Algorytmy w bioinformatyce sekwencji

analiza danych wielowymiarowych; grupowanie; drzewa filogenetyczne

Robert Nowak

2024

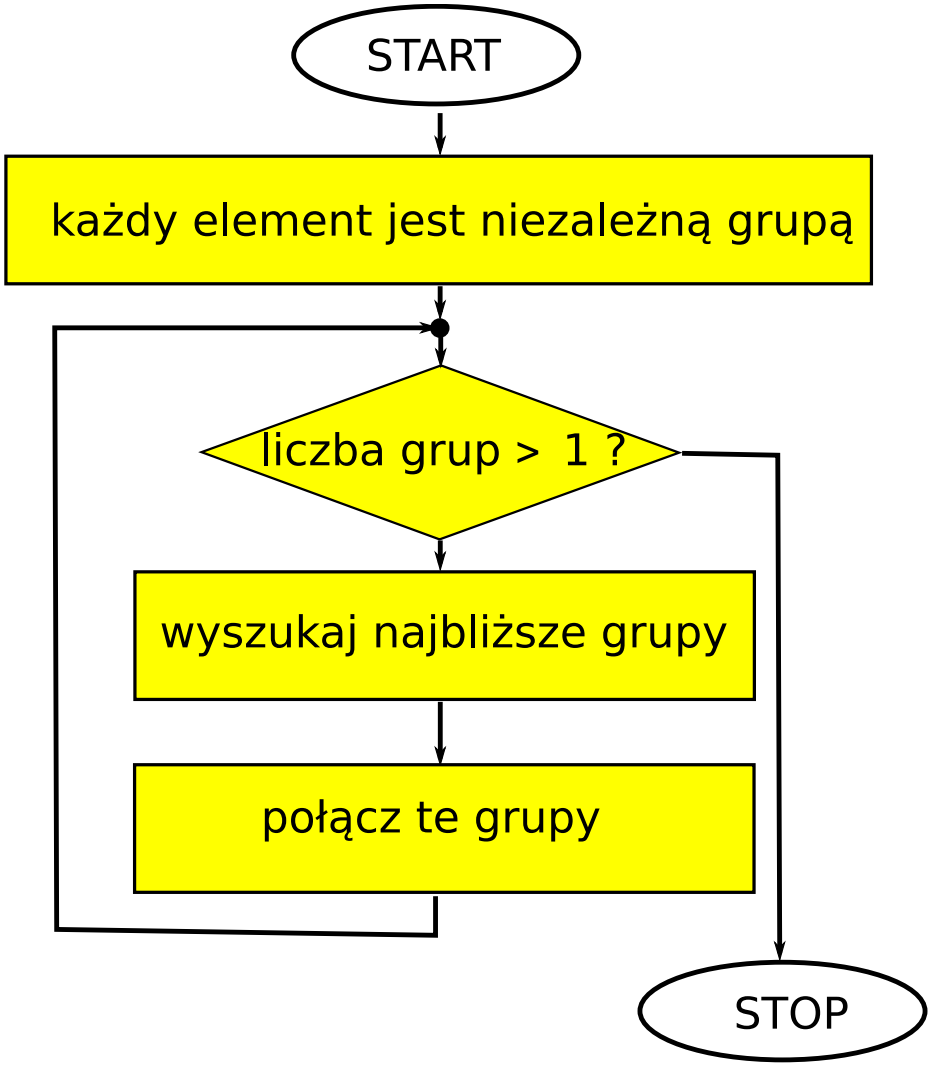
Grupowanie

grupowanie (ang. clustering)

Grupowanie to tworzenie klastrów (grup) obiektów o podobnych cechach.

- wymaga definicji, co to znaczy, że obiekty są „podobne”.
- obiekt jest „podobny” do dowolnego obiektu z tej samej grupy, i jest „niepodobny” do każdego obiektu z innych grup.
- istnieją miary jakości grupowania oceniające jednocześnie
  - czy klastry są zwarte („podobne” obiekty w grupach)
  - czy klastry są rozłączne („niepodobne” obiekty pomiędzy grupami)

algorytm grupowania hierarchicznego



- wymaga definicji odległości pomiędzy elementami
- wymaga definicji odległości pomiędzy grupami

definicja odległości

x, y oznaczają punkty ze zbiorów danych, każdy punkt ma m cech (atrybutów) $x =< x_1, x_2, ..., x_m >$

