z verjetnostjo 0.8
$(0.2)^2$	7-krat vržemo G z verjetnostjo 0.2
$(0.4)^{10}$	10-krat ne zamenjamo kovanca B z verjetnostjo 0.4

$$p(B) = (0.8)^7 \cdot (0.2)^4 \cdot (0.4)^{10} = 0.00134$$

Za vse mete smo uporabili kovanec C

$(0.2)^7$	7-krat vržemo C z verjetnostjo 0.8
$(0.8)^2$	7-krat vržemo G z verjetnostjo 0.2
$(0.4)^{10}$	10-krat ne zamenjamo kovanca C z verjetnostjo 0.4

$$p(C) = (0.2)^7 \cdot (0.8)^4 \cdot (0.4)^{10} = 0.0000209$$

Π = AAACBCCBBCA

$(0.3)^6$	6-krat ostanemo v istem kovancu (vsi kovanci imajo enake verjetnosti)
$(0.4)^4$	4-krat zamenjamo kovanec (tudi tu imajo enako verjetnosti)
$(0.75)^4$	4-krat vržemo kovanec A, vsakič vržemo cifro z verjetnostjo 0.75
$(0.8)^3$	3-krat vržemo kovanec B, vsakič vržemo cifro z verjetnostjo 0.8
$(0.8)^4$	4-krat vržemo kovanec C, vsakič vržemo grb z verjetnostjo 0.8

$$p(\Pi) = (0.3)^6 \cdot (0.4)^4 \cdot (0.75)^4 \cdot (0.8)^3 \cdot (0.8)^4 = 0.00000124$$

Rešitev: Najbolj verjetna je možnost z največjo verjetnostjo. To je možnost (a) z verjetnostjo 0.00209.

2. *Dani imamo zaporedji $s = GAGTACA$ in $t = TGATTACA$ ter vrednostno funkcijo s parametroma $\mu = 4, \sigma = 2$ in nagrado za ujemanje 2.*

   (a) *Z uporabo Needleman-Wunsch-evega algoritma za globalno poravnavo smo dobili naslednjo tabelo:*

		-	G	A	G	T	A	C	A
		0	1	2	3	4	5	6	7
-	0	0	-2	-4	-6	-8	-10	-12	-14
T	1	-2	-4	-6	-8	-4	-6	-8	-10
G	2	-4	0	-2	-4	-6	-8	-10	-12
A	3	-6	-2	2	0	-2	-4	-6	-8
T	4	-8	-4	0	-2	2	0	-2	-4
T	5	-10	-6	-2	-4	0	-2	-4	-6
A	6	-12	-8	-4	-6	-2	2	0	-2
C	7								
A	8								

   Dopolnite tabelo tako, da poračunate vrednosti (in ustrezne puščice) za zadnji dve vrstici.

		–	G	A	G	T	A	C	A
		0	1	2	3	4	5	6	7
–	0	0	-2	-4	-6	-8	-10	-12	-14
T	1	-2	-4	-6	-8	-4	-6	-8	-10
G	2	-4	0	-2	-4	-6	-8	-10	-12
A	3	-6	-2	2	0	-2	-4	-6	-8
T	4	-8	-4	0	-2	2	0	-2	-4
T	5	-10	-6	-2	-4	0	-2	-4	-6
A	6	-12	-8	-4	-6	-2	2	0	-2
C	7	-14	-10	-6	-8	-4	0	4	2
A	8	-16	-12	-8	-10	-6	-2	2	6

   (b) *Koliko optimalnih globalnih poravnav dobite? Izpišite vse rešitve.*

Dobim dve optimalni globalni poravnavi. Mesto vrzeli se spremeni in sicer iz mesta 4 na mesto 3 (in obratno):

s	-	G	A	G	T	-	A	C	A
t	T	G	A	-	T	T	A	C	A

s	-	G	A	G	-	T	A	C	A
t	T	G	A	-	T	T	A	C	A

Z matrikami to izgleda tako:

		–	G	A	G	T	A	C	A
		0	1	2	3	4	5	6	7
–	0	0	-2	-4	-6	-8	-10	-12	-14
T	1	-2	-4	-6	-8	-4	-6	-8	-10
G	2	-4	0	-2	-4	-6	-8	-10	-12
A	3	-6	-2	2	0	-2	-4	-6	-8
T	4	-8	-4	0	-2	2	0	-2	-4
T	5	-10	-6	-2	-4	0	-2	-4	-6
A	6	-12	-8	-4	-6	-2	2	0	-2
C	7	-14	-10	-6	-8	-4	0	4	2
A	8	-16	-12	-8	-10	-6	-2	2	6

		–	G	A	G	T	A	C	A
		0	1	2	3	4	5	6	7
–	0	0	-2	-4	-6	-8	-10	-12	-14
T	1	-2	-4	-6	-8	-4	-6	-8	-10
G	2	-4	0	-2	-4	-6	-8	-10	-12
A	3	-6	-2	2	0	-2	-4	-6	-8
T	4	-8	-4	0	-2	2	0	-2	-4
T	5	-10	-6	-2	-4	0	-2	-4	-6
A	6	-12	-8	-4	-6	-2	2	0	-2
C	7	-14	-10	-6	-8	-4	0	4	2
A	8	-16	-12	-8	-10	-6	-2	2	6

3. *Dano imamo naslednjo matriko izražanja:*

	T_1	T_2	T_3	T_4	T_5	T_6
g_1	2	2	6	2	3	4
g_2	3	7	3	1	9	3
g_3	2	2	7	2	6	3
g_4	3	2	3	2	1	3
g_5	2	1	5	1	0	4
g_6	3	5	5	8	2	3
g_7	1	3	1	5	4	2
g_8	5	4	2	4	7	5

Določite gruče z uporabo metode voditeljev, če je začetna množica voditeljev enaka $X = \{g_1, g_5, g_6\}$. Vsak gen lahko obravnavamo kot vektor $g_i = (T_1, \ldots, T_6)$.

Prva iteracija

Prvi korak Najprej je potrebno izračunati razdalje med geni. (Evklidska) razdalja med vsakim genom je definirana kot:

$$d(g_i, g_j) = \sqrt{\left(T_i^{(1)} - T_j^{(1)}\right)^2 + \cdots + \left(T_i^{(6)} - T_j^{(6)}\right)^2} \quad ; \quad 1 \leq i, j \leq 8 \tag{1}$$

Primer za prvi gen:

$$d(g_1, g_1) = \sqrt{(2-2)^2 + (2-2)^2 + (6-6)^2 + (2-2)^2 + (3-3)^2 + (4-4)^2} = 0$$

Očitno je razdalja med istim genom 0, kar pravi tudi prva lastnost metrike: $d(x, y) = 0 \iff x = y$. Ko poračunamo vse, dobimo matriko razdalj. Potrebujemo razdalje samo do voditeljev:

	$d(g_1, g_i)$	$d(g_5, g_i)$	$d(g_6, g_i)$
g_1	0	18.166	17.493
g_2	8.544	11.091	10.296
g_3	3.317	6.557	8.124
g_4	3.873	3.000	7.071
g_5	18.166	0	10.198
g_6	17.493	10.198	0
g_7	5.657	7.071	5.568
g_8	7.071	9.274	7.681

Drugi korak Za vsak gen izberemo najkrajšo razdaljo med njim in voditeljem.

	$d(g_1, g_i)$	$d(g_5, g_i)$	$d(g_6, g_i)$
g_1	0	18.166	17.493
g_2	8.544	11.091	10.296
g_3	3.317	6.557	8.124
g_4	3.873	3.000	7.071
g_5	18.166	0	10.198
g_6	17.493	10.198	0
g_7	5.657	7.071	5.568
g_8	7.071	9.274	7.681

Tretji korak Iz vsakega stolpca odčitamo nove voditelje (vrednosti označene z rdečo), katere označimo s $C_i, i \in \mathbb{N}$, in sicer:

$$\begin{aligned} C_1 &= \{g_1, g_2, g_3, g_8\} \\ C_2 &= \{g_4, g_5\} \\ C_3 &= \{g_6, g_7\} \end{aligned}$$

Četrti korak Za gručo C_i z n geni $\{g_1, \ldots, g_k \mid 1 \leq k \leq 8\}$, izračunamo nov vektor vrednosti $v_i = (v_{i1}, \ldots, v_{in})$ z enačbo:

$$v_i = \frac{1}{n} \sum_{j=1}^{n} g_k \tag{2}$$

Nove vrednosti so torej aritmetična sredina vseh genov v gruči:

$$\begin{aligned} v_1 &= (3, 3.75, 4.5, 2.25, 6.25, 3.75) \\ v_2 &= (2.5, 1.5, 4, 1.5, 0.5, 3.5) \\ v_3 &= (2, 4, 3.5, 6.5, 3, 2.5) \end{aligned}$$

Zdaj ponovimo korake dokler ne dosežemo **konvergence**:

- ko se gruče med iteracijama ne spremenijo
- ko postanejo razlike med radaljami gruč manjše od neke vnaprej določene vrednosti.