- odległość Euklidesowa

$$d_{xy} = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

- odległość Manhattan

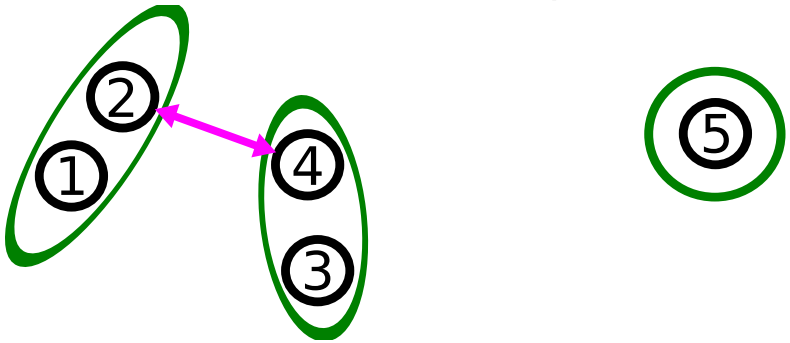
$$d_{xy} = \sum_{i=1}^{m}|x_i - y_i|$$

- korelacja

$$d_{xy} = \sum_{i=1}^{m} x_i y_i$$

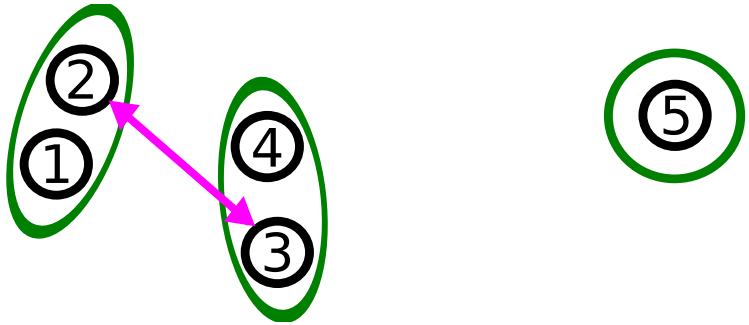
odległości grup

pojedyncze wiązanie (najmniejsza odległość)



$$d_s(X, Y) = \min_{x \in X, y \in Y} d_{xy}$$

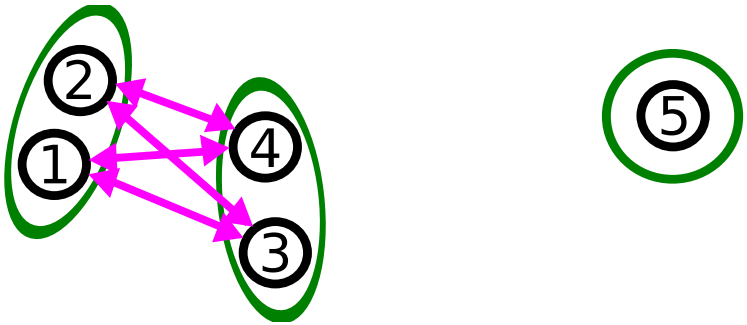
pełne wiązanie (największa odległość)



$$d_f(X, Y) = \max_{x \in X, y \in Y} d_{xy}$$

odległości grup (2)

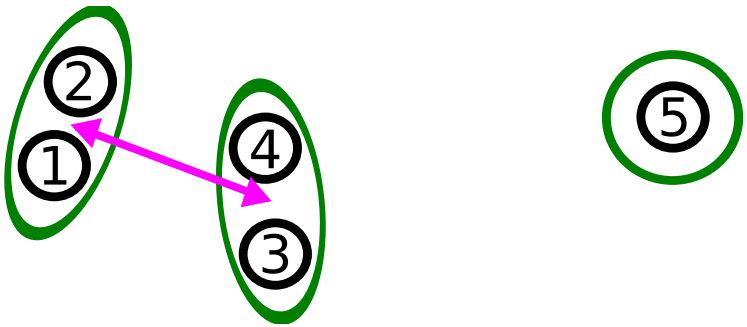
średnia odległość



$$d_a(X, Y) = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d_{xy}$$

gdzie $|X|$ to ilość elementów w X

odległość pomiędzy środkami



$$d_c(X, Y) = d_{x'y'}$$

gdzie x' to średnia elementów $x \in X$

algorytm grupowania hierarchicznego, przykład

Tabela odległości pomiędzy obiektami A, B, C, D:

	A	B	C	D
A	0	2	6	11
B		0	4	9
C			0	5
D				0

Dla dwóch grup:

- jeżeli odległość między grupami to pojedyncze wiązanie (minimalna odległość pomiędzy elementami)?
  {A,B,C}{D}
- jeżeli odległość między grupami to pełne wiązanie (maksymalna odległość pomiędzy elementami)?
  {A,B}{C,D}

algorytm grupowania hierarchicznego - złożoność

- liczba przykładów: n, liczba atrybutów: m
- liczba kroków algorytmu $n - 1$
- każda iteracja:
  - $n(n-1)/2$ razy oblicza odległość
  - koszt obliczenia odległości $O(m)$
  - koszt iteracji $O(n^2 m)$

$$\mathbf{O(n^3 m)}$$

algorytm K-średnich (grupowanie niehierarchiczne)