Druga iteracija

Prvi + drugi korak

	$d(v_1, g_i)$	$d(v_2, g_i)$	$d(v_3, g_i)$
g_1	3.808	3.354	5.723
g_2	4.743	10.209	8.761
g_3	3.317	6.344	6.764
g_4	5.788	1.500	5.454
g_5	7.036	1.500	7.263
g_6	7.314	7.632	2.784
g_7	5.148	5.958	2.784
g_8	3.937	8.185	6.305

Tretji korak

$$\begin{aligned} C_4 &= \{g_1, g_2, g_3, g_8\} \\ C_5 &= \{g_4, g_5\} \\ C_6 &= \{g_6, g_7\} \end{aligned}$$

Ker so gruče enake kot v prejšnji iteraciji, lahko postopek tu končamo...

Rešitev: gruče določene z metodo voditeljev s $k = 3$ za dano matriko izražanja so

$$\begin{aligned} &\{g_1, g_2, g_3, g_8\} \\ &\{g_4, g_5\} \\ &\{g_6, g_7\} \end{aligned}$$

4. *Izračunajte drevo hierarhičnega gručenja z uporabo algoritma UPGMA.*

Naj bo množica vseh genov $G = \{g_1, \ldots, g_8\}$. Osnova za algoritem UPGMA je matrika razdalj genov, za katero uporabimo sledečo enačbo

$$d_{\mathrm{avg}}(C, C^*) = \frac{1}{|C||C^*|} \sum_{x \in C, y \in C^*} d(x, y) \tag{3}$$

kjer sta C in C^* dve gruči (na začetku so to geni). d je ista kot v (1).

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
g_1	0	8.544	3.317	3.873	3.464	7.000	5.657	7.071
g_2		0	8.124	10.198	11.091	10.296	10.677	8.602
g_3			0	6.325	6.557	8.124	6.403	7.000
g_4				0	3.000	7.071	5.099	6.403
g_5					0	10.198	7.071	9.274
g_6						0	5.568	7.681
g_7							0	5.916
g_8								0

Gručenje deluje tako, da vsako iteracijo izberemo najbližja gena in ju združimo v gručo. Na začetku je vsak gen v svoji gruči, do konca postopka pa ustvarimo eno celovito gručo, ki bo vsebovala vse gene.

Prva iteracija

Prvi korak Izberemo najmanjšo vrednost med razdaljami.

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
g_1	0	8.544	3.317	3.873	3.464	7.000	5.657	7.071
g_2		0	8.124	10.198	11.091	10.296	10.677	8.602
g_3			0	6.325	6.557	8.124	6.403	7.000
g_4				0	3.000	7.071	5.099	6.403
g_5					0	10.198	7.071	9.274
g_6						0	5.568	7.681
g_7							0	5.916
g_8								0

Drugi korak Dodamo gena v gručo C_1, torej $C_1 = \{g_4, g_5\}$ in poračunamo novi vektor v_1, ki bo predstavljal novo gručo v matriki razdalj (tako kot prej uporabimo aritmetično sredino komponent vrednosti iz matrike izražanja):

$$v_1 = (2.5,\ 1.5,\ 4,\ 1.5,\ 0.5,\ 3.5)$$

Tretji korak Gena v gruči združimo, tako da je $G = \{g_1, \ldots, C_1, \ldots, g_8\}$. Izračunamo razdaljo gruče (C_1) do vseh ostalih gruč s pomočjo enačbe (3).

	g_1	g_2	g_3	C_1	g_6	g_7	g_8
g_1	0	8.544	3.317	3.669	7.000	5.657	7.071
g_2		0	8.124	10.645	10.296	10.677	8.602
g_3			0	6.441	8.124	6.403	7.000
C_1				0	8.635	6.085	7.839
g_6					0	5.568	7.681
g_7						0	5.916
g_8							0

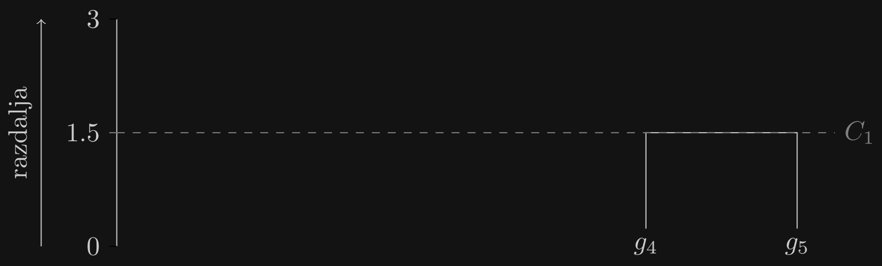
Četrti korak Gručo povežemo na dendrogramu na višini, ki jo izračunamo z enačbo:

$$h(C) = \frac{D(C_1, C_2)}{2} \tag{4}$$

Povezavi (C_1, g_4) dodelimo višino $h(C_1) - h(g_4) = h(C_1)$ ter povezavi (C_1, g_5) višino $h(C_1) - h(g_5) = h(C_1)$.

$$h(C_1) = \frac{D(g_4, g_5)}{2} = \frac{\frac{1}{|g_4||g_5|} d(g_4, g_5)}{2} = \frac{d(g_4, g_5)}{2} = \frac{3.000}{2} = 1.500$$

Dendrogramu dodamo vozlišče:



Ponavljamo vse korake, dokler obstaja več kot ena gruča.

Druga iteracija

Prvi korak

	g_1	g_2	g_3	C_1	g_6	g_7	g_8
g_1	0	8.544	3.317	3.669	7.000	5.657	7.071
g_2		0	8.124	10.645	10.296	10.677	8.602
g_3			0	6.441	8.124	6.403	7.000
C_1				0	8.635	6.085	7.839
g_6					0	5.568	7.681
g_7						0	5.916
g_8							0

Drugi korak Ustvarimo novo gručo $C_2 = g_1, g_3$ in poračunamo vektor v_2:

$$v_2 = (2,\ 2,\ 6.5,\ 2,\ 4.5,\ 3.5)$$

Tretji korak Posodobimo množico genov $G = \{C_2, g_2, C_1, \ldots, g_8\}$. Poračunamo nove razdalje, prepišemo ostale:

	C_2	g_2	C_1	g_6	g_7	g_8
C_2	0	8.334	5.055	7.562	6.030	7.036
g_2		0	10.645	10.296	10.677	8.602
C_1			0	8.635	6.085	7.839
g_6				0	5.568	7.681
g_7					0	5.916
g_8						0

Četrti korak Izračunamo višino povezave.

$$h(C_2) = \frac{D(g_1, g_3)}{2} = \frac{d(g_1, g_3)}{2} = \frac{3.317}{2} = 1.659$$

Posodobimo dendrogram:

Tretja iteracija

Prvi korak

	C_2	g_2	C_1	g_6	g_7	g_8
C_2	0	8.334	5.055	7.562	6.030	7.036
g_2		0	10.645	10.296	10.677	8.602
C_1			0	8.635	6.085	7.839
g_6				0	5.568	7.681
g_7					0	5.916
g_8						0

Drugi korak Zdaj pa združimo gruči $C_3 = \{C_1, C_2\} = \{g_1, g_3, g_4, g_5\}$.
Poračunamo vektor:

$$v_3 = (2.25,\ 1.75,\ 5.25,\ 1.75,\ 2.5,\ 3.5)$$

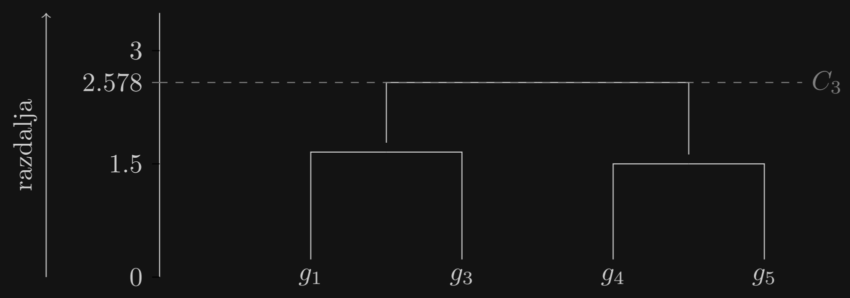
Tretji korak Posodobimo množico: $G = \{C_3, g_2, \ldots, g_8\}$.
Nove razdalje:

	C_3	g_2	g_6	g_7	g_8
C_3	0	9.490	8.099	6.058	7.438
g_2		0	10.296	10.677	8.602
g_6			0	5.568	7.681
g_7				0	5.916
g_8					0