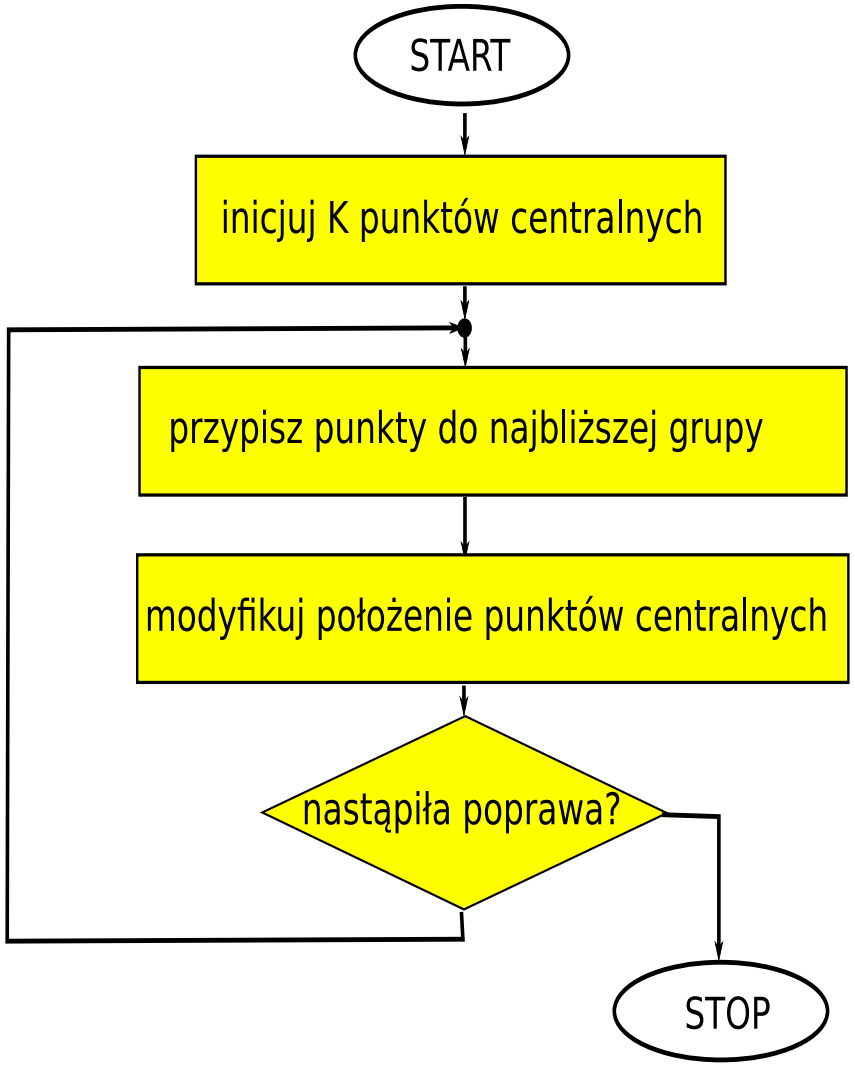
zakłada podział na K grup

- odległość
  $x =< x_1, x_2, ..., x_m >$ od c

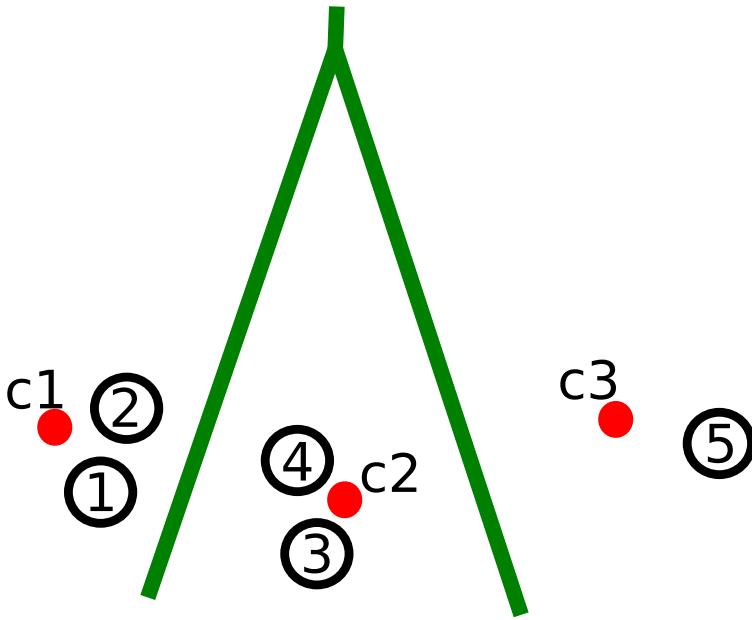
$$d_{xc} = \sum_{i=1}^{m} (x_i - c_i)^2$$

- funkcja błędu (którą minimalizujemy)

$$E = \sum_{c} \sum_{x \in c} d_{xc}$$



algorytm K-średnich (2)

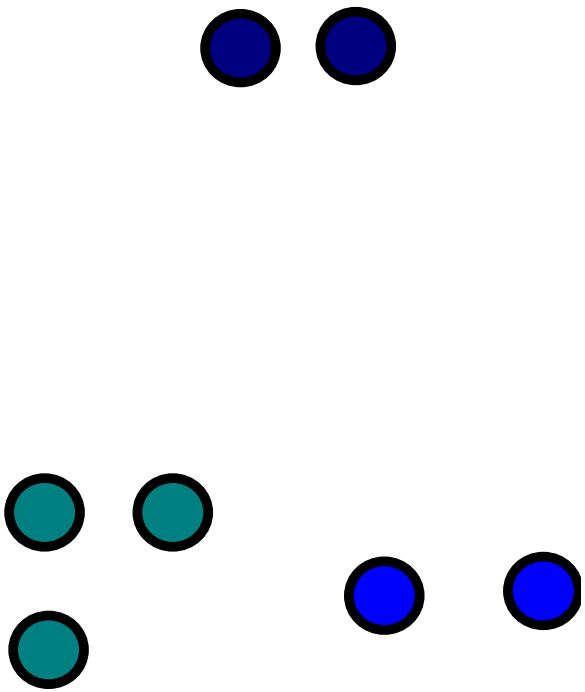


Algorytm optymalizacyjny:

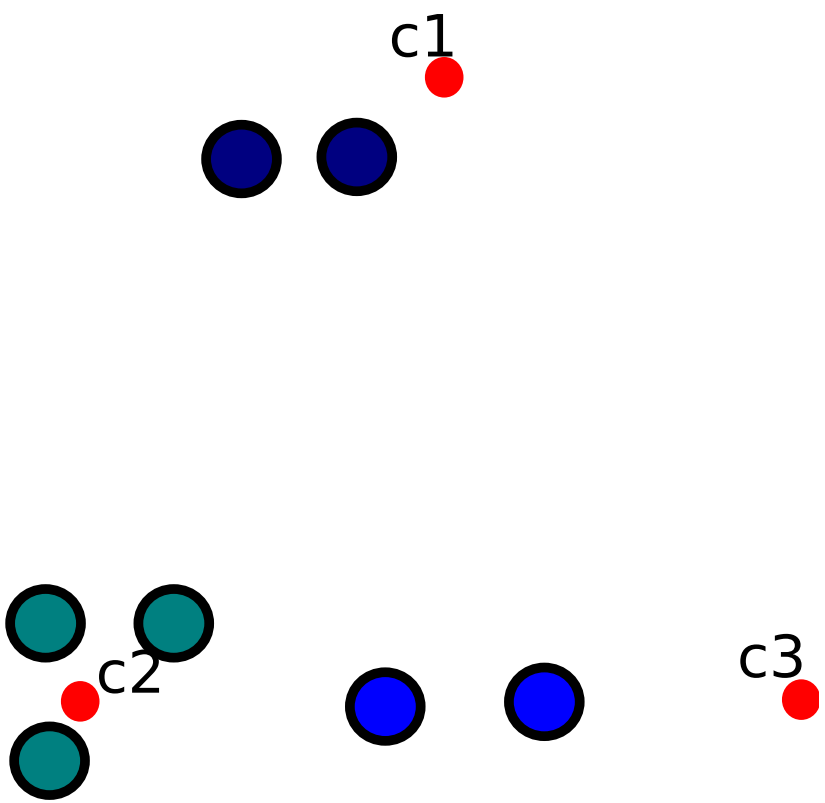
- inicjacja: losowa
- funkcja celu:

$$\arg\min_{C_1, C_2, \ldots, C_k} \sum_{c=1}^{k} \sum_{x_i \in C_c} \sum_{j=1}^{m} (x_{ij} - c_{cj})^2$$

algorytm K-średnich (przykład)

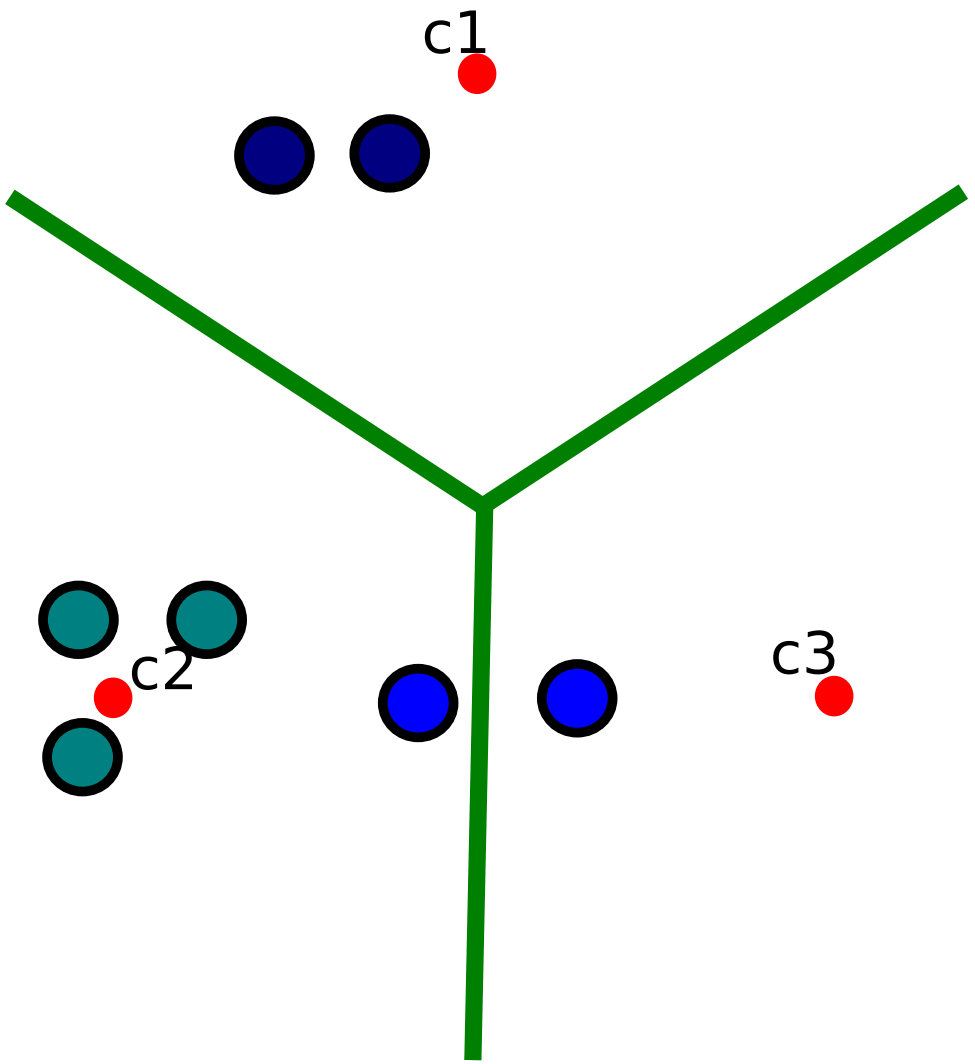


algorytm K-średnich (przykład)



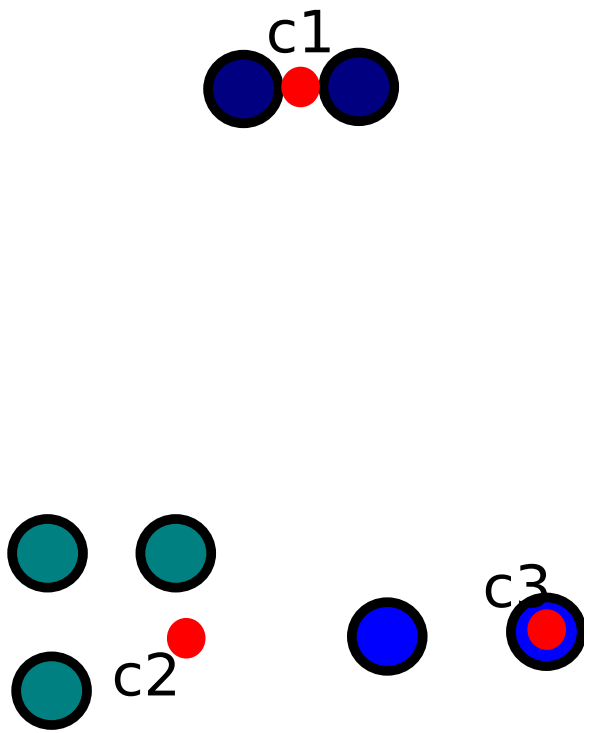
- poszukiwane 3 grupy
- inicjacja punktów centralnych

algorytm K-średnich (przykład)



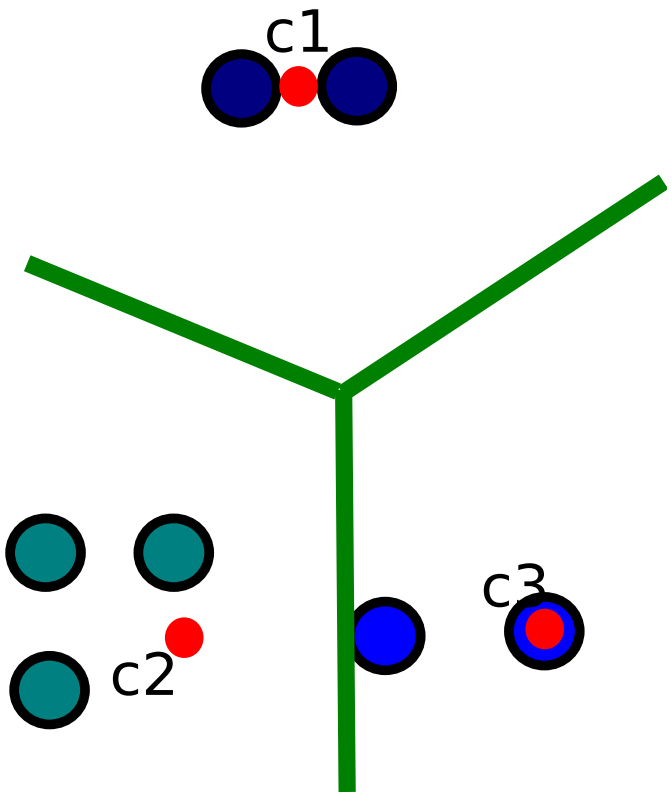
obliczenie przykładów należących do danej grupy

algorytm K-średnich (przykład)



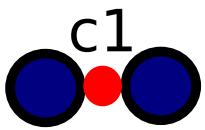
aktualizacja położenia
punktów centralnych

algorytm K-średnich (przykład)

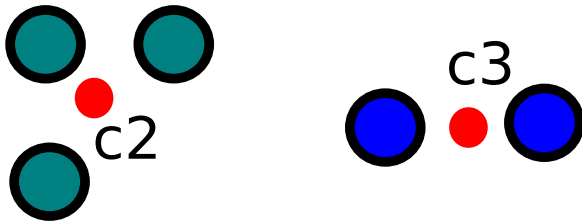


obliczenie przykładów należących do danej grupy

algorytm K-średnich (przykład)



aktualizacja punktów centralnych



algorytm K-średnich - złożoność

- liczba przykładów: n, liczba atrybutów m, liczba kroków algorytmu $p \approx n$
- każda iteracja:
  - koszt obliczenia odległości $O(m)$
  - koszt znalezienia grupy: dla n punktów k razy oblicza odległość od punktu centralnego, więc $O(knm)$
  - koszt obliczenia nowego punktu środkowego, dla n punktów oblicza średnią $O(nm)$
  - iteracja $O(knm + nm) = O(knm)$

$$\mathbf{O(kn^2m)}$$

algorytm znacznie wydajniejszy niż grupowanie hierarchiczne $O(n^3m)$, ponieważ $k \ll n$, ale problemy z właściwą inicjacją

Miara jakości grupowania

Uwzg. odchylenie wewnątrzklastrowe (klastry zwarte)

$$wc = \sum_{i}^{n} \sum_{x \in C_i} d(x, r_i),\ r_i \text{ to środek klastra } C_i$$

oraz odchylenie międzyklastrowe (klastry rozłączne)

$$bc = \sum_{i,j \neq i}^{n} d(r_i, r_j)$$

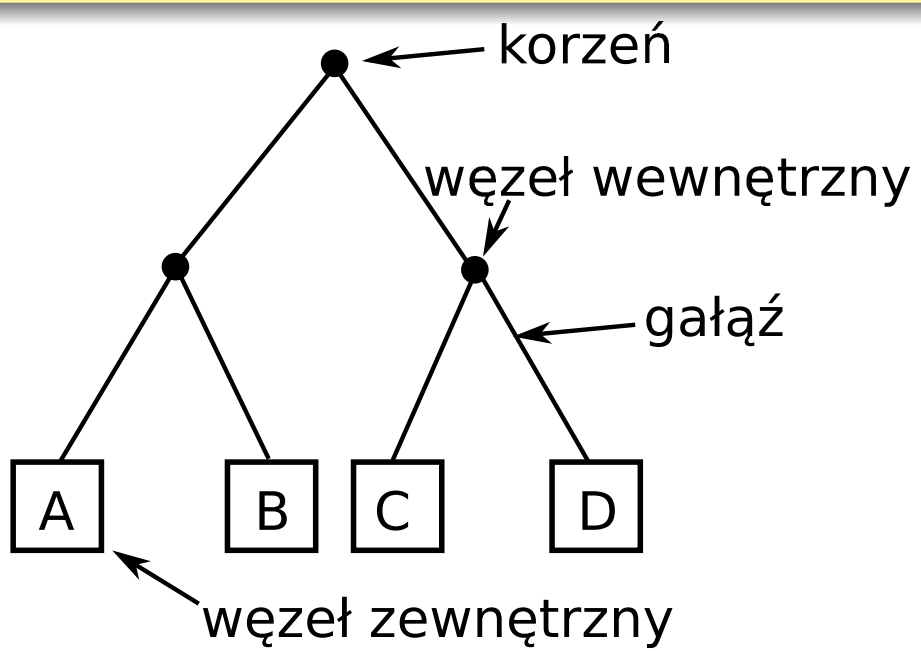
Przykłady miar:

- $$\frac{bc}{wc},$$
- indeks Daviesa-Bouldina,
- indeks Dunna.

Drzewa filogenetyczne

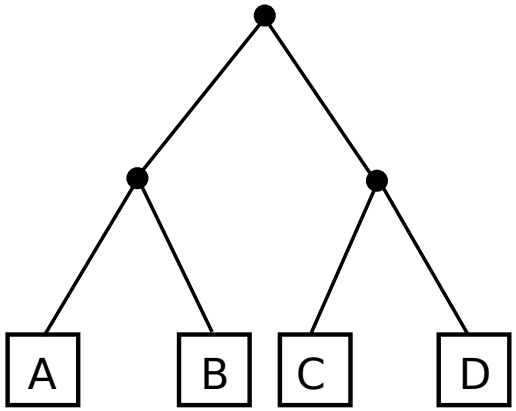
Drzewa filogenetyczne

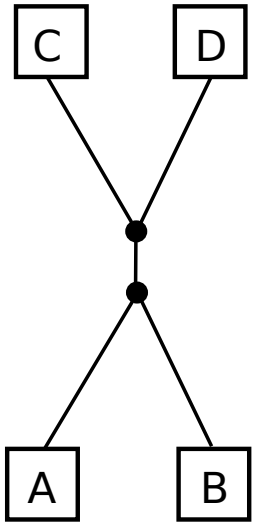
Drzewo filogenetyczne przedstawia zależności pomiędzy wieloma sekwencjami.



- drzewa ukorzenione i nieukorzenione
- drzewa binarne i inne
- krawędzie obrazują lub nie obrazują odległość

Drzewa ukorzenione i nieukorzenione





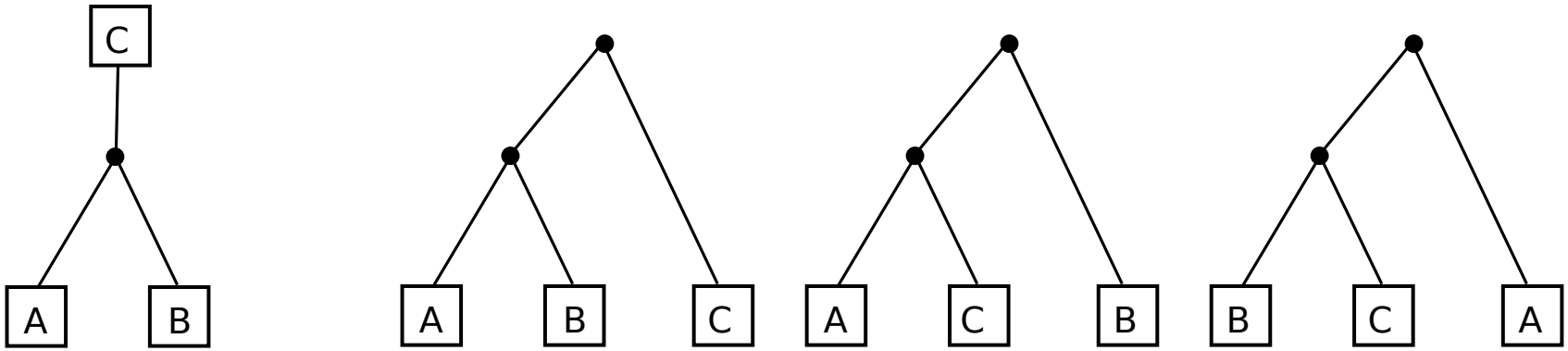
- liczba drzew ukorzenionych, gdzie n to liczba taksonów (węzłów zewnętrznych)

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

- liczba drzew nieukorzenionych

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Drzewa ukorzenione i nieukorzenione (2)



n	N_U	N_R
3	1	3
4	3	15
6	105	945
10	$2 * 10^6$	$34 * 10^6$

n - liczba taksonów, N_U liczba drzew nieukorzenionych, N_R liczba drzew ukorzenionych

Algorytmy do konstrukcji drzew

- metody bazujące na odległościach sekwencji
  - metoda średnich połączeń (UPGMA - unweighted pair group method with arithmetic mean)
  - metoda przyłączania sąsiadów (NJ - neighbour joining)
- metody bazujące na analizie symboli
  - metoda parsymonii (Parsimony)
  - metoda największej wiarygodności (Maximum likelihood)

UPGMA - metoda grupowania parami ze średnią arytmetyczną

- bazuje na odległościach pomiędzy sekwencjami
- grupuje najbliższe taksony
- zakłada jednakową odległość taksonów od korzenia
- bardzo wydajna

Algorytm UPGMA identyczny jak algorytm grupowania hierarchicznego.

UPGMA - przykład (1)

Przykładowa macierz odległości pomiędzy taksonami A, B, C, D:

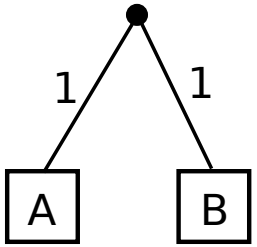
	B	C	D
A	2	6	11
B		4	9
C			4

Najbliższe taksony *A* i *B*

Zredukowana macierz odległości:

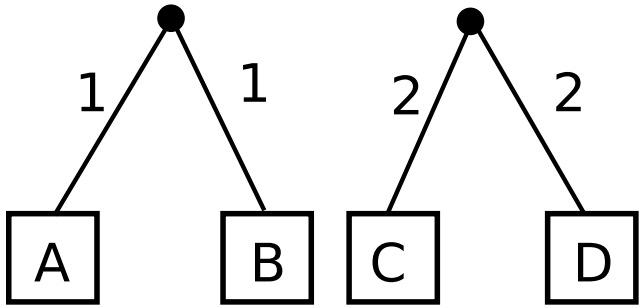
	C	D
{A, B}	5	10
C		4

Drzewo po utworzeniu taksonu {A, B}



Węzeł wewnętrzny umieszczamy w połowie odległości.

Drzewo w kroku nr 2.



UPGMA - przykład (2)

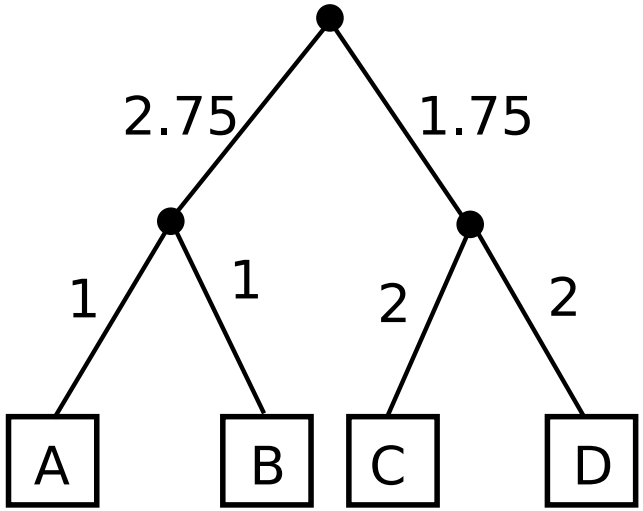


Tabela odległości odtworzona z drzewa:

	B	C	D
A	2	7.5	7.5
B		7.5	7.5
C			4

tabela pierwotna:

	B	C	D
A	2	6	11
B		4	9
C			4

NJ - metoda przyłączania sąsiadów

- bazuje na odległościach pomiędzy sekwencjami
- grupuje najbliższe taksony
- nie zakłada jednakowej odległość taksonów od korzenia

Algorytm NJ podobny do UPGMA

algorytm NJ - obliczenia

- korekta odległości pomiędzy taksonami

$$\begin{aligned} d_{ij} &\quad \text{odległość pomiędzy taksonami} \\ r_i &= \textstyle\sum_{i \neq j} d_{ij} \text{ odległość dla taksonu} \\ d'_{ij} &= d_{ij} - \frac{r_i + r_j}{2} \text{ korekta odległości} \end{aligned}$$

- dodanie nowego węzła

$$\begin{aligned} r'_i &= \frac{r_i}{n-2}, \text{ gdzie} n \text{ to liczba taksonów} \\ d_{iu} &= \frac{d_{ij} + r'_i - r'_j}{2} \text{ nowy węzeł } u \\ d_{ju} &= \frac{d_{ij} + r'_i - r'_j}{2} \text{ nowy węzeł } u \end{aligned}$$

- redukcja macierzy (dodano takson u pomiędzy i i j)

$$d_{ku} = \frac{d_{ik} - d_{iu} + d_{jk} - d_{ju}}{2} \text{ dla każdego } k \neq i \wedge k \neq j$$

algorytm NJ - przykład (1)

Przykładowa macierz odległości pomiędzy taksonami A, B, C, D:

	B	C	D
A	2	5	4
B		3	2
C			3

Macierz po korekcie:

	B	C	D
A	-7	-6	-6
B		-6	-6
C			-7

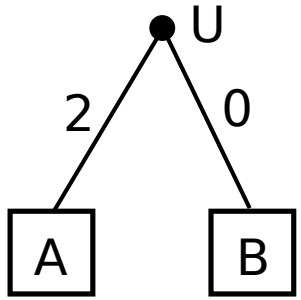
$$
\begin{aligned}
r_A &= 2 + 5 + 4 = 11 \\
r_B &= 7 \\
r_C &= 11 \\
r_D &= 9 \\
d'_{AB} &= d_{AB} - \frac{r_A + r_B}{2} = -7
\end{aligned}
$$

Minimalna odległość to d_{AB} lub d_{CD}, wybieramy dowolną z tych par, tutaj d_{AB}

Algorytm NJ - przykład (2)

Nowy węzeł U pomiędzy A i B

$$
\begin{aligned}
r'_A &= \frac{r_A}{4-2} = 5.5 \\
r'_B &= 3.5 \\
d_{AU} &= \frac{d_{AB}+r'_A-r'_B}{2} = 2 \\
d_{BU} &= \frac{d_{AB}+r'_B-r'_A}{2} = 0
\end{aligned}
$$



Redukcja:

$$
\begin{aligned}
d_{CU} &= \frac{d_{AC}-d_{AU}+d_{BC}-d_{BU}}{2} = \frac{5-2+3-0}{2} = 3 \\
d_{DU} &= \frac{d_{AD}-d_{AU}+d_{BD}-d_{BU}}{2} = 2
\end{aligned}
$$

Macierz po redukcji:

	C	D
U	3	2
C		3

Algorytm NJ - przykład (3)

Macierz po redukcji:

	C	D
U	3	2
C		3

$$r_U = 3 + 2 = 5$$
$$r_C = 6$$
$$r_D = 5$$

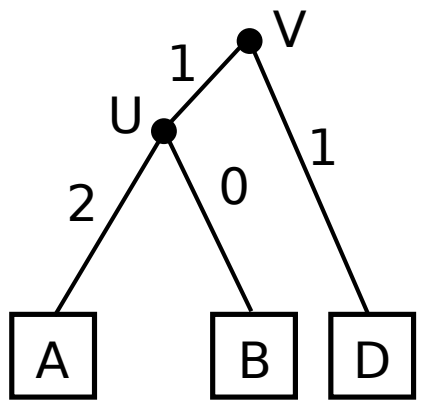
Macierz po korekcie:

	C	D
U	-2.5	-3
C		-2.5

Minimalna odległość to d_{UD}

Nowy węzeł V pomiędzy U i D

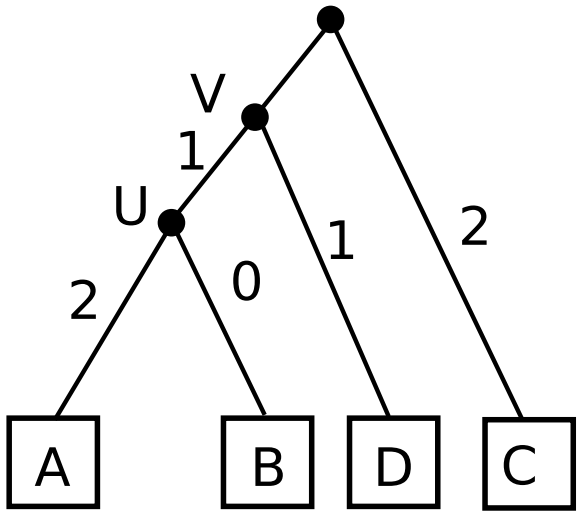
$$r'_U = \frac{r_U}{3-2} = 5$$
$$r'_D = 6$$
$$d_{UV} = \frac{d_{UD}+r'_U-r'_D}{2} = 1$$
$$d_{DV} = \frac{d_{UD}+r'_D-r'_U}{2} = 1$$



Algorytm NJ - przykład (4)

Po redukcji:

$$d_{CV} = \frac{d_{UC} - d_{UV} + d_{CD} - d_{DV}}{2} = 2$$



	B	C	D
A	2	5	4
B		3	2
C			3

Tabela odległości odtworzona z drzewa jest identyczna z tabelą pierwotną

Dziękuję