Četrti korak Višina povezave:

$$
\begin{aligned}
h(C_3) &= \frac{D(C_1, C_2)}{2} \\
&= \frac{1}{2|C_1||C_2|} \sum_{x \in C_1, y \in C_2} d(x, y) \\
&= \frac{1}{2|C_1 \times C_2|} \sum_{u \in C_1 \times C_2} d(u) \\
&= \frac{1}{8} \cdot (d(g_4, g_1) + d(g_4, g_3) + d(g_5, g_1) + d(g_5, g_3)) \\
&= \frac{1}{8} \cdot (3.873 + 6.325 + 3.464 + 6.557) \\
&= \frac{1}{8} \cdot 20.219 \\
&= 2.5278
\end{aligned}
$$

Posodobimo dendrogram:



Četrta iteracija

Prvi korak

	C_3	g_2	g_6	g_7	g_8
C_3	0	9.490	8.099	6.058	7.438
g_2		0	10.296	10.677	8.602
g_6			0	5.568	7.681
g_7				0	5.916
g_8					0

Drugi korak Ustvarimo gručo $C_4 = \{g_6, g_7\}$.

Poračunamo vektor:

$$v_4 = (2.0,\ 4.0,\ 3.5,\ 6.5,\ 3.0,\ 2.5)$$

Tretji korak Posodobimo množico: $G = \{C_3, g_2, C_4, g_8\}$.

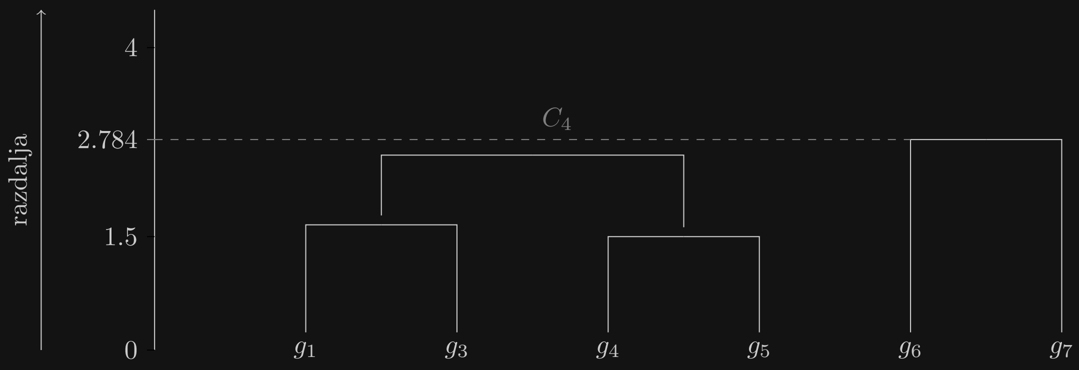
Nove razdalje:

	C_3	g_2	C_4	g_8
C_3	0	9.490	7.079	7.438
g_2		0	10.487	8.602
C_4			0	6.799
g_8				0

Četrti korak Višina povezave:

$$h(C_4) = \frac{D(g_6, g_7)}{2} = \frac{\frac{1}{|g_6||g_7|} d(g_6, g_7)}{2} = \frac{d(g_6, g_7)}{2} = \frac{5.568}{2} = 2.784$$

Posodobimo dendrogram:



Peta iteracija

Prvi korak

	C_3	g_2	C_4	g_8
C_3	0	9.490	7.079	7.438
g_2		0	10.487	8.602
C_4			0	6.799
g_8				0

Drugi korak Ustvarimo gručo $C_5 = \{C_4, g_8\} = \{g_6, g_7, g_8\}$.
Poračunamo vektor:

$$v_5 = (4.0,\ 3.75,\ 2.0,\ 4.25,\ 6.25,\ 4.25)$$

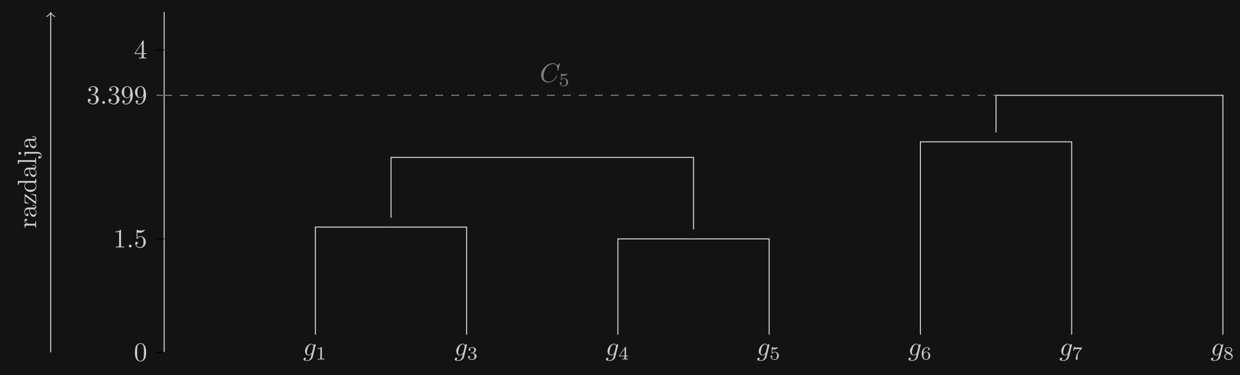
Tretji korak Posodobimo množico: $G = \{C_3, g_2, C_5\}$.
Nove razdalje:

	C_3	g_2	C_5
C_3	0	9.490	7.258
g_2		0	9.545
C_5			0

Četrti korak Višina povezave:

$$\begin{aligned} h(C_5) &= \frac{D(C_4, g_8)}{2} \\ &= \frac{1}{2} \cdot 6.799 \\ &= 3.399 \end{aligned}$$

Posodobimo dendrogram:



Šesta iteracija

Prvi korak

	C_3	g_2	C_5
C_3	0	9.490	7.258
g_2		0	9.545
C_5			0

Drugi korak Ustvarimo gručo $C_6 = \{C_3, C_5\} = \{g_1, g_3, g_4, g_5, g_6, g_7, g_8\}$.
Poračunamo vektor:

$$v_6 = (3.125,\ 2.75,\ 3.625,\ 3.0,\ 4.375,\ 3.875)$$

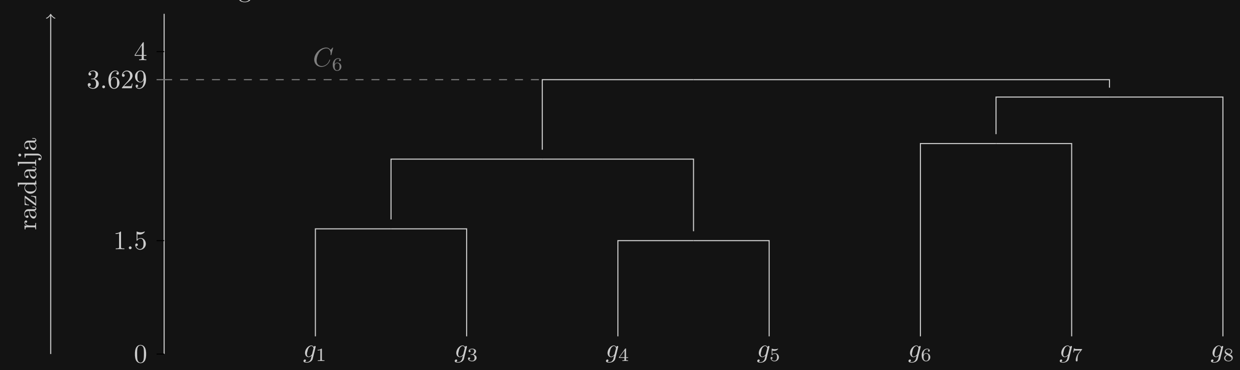
Tretji korak Posodobimo množico: $G = \{g_2, C_6\}$.
Nove razdalje:

	C_6	g_2
C_6	0	9.518
g_2		0

Četrti korak Višina povezave:

$$\begin{aligned} h(C_6) &= \frac{D(C_3, C_5)}{2} \\ &= \frac{1}{2} \cdot 7.258 \\ &= 3.629 \end{aligned}$$

Posodobimo dendrogram:



Sedma iteracija

Prvi korak

	C_6	g_2
C_6	0	9.518
g_2		0

Drugi korak Ustvarimo gručo $C_7 = \{g_2, C_6\} = \{g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8\}$.
Poračunamo vektor:

$$v_7 = (3.0625,\ 4.875,\ 3.3125,\ 2.0,\ 6.6875,\ 3.4375)$$

Tretji korak Posodobimo množico: $G = \{C_7\}$.
Nove razdalje: ker smo združili vse gene, smo odstranili še zadnje elemente matrike.

Četrti korak Višina povezave:

$$\begin{aligned} h(C_7) &= \frac{D(C_6, g_2)}{2} \\ &= \frac{1}{2} \cdot 9.518 \\ &= 4.759 \end{aligned}$$

Končni dendrogram: