Metody detekcji i analizy patogennych zmian strukturalnych w genomie człowieka

Anna Gambin,
Instytut Informatyki,
Uniwersytet Warszawski

Wykład 2: algorytmy grupowania

- wstęp
- przykłady
- podejścia do problemu klasteryzacji
- podstawowe algorytmy
- jak ocenić jakość grupowania ?

- grupowanie = klasteryzacja = uczenie bez nadzoru

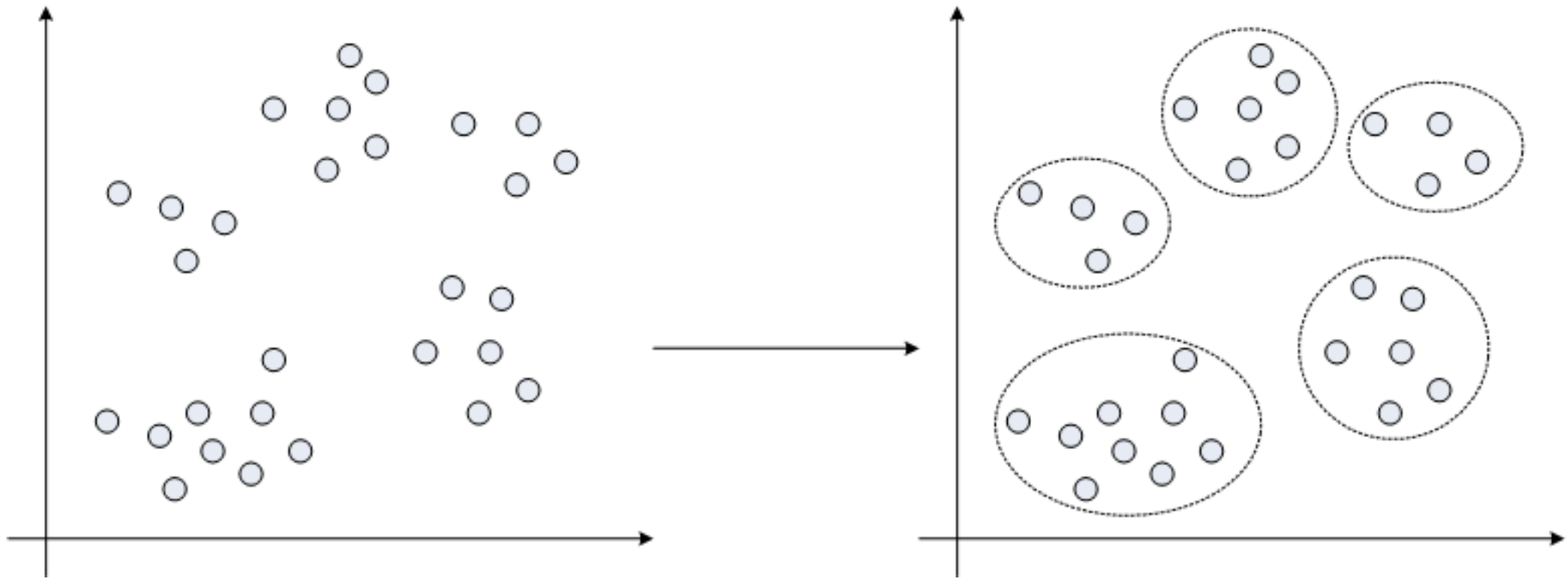
- w odróżnieniu od problemu klasyfikacji nie dysponujemy wiedzą o przynależności do klas

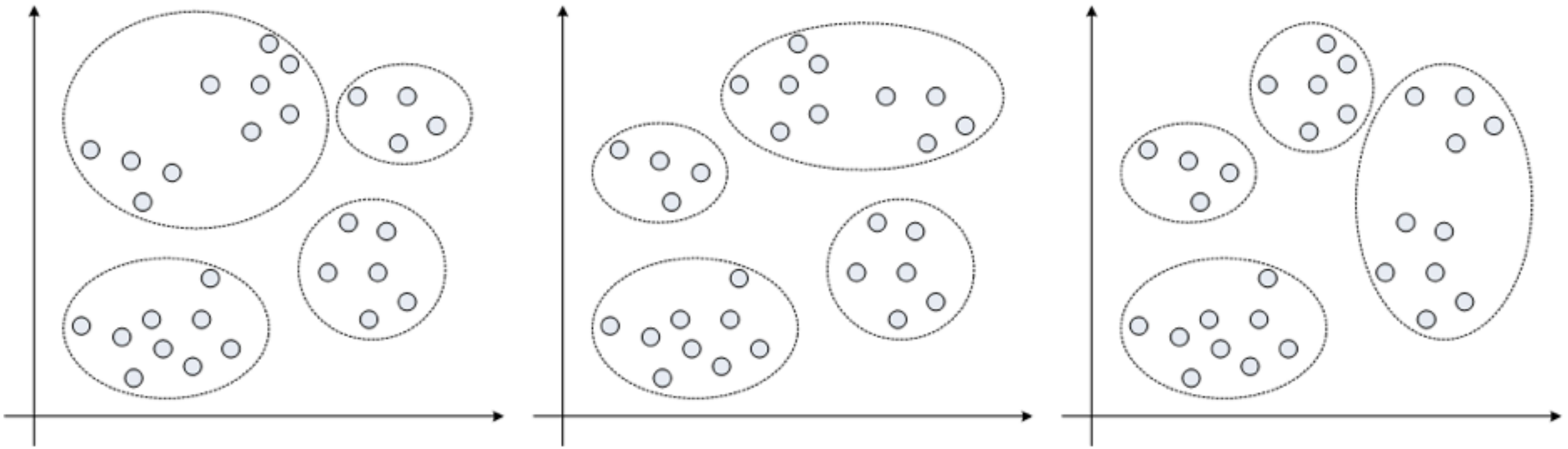
- klaster = zbiór podobnych obiektów

- najczęściej grupujemy w oparciu o odległość pomiędzy obiektami

	Adelaide	Amundsen-Scott	Buenos Aires	Bouvet Island	Capetown	Casey	Christchurch	Commonwealth Bay	Davis	Dumont d'Urville	Edgeworth David Base
Adelaide	6133										
Amundsen-Scott	7925	6165									
Buenos Aires	9212	3968	5185								
Bouvet Island	9845	6246	6895	2564							
Capetown	3936	2647	8772	5298	6693						
Casey	3072	5178	9915	9098	8987	4450					
Christchurch	3341	2803	8815	6324	8035	1475	3019				
Commonwealth Bay	5249	2392	8007	3965	5309	1400	5766	2747			
Davis	3529	2605	8660	6091	7811	1302	3235	233	2533		
Dumont d'Urville	4264	2651	8674	4948	6259	445	4895	1911	998	1728	
Edgeworth David Base											

przykład





- redukcja danych
- odkrywanie „naturalnych” grup i ich własności
  - grupowanie profili ekspresji genów
  - interpretacja danych spektrometrycznych
- znajdowanie elementów odstających (*ang. outliers*)

Jako przykład obiektów rozważmy dane o ekspresji genów pochodzące z eksperymentu mikromacierzowego. Niech $X = [x_{ij}] \in \mathcal{R}$ będzie macierzą liczb rzeczywistych. Indeks $i = 1 \ldots n$ odpowiada badanym genom, natomiast indeks $j = 1 \ldots m$ numeruje kolejne eksperymenty. Na podstwie macierzy X, która określana jest mianem *surowych danych* buduje się często *macierz podobieństwa* pomiędzy genami gdzie wartość $x_{ij} \in \mathcal{R}$ dla $i, j = 1 \ldots n$ odpowiada np. korelacji Pearsona dla profili ekspresji genów i oraz j.

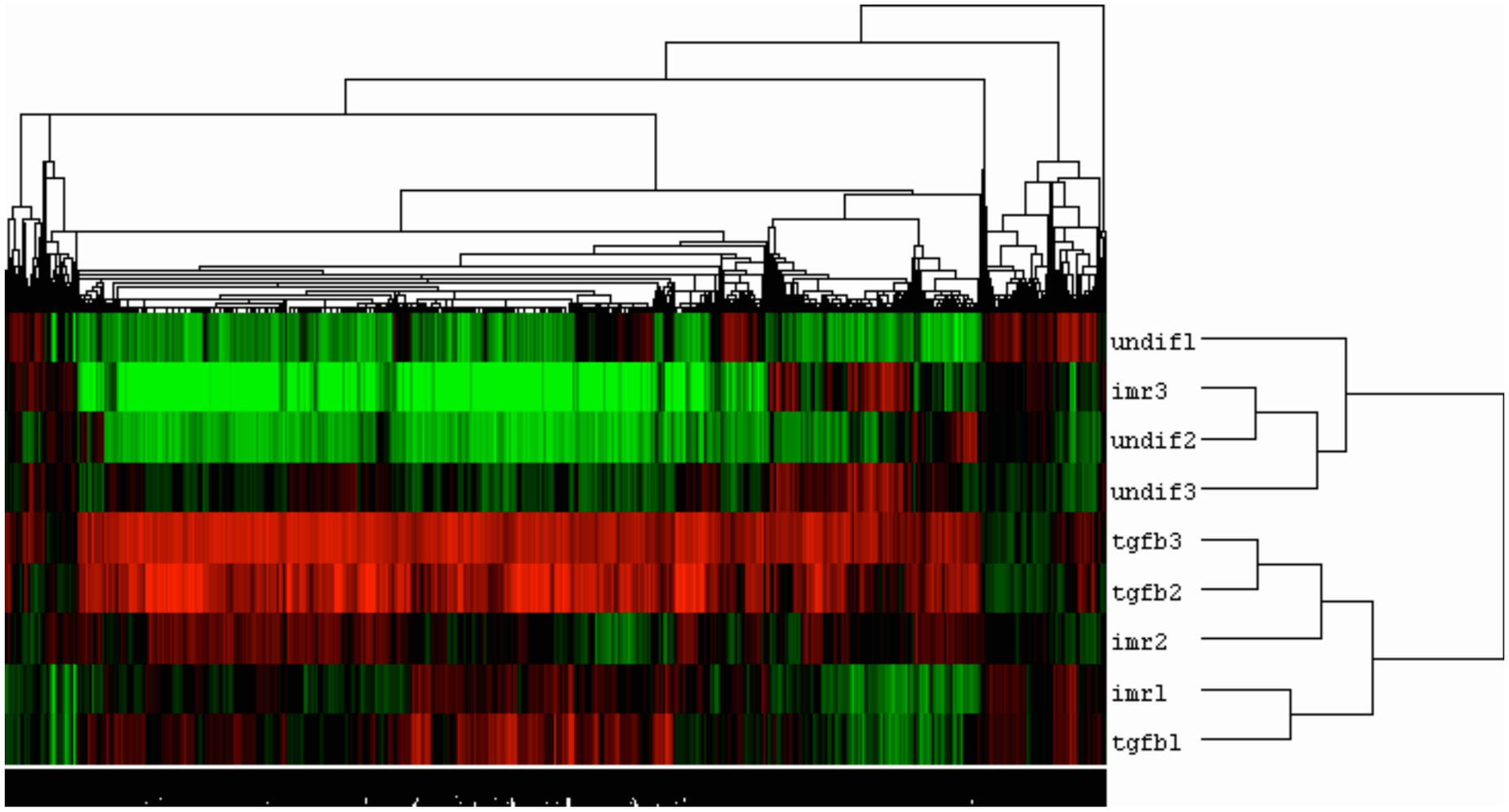
Korelacja Pearsona dla dwóch prób losowych zdefiniowana jest jako:

$$r_{ij} = \frac{s_{ij}^2}{\sqrt{s_i^2}\sqrt{s_j^2}}$$

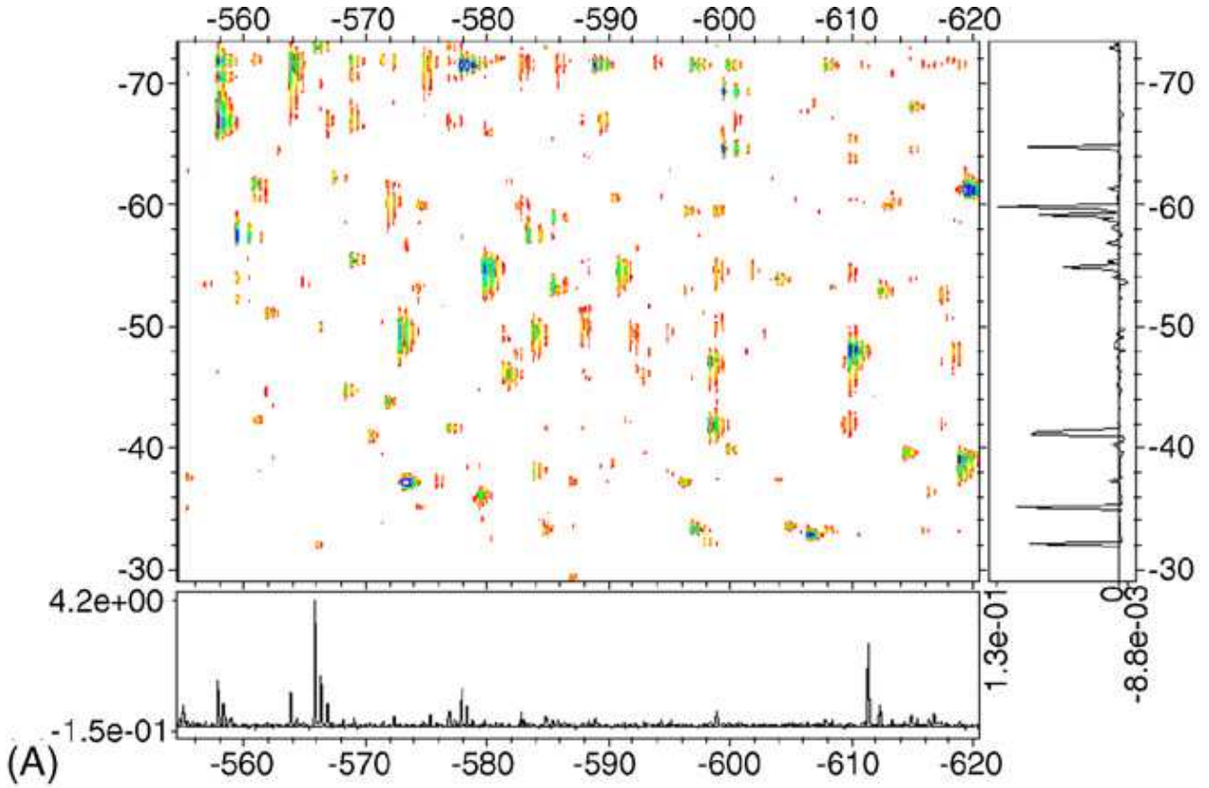
gdzie $s_i^2 = \frac{1}{m}\Sigma_{k=1}^{m}(x_{ki} - \bar{x_i})^2$ jest wariancją z próby $X_i = (x_{1i}, \ldots x_{mi})$, średnia $\bar{x_i} = \frac{1}{m}\Sigma_{k=1}^{m} x_{ki}$ oraz

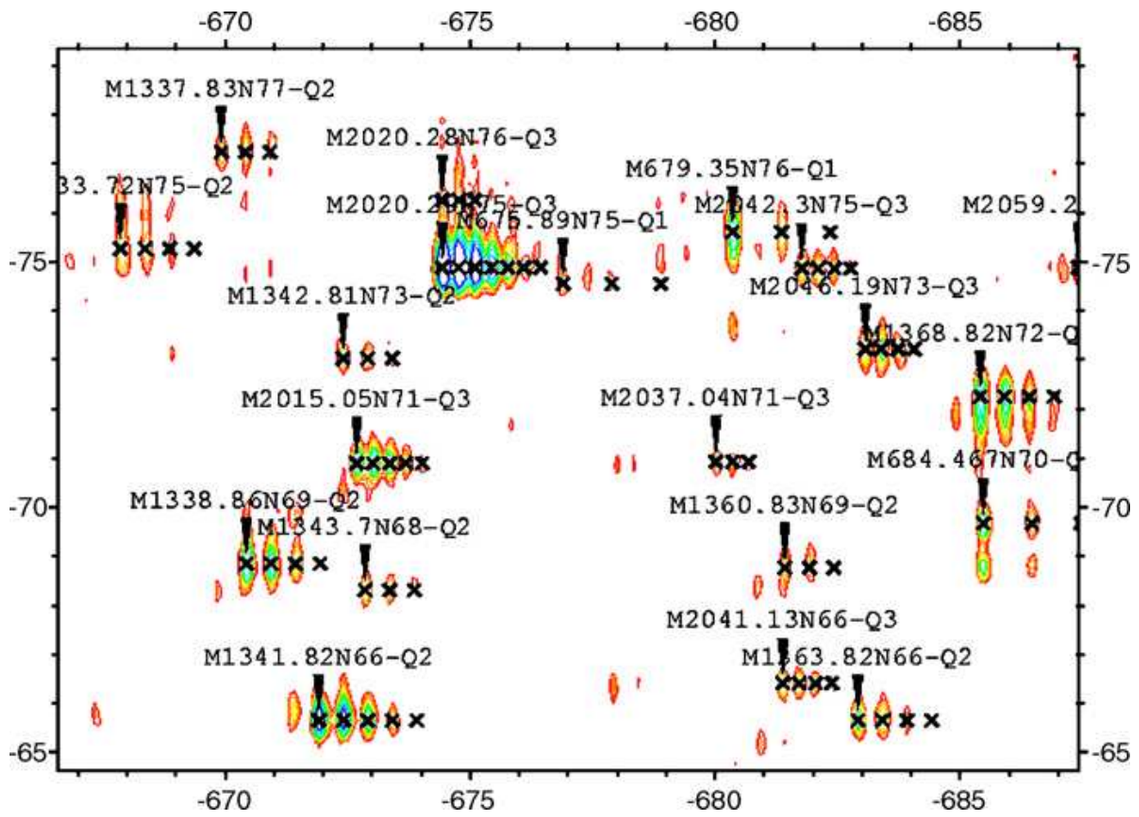
$$s_{ij}^2 = \frac{1}{m}\Sigma_{k=1}^{m}(x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j})$$

jest kowariancją dla X_i oraz X_j.


undif1
imr3
undif2
undif3
tgfb3
tgfb2
imr2
imr1
tgfb1

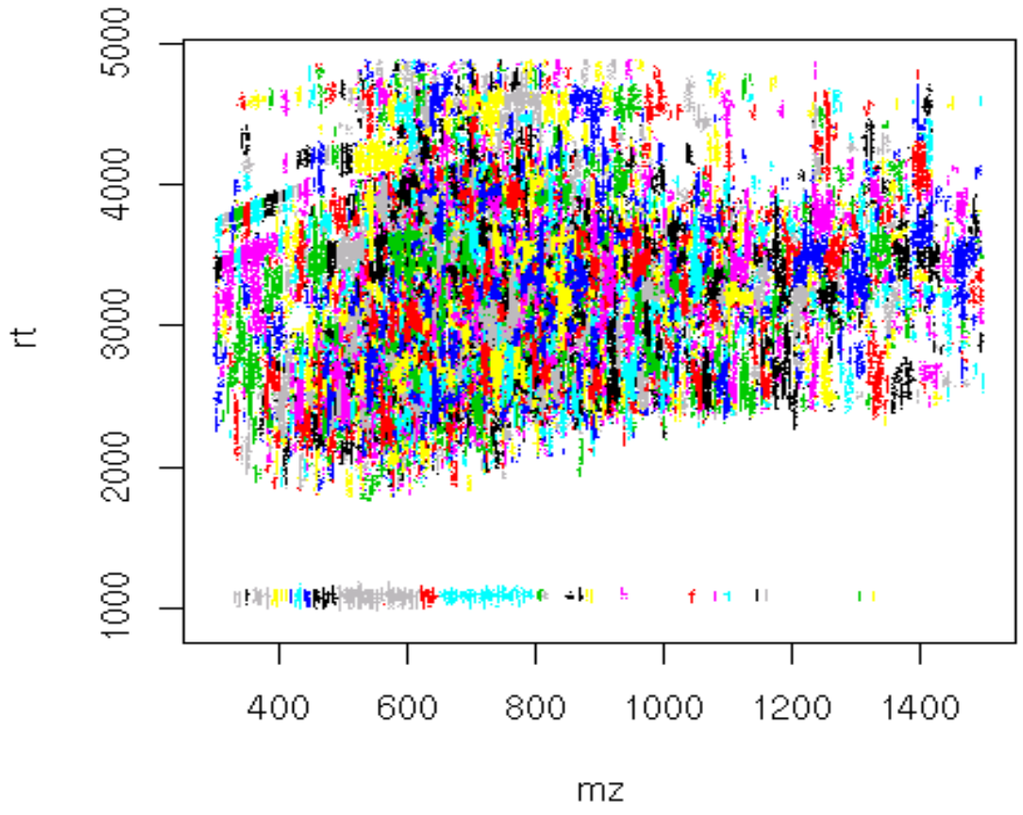
przykład: LC-MS

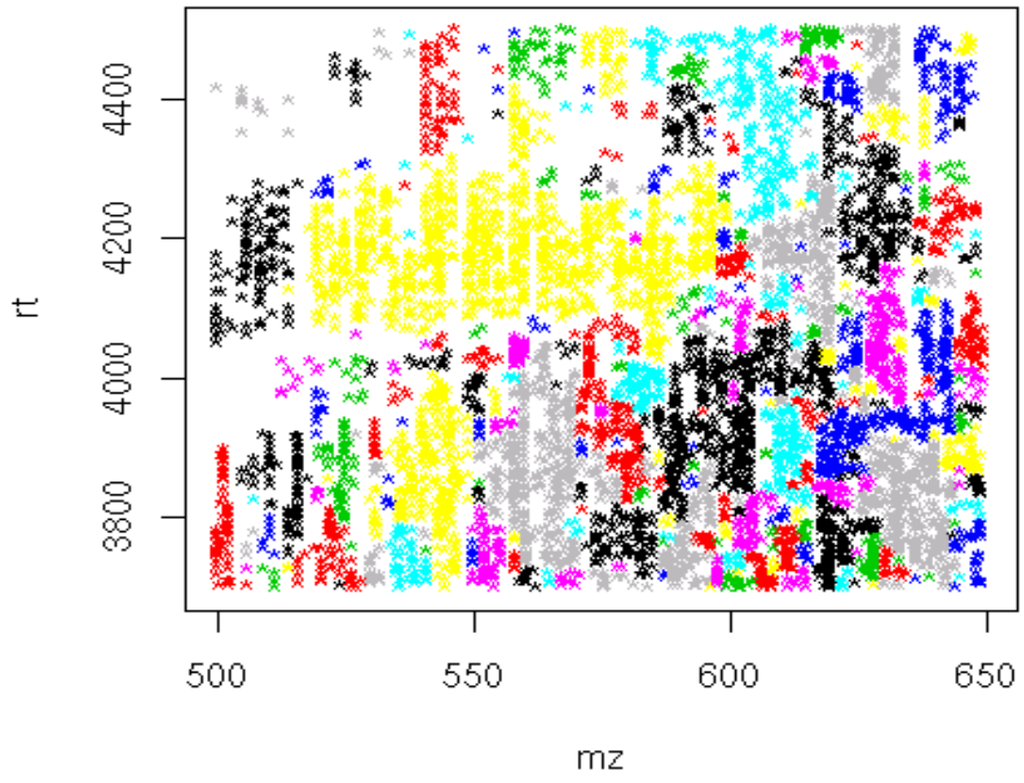




A. Gambin et al. / International Journal of Mass Spectrometry 260 (2007) 20–30

przykład: uliniowienie próbek LC-MS





Statistical Applications in Genetics and Molecular Biology, Vol. 8 [2009], Iss. 1, Art. 15

przykład

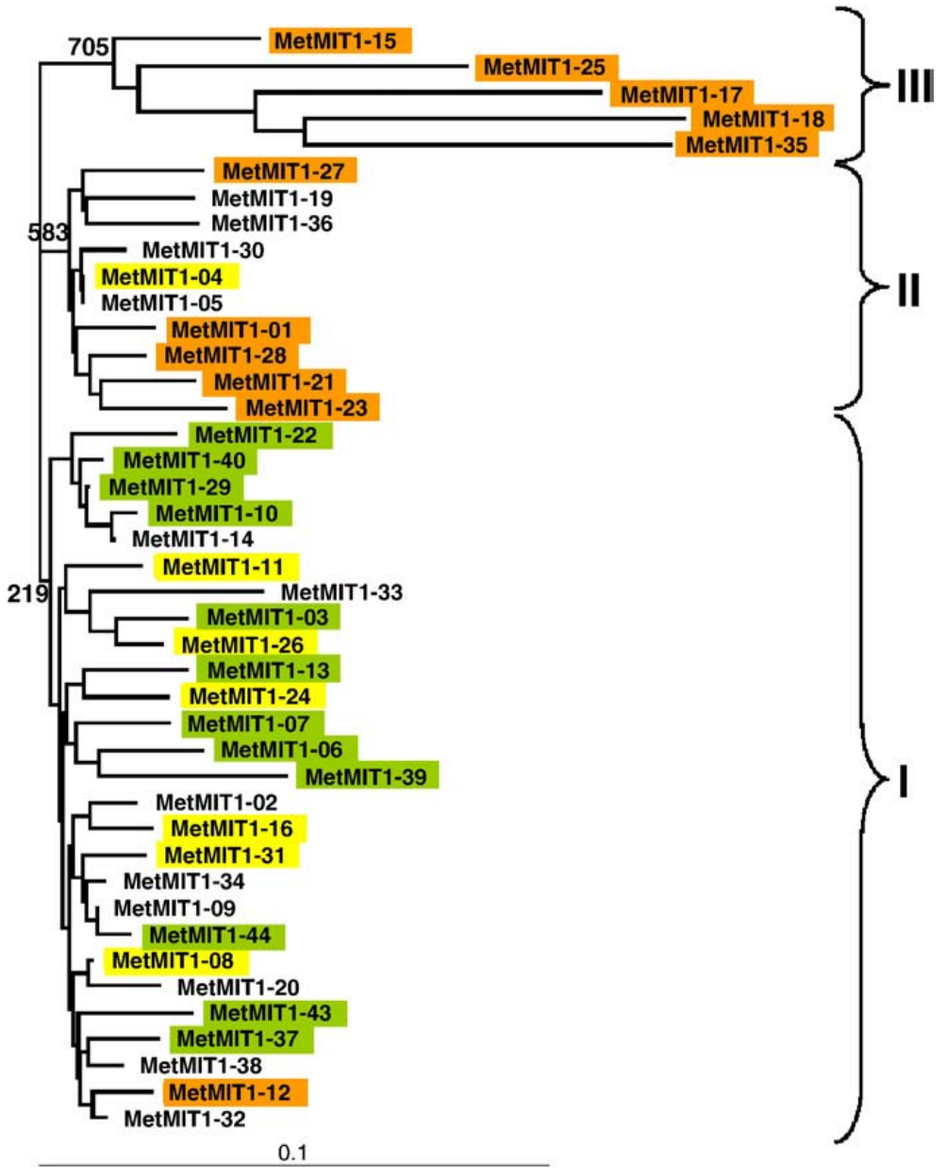


Fig. 3. Bootstrap neighbor-joining tree of the *MetMIT1* family. Elements representing insertions revealed as fixed, polymorphic among *M. truncatula* ecotypes, and unique to A17 'Jemalong' are marked orange, green, and yellow, respectively. Numbers show bootstrap values for the three major clusters.

D. Grzebelus et al. / Gene 448 (2009) 214–220

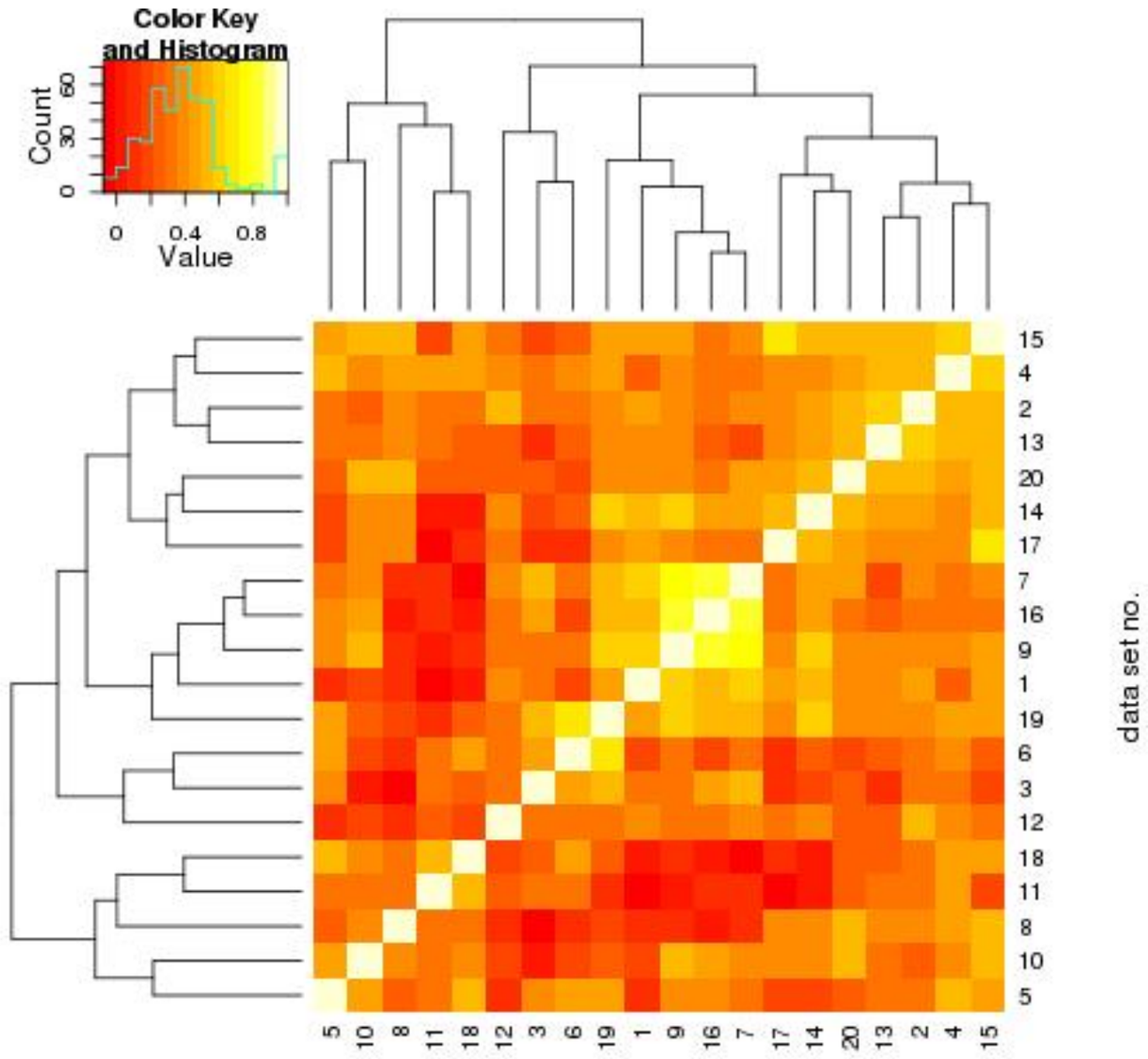


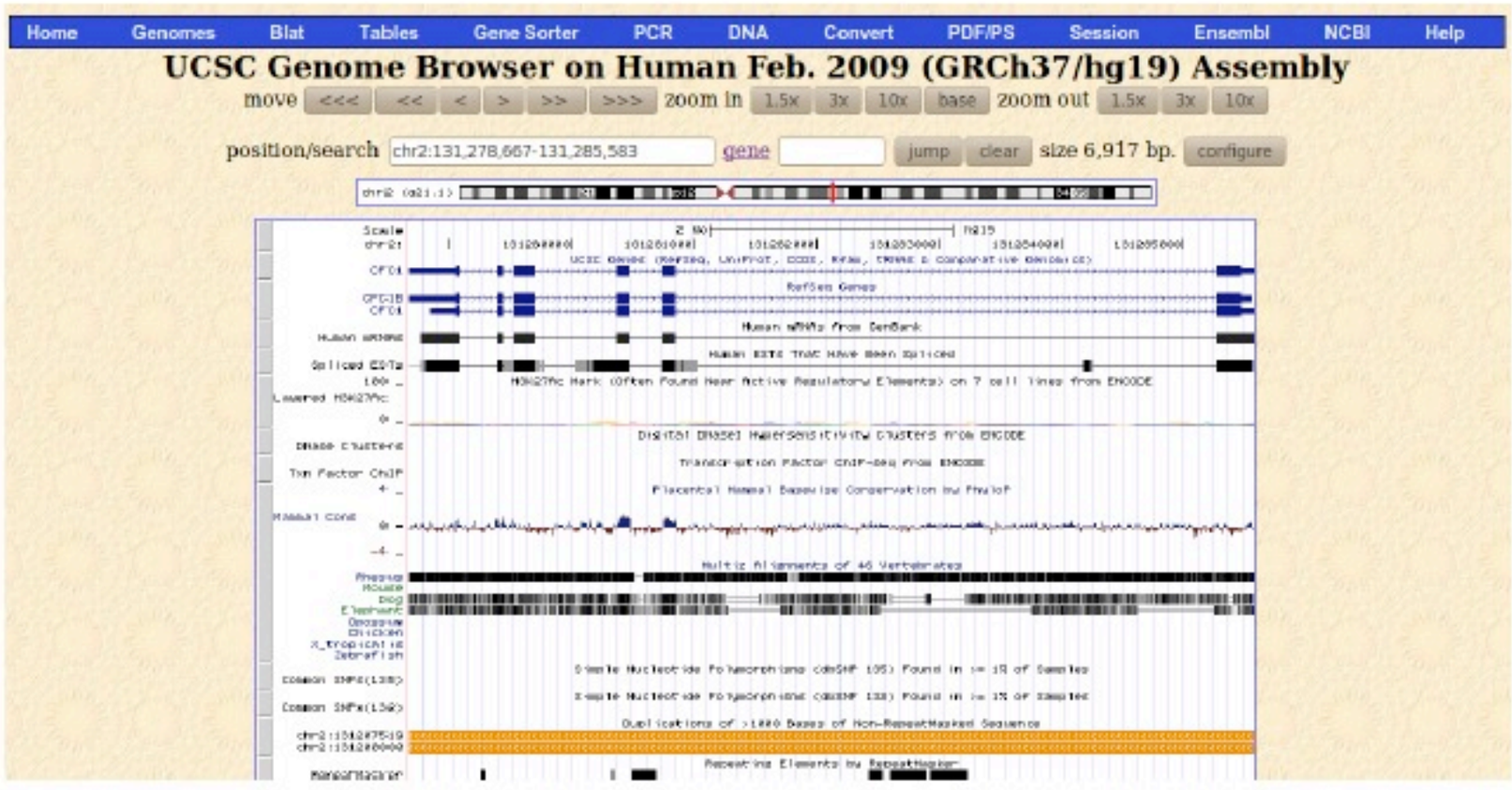
Figure 17: Ranks of peptidases for data sets.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 2, 2009

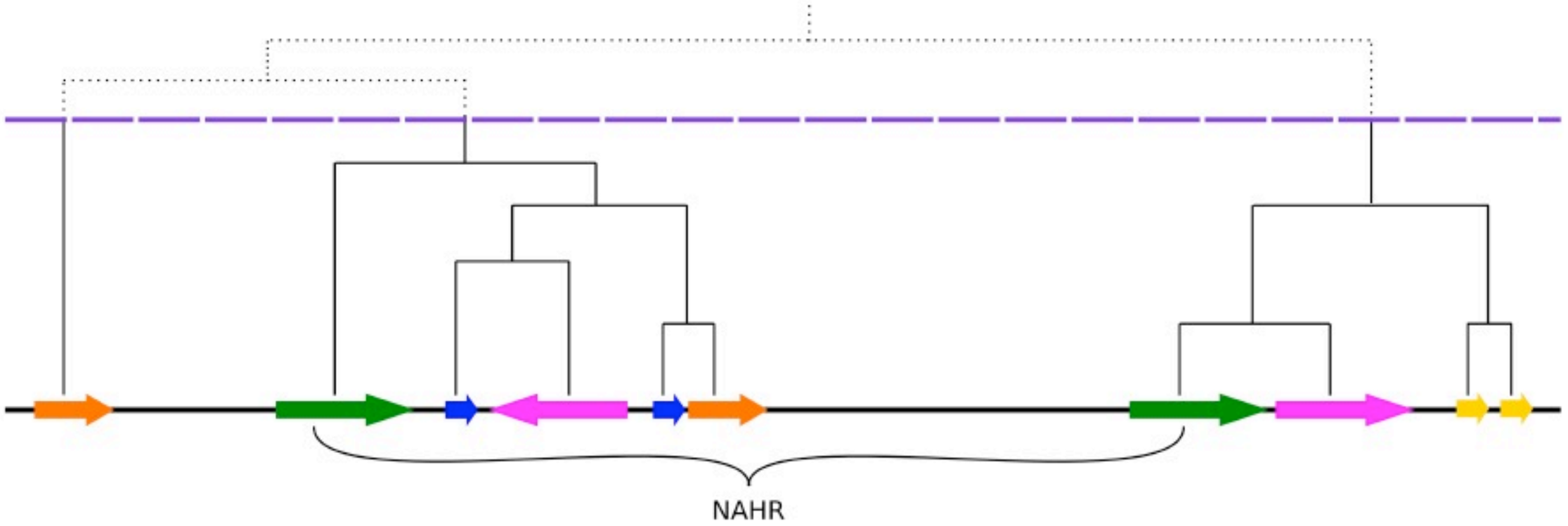
Pp. 395–406
DOI: 10.1089/cmb.2008.22TT

DP-LCRs

- direct paralogous LCRs (from SegDups)
- fraction matching (fractMatch) measure between paralogous elements at least 95%,
- rearrangement length: 50kb-10Mb
- not spanning centromeres
- minimal length of an element: 8 kb

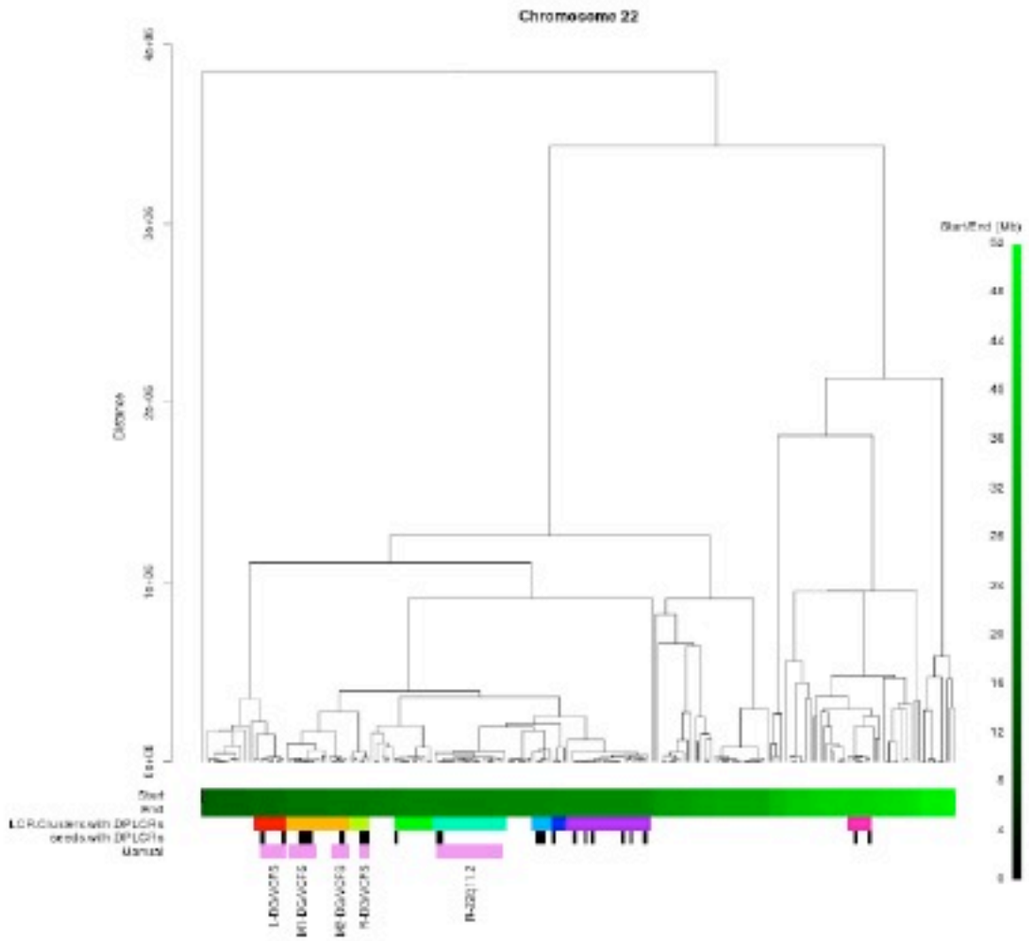


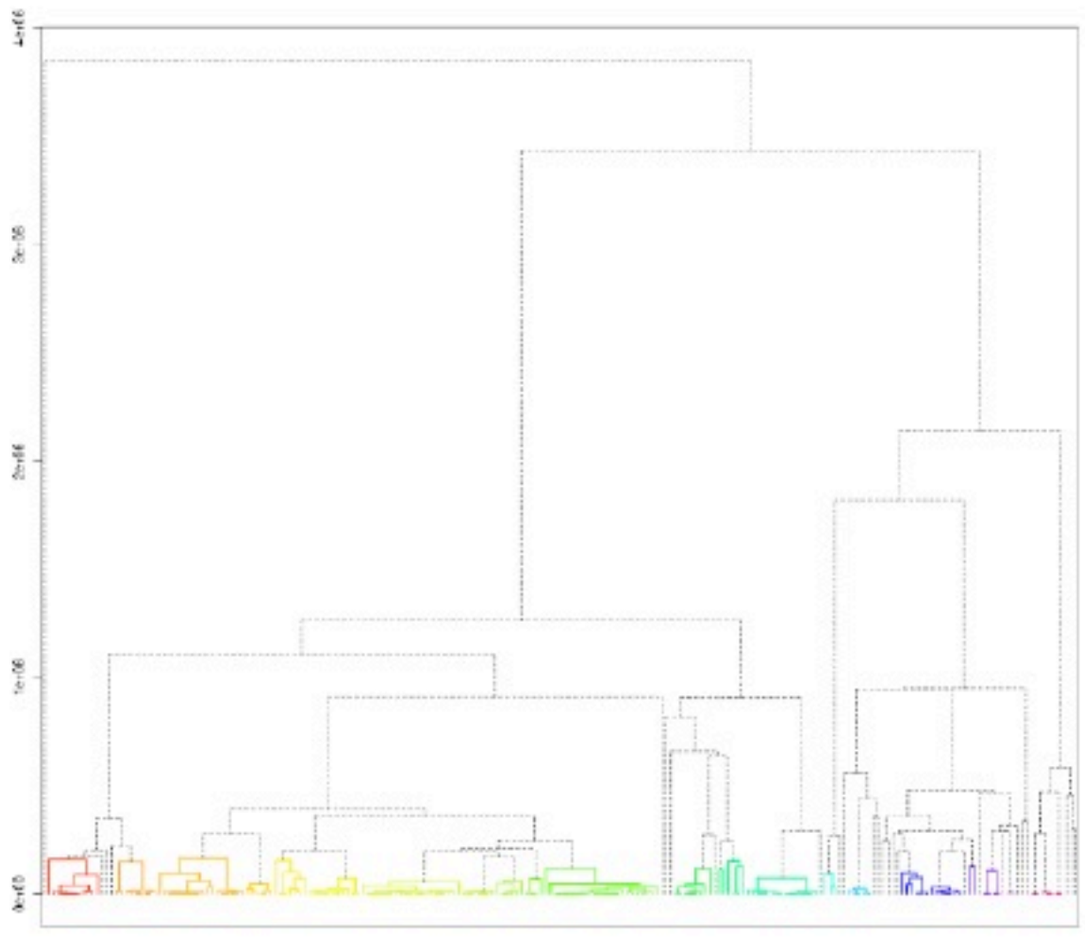
Schematic representation of LCR clustering



- arrows indicate LCR elements and their orientation,
- the same color represents a pair of LCRs
- hierarchical clustering tree is depicted above and horizontal line (violet) shows the height threshold for cutting this tree
- directly oriented paralogous LCRs within the clusters (green) potentially mediate NAHR event.

LCR clusters - chrom22





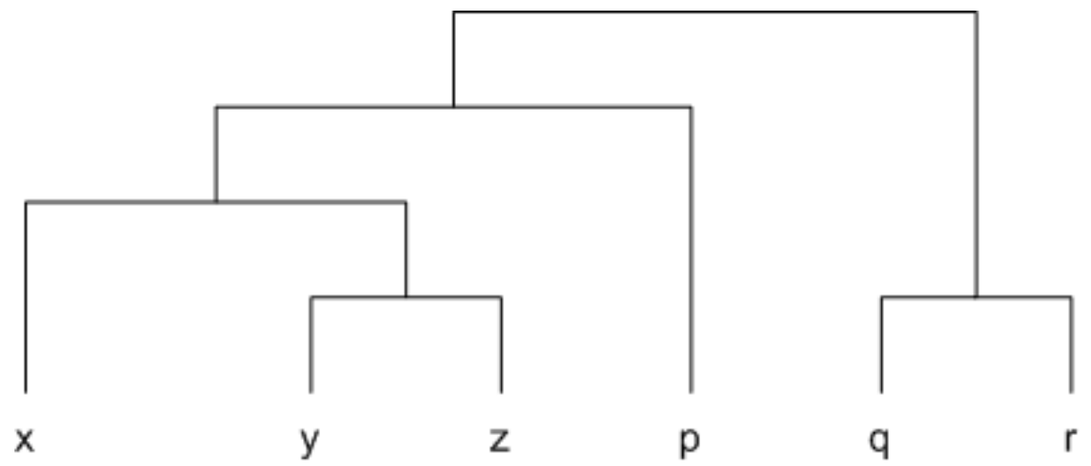
- metody hierarchiczne (np. UPGMA)
- metody grafowe (znajdowanie klik)
- metody wykorzystujące model probabilistyczny (lub prostsze: alg k-średnich, k-median)
- metody spektralne / wykorzystujące łańcuchy Markowa (MCL)
- metody wykorzystujące gęstość (np. DBSCAN)

Definicja 2 *Grupowaniem (podziałem) zbioru* V *nazwiemy rodzinę jego parami rozłącznych niepustych podzbiorów* $\{V_1, V_2, \ldots, V_d\}$*, zwanych też blokami (grupami), dla których* $\cup_{i=1}^{d} V_i = V$*.*

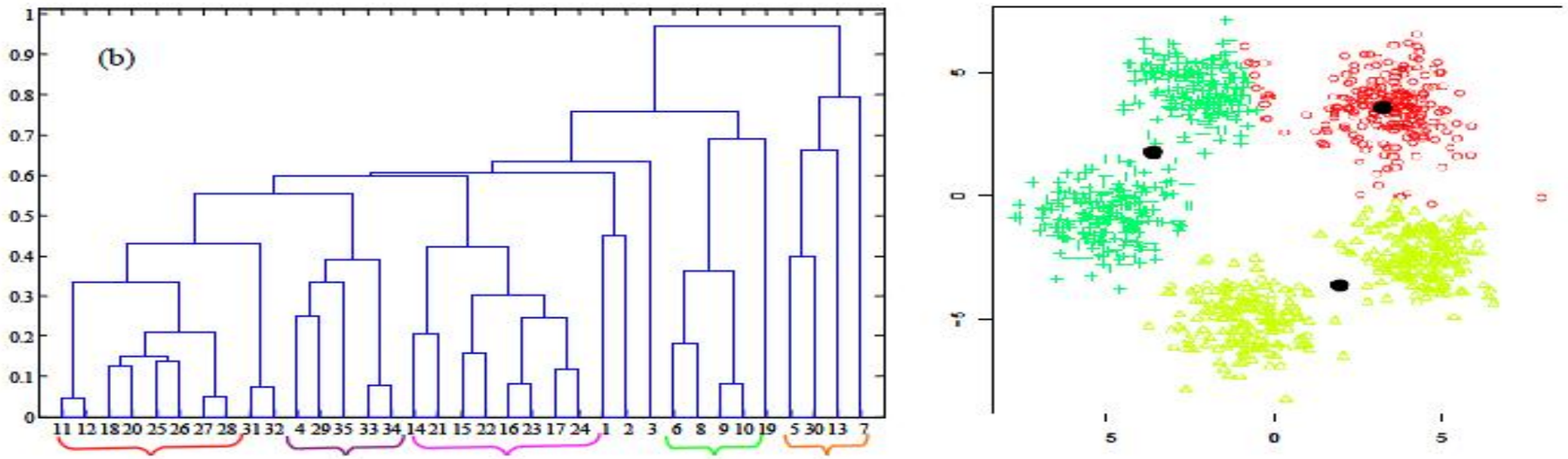
Definicja 3 *Niech* $\theta \in \mathcal{R}$*. Na podstawie macierzy podobieństwa* $S = [s_{ij}]$ *zdefiniujmy graf podobieństwa obiektów* $v \in V$ *jako*

$$G_\theta(V, E_\theta) \quad gdzie \quad (i,j) \in E_\theta \iff s_{ij} > \theta$$

Definicja 4 *Grupowanie hierarchiczne to skończona uporządkowana lista zagnieżdżonych podziałów P_i zbioru V $(i = 1 \ldots n)$. $P_1 = \{\{v_1\}, \{v_2\}, \ldots \{v_n\}\}$ oraz $P_n = \{V\}$. Dla dowolnych i oraz j, $1 \leq i < j \leq n$ bloki podziału P_j możemy otrzymać łącząc pewne bloki podziału P_i.*



- Dla wielu zagadnień ewolucyjnych podejście hierarchiczne jest naturalne
- Wynikiem klastrowania hierarchicznego jest drzewo, wygodne w analizie eksperckiej



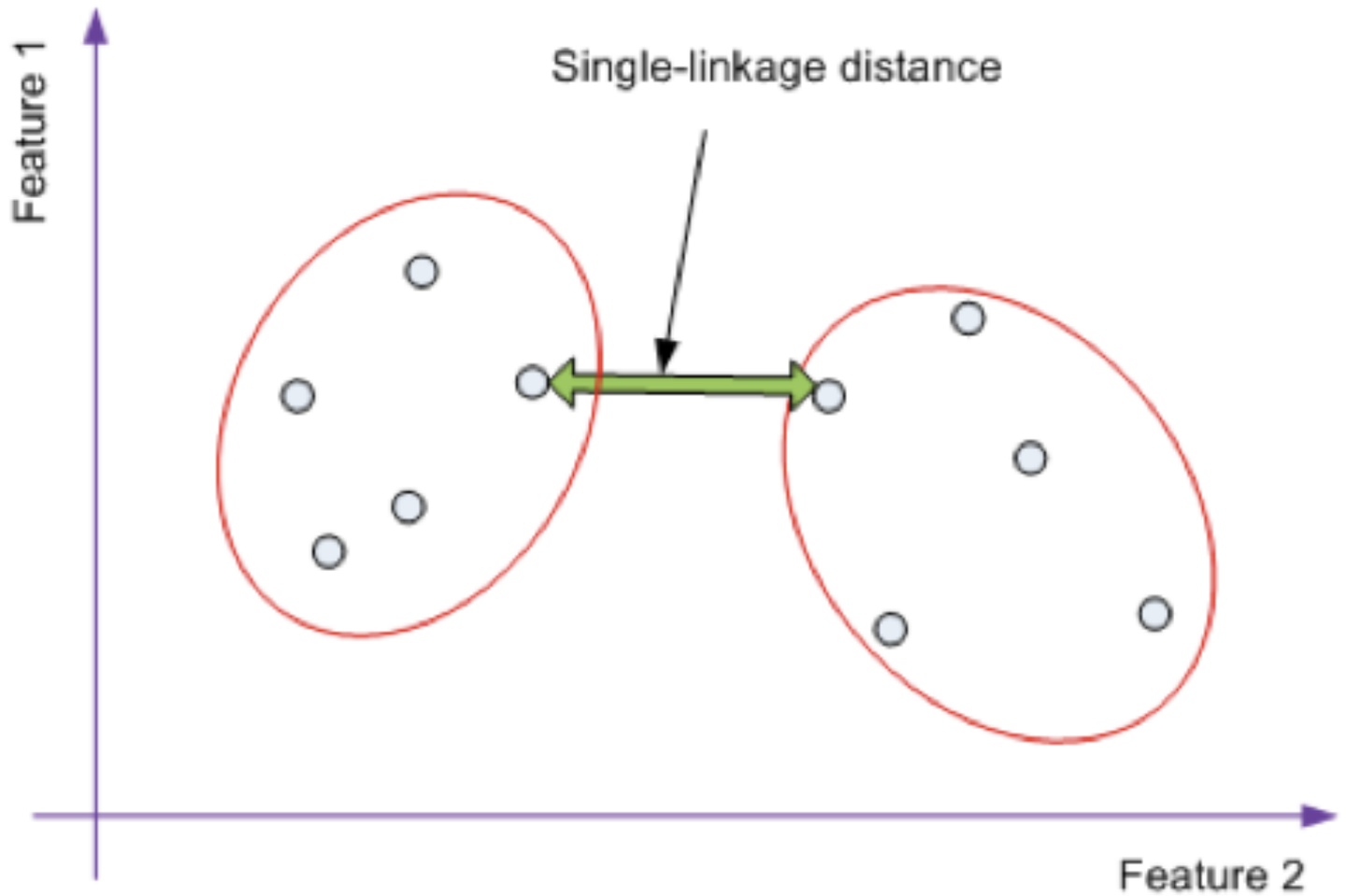
Rysunek: Przykładowe wyniki klastrowania uzyskane metodą hierarchiczną i algorytmem k-średnich

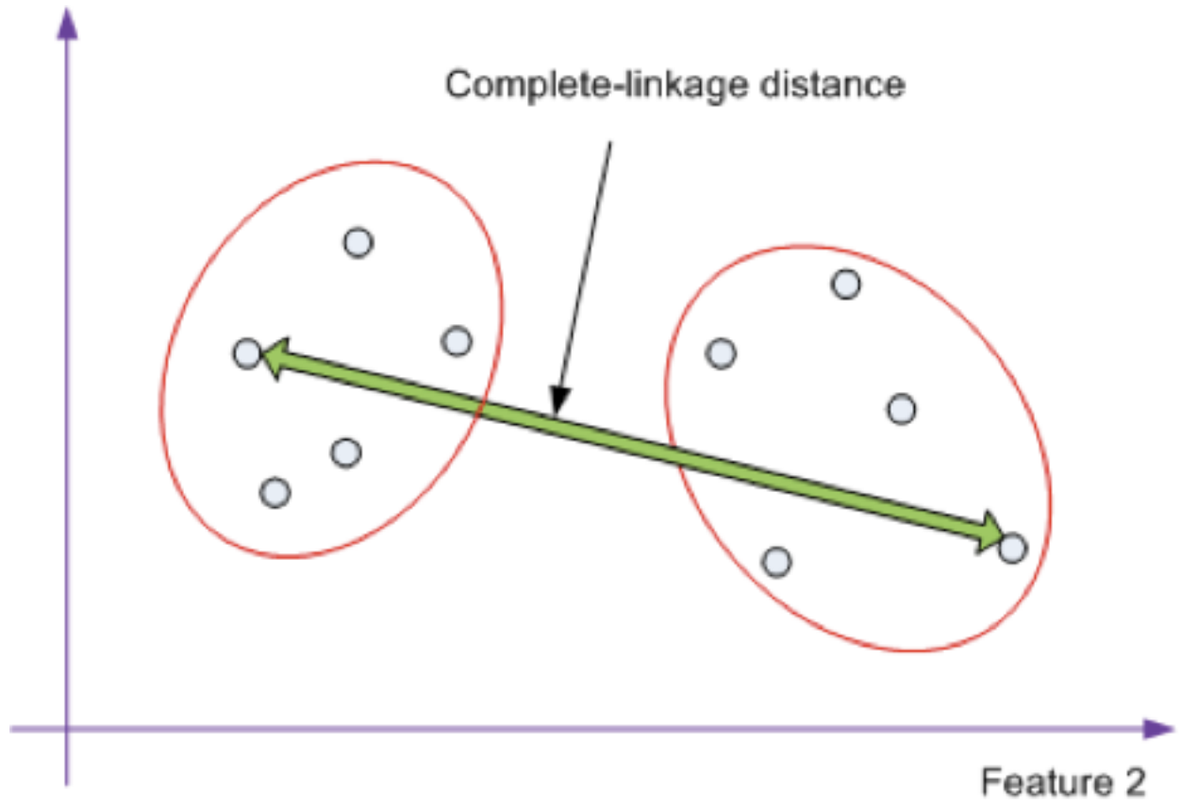
Ogólny schemat działania algorytmów hierarchicznych:

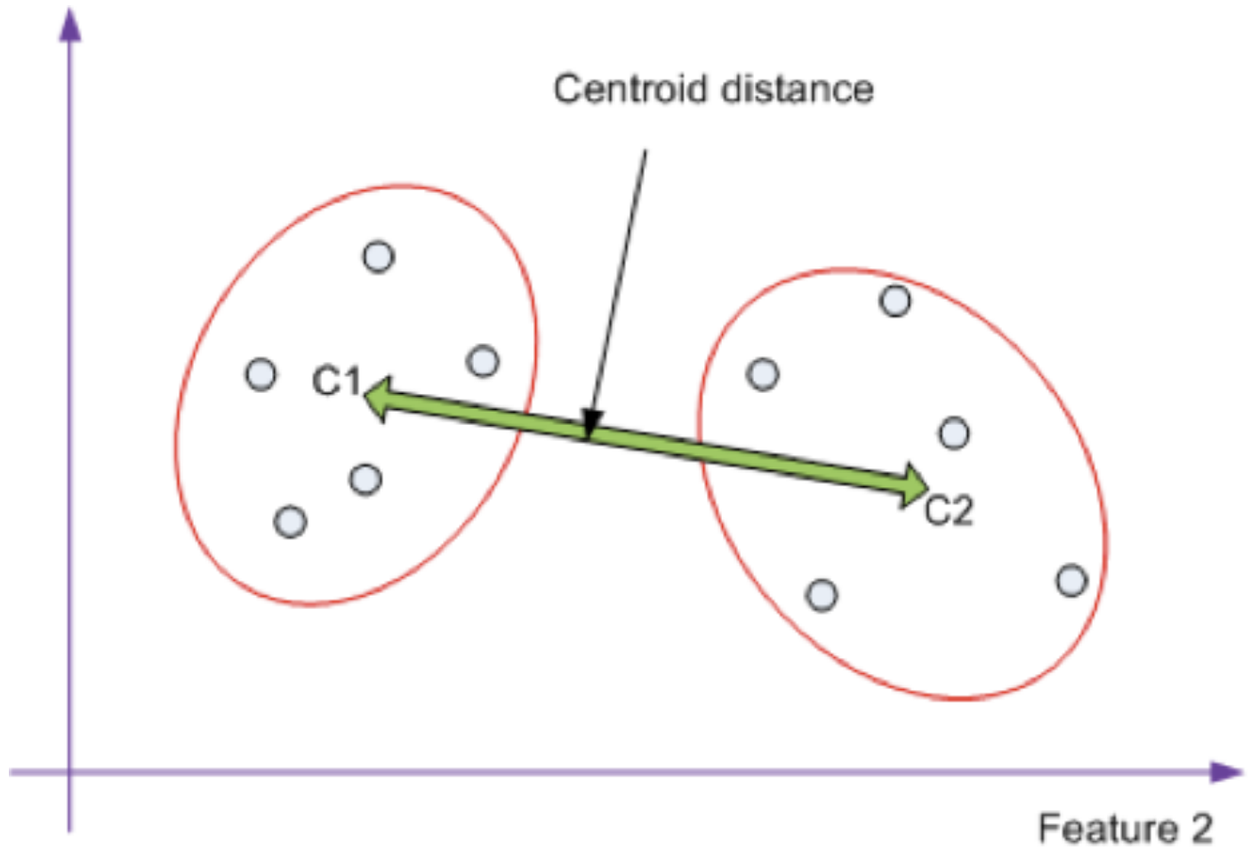
```
Inicjalizacja: każdy obiekt umieść w oddzielnym klastrze

for(i in 1..(n-1))
      Połącz 2 najbliższe klastry
      Odnów macierz odległości pomiędzy klastrami
```

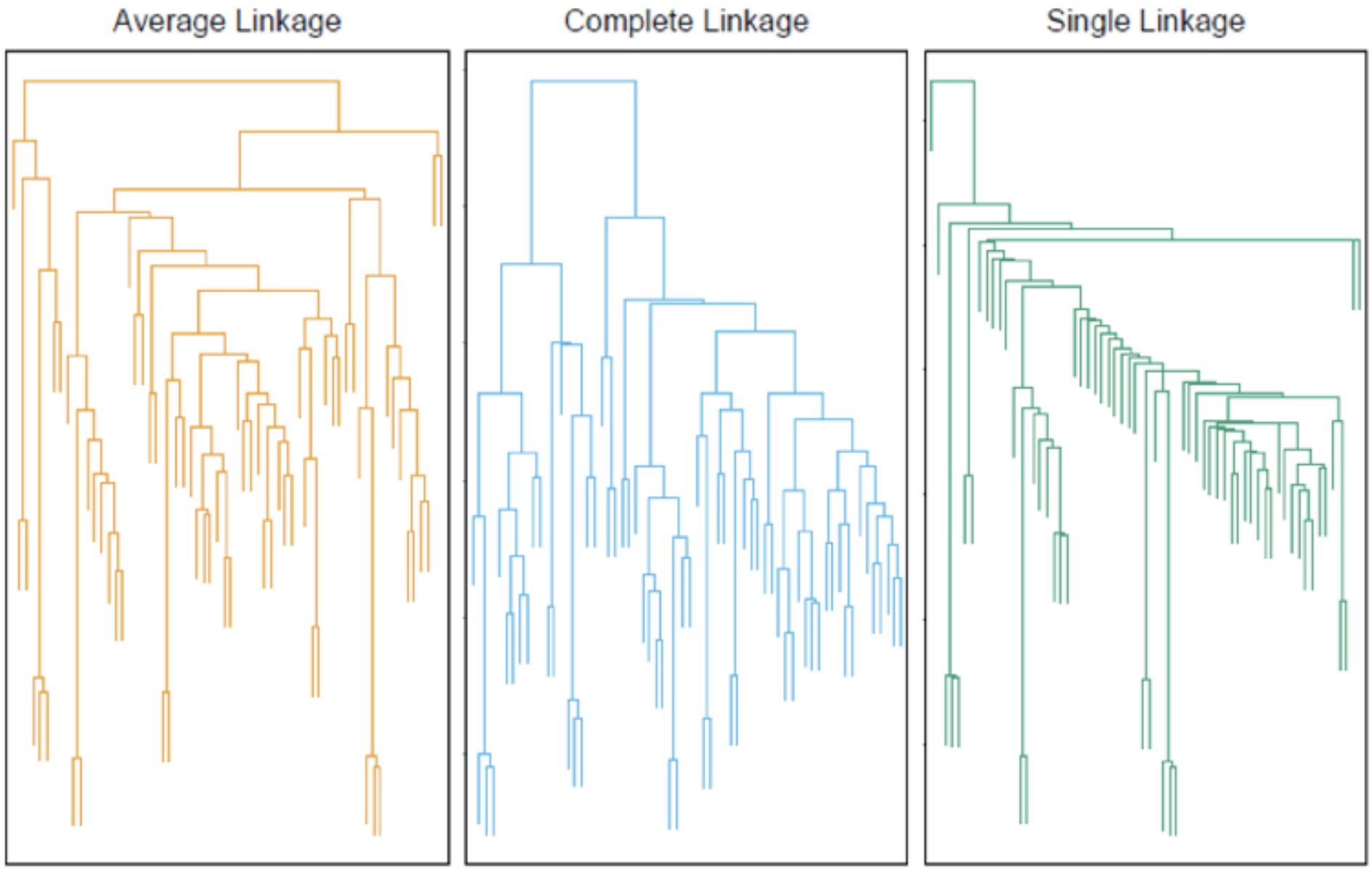
Podstawowe metody liczenia odległości pomiędzy klastrami:

- single-link: $D_{k.ij} = min(D_{k.i}, D_{k.j})$
- complete-link: $D_{k.ij} = max(D_{k.i}, D_{k.j})$
- average-link: $D_{k.ij} = \frac{n_i}{n_i+n_j} D_{k.i} + \frac{n_j}{n_i+n_j} D_{k.j}$


Feature 1
Single-linkage distance
Feature 2


Complete-linkage distance
Feature 2


Centroid distance
C1
C2
Feature 2


Average Linkage
Complete Linkage
Single Linkage

$$D_{ij} = \frac{1}{|C_i| \cdot |C_j|} \Sigma_{p \in C_i} \Sigma_{q \in C_j} D_{pq}$$

Algorithm 1 UPGMA — Unweighted Pair Group Method using Arithmetic averages

1 w wynikowym drzewie przypisz liściom gatunki (wiersze macierzy D); każdy z liści staje się jednoelementowym klastrem;

repeat

2 znajdź C_i oraz C_j o najmniejszej odległości D_{ij};

3 połącz je w jeden klaster C_k zawierający $|C_i| + |C_j|$ sekwencji;

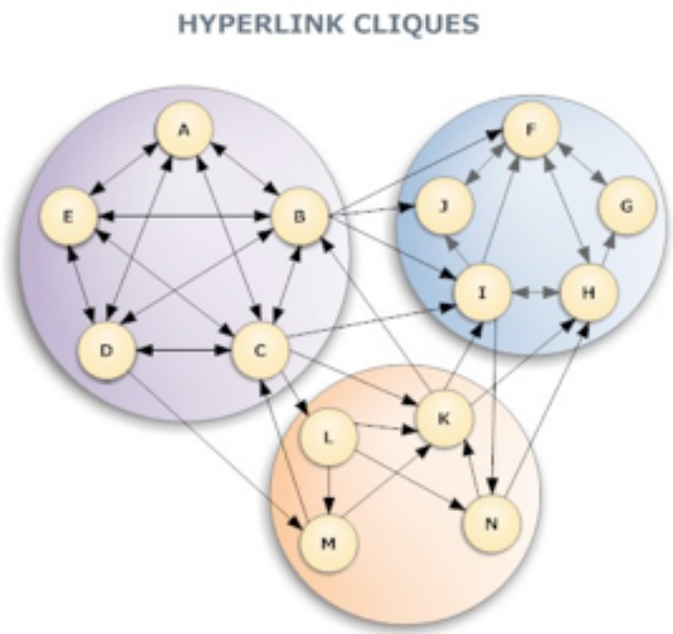
4 dodaj w drzewie wierzchołek odpowiadający nowemu klastrowi krawędziom łączącym go z synami C_i oraz C_j nadaj wagi $\frac{D_{ij}}{2}$;

5 policz odległość pomiędzy nowym klastrem a pozostałymi (z pominięciem klastrów C_i oraz C_j) jako:

$$D_{kl} = (\frac{|C_i|}{|C_i| + |C_j|}) D_{il} + (\frac{|C_j|}{|C_i| + |C_j|}) D_{jl} \quad (3)$$

6 Usuń z macierzy D kolumny i wiersze odpowiadające klastrom C_i oraz C_j oraz dodaj kolumnę i wiersz dla nowego klastra C_k.

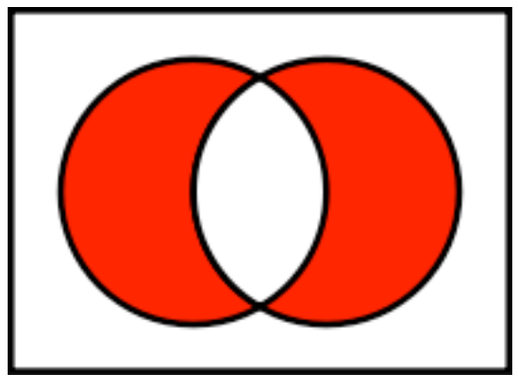
until pozostanie tylko jeden klaster



Definicja 7 Graf klikowy *to suma rozłącznych klik.*

Będziemy zakładać, że prawdziwe (idealne) grupowanie jest reprezentowane przez graf klikowy Q natomiast macierz podobieństwa wyprowadzona z danych eksperymentalnych indukuje graf G, będący zaburzonym grafem klikowym.

metody grafowe



problem 1: edycja grafu klikowego.
INPUT: $G(V, E)$
OUTPUT: $Q(V, F)$ graf klikowy, który minimalizuje wartość $|E \triangle F|$, gdzie $\triangle$ oznacza różnicę symetryczną.

problem 2: uzupełnienie do grafu klikowego.
INPUT: $G(V, E)$
OUTPUT: $Q(V, F)$ graf klikowy, gdzie $E \subseteq F$ minimalizuje$|F \setminus E|$.

Problem uzupełnienia do grafu klikowego można rozwiązać efektywnie znajdując spójne składowe grafu G, a następnie dodając brakujące krawędzie.

złożoność

problem 3: odchudzenie do grafu klikowego.
INPUT: $G(V, E)$
OUTPUT:$Q(V, F)$ graf klikowy, gdzie $F \subseteq E$ minimalizuje$|E \setminus F|$.

- zakładamy, że dane pochodzą z pewnego rozkładu prawdopodobieństwa, np: każdy klaster jest modelowany jako dwuwymiarowy rozkład normalny.

- na podstawie obserwowanych danych estymujemy nieznane parametry modelu: czyli średnie (środki) i wariancje (średnice) klastrów.

Warunkową wartością oczekiwaną zmiennej losowej Y dla ustalonej wartości innej zmiennej $X = x$ nazywamy wartość oczekiwaną **względem rozkładu warunkowego:**

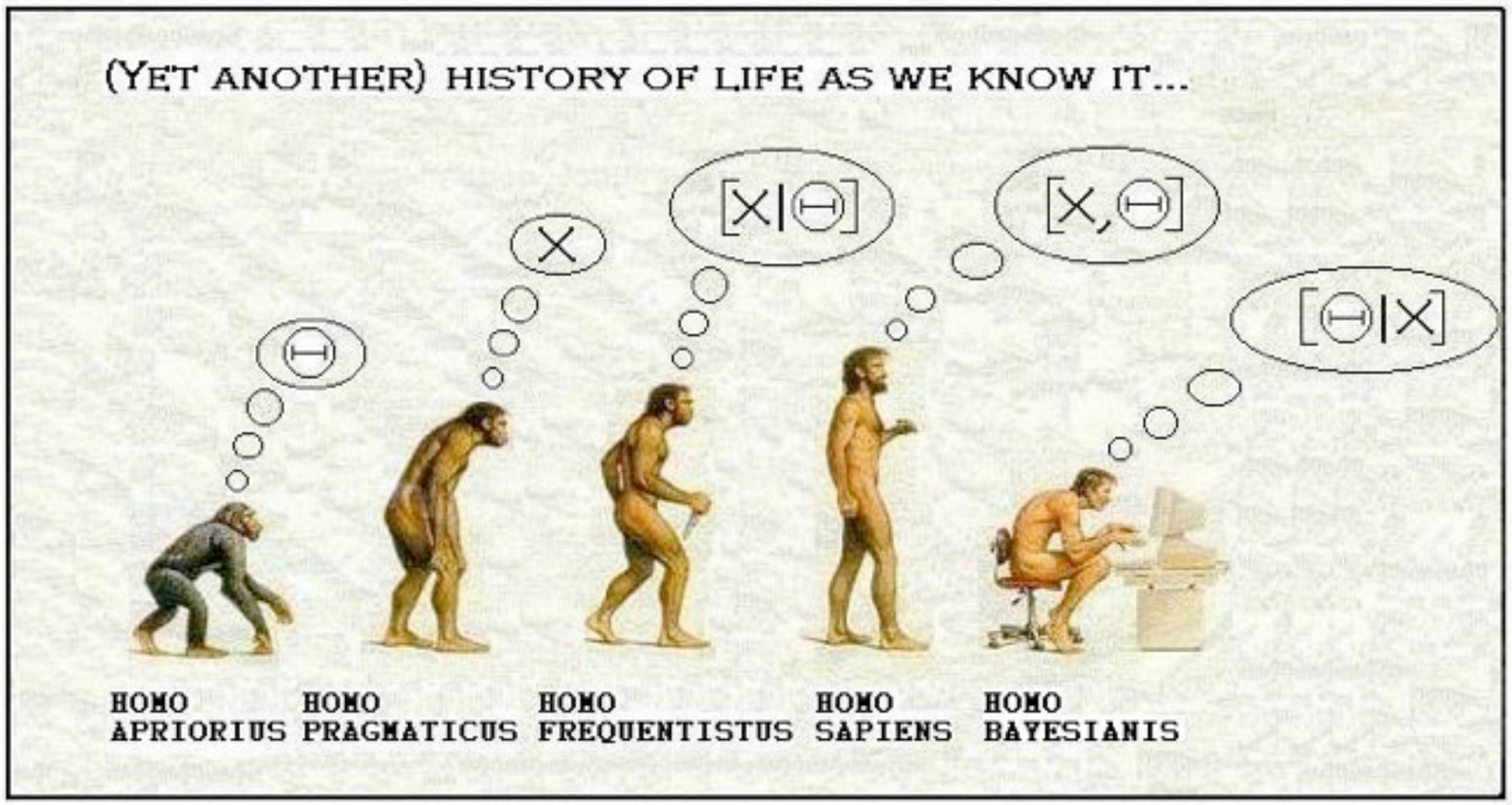
$$E(Y|X=x) = \begin{cases} \sum_k yf(y|x), \text{ jeśli Y jest zmienną dyskretną} \\ \int f(y|x)dy, \text{ jeśli Y jest zmienną ciągłą} \end{cases}$$

gdzie $f(y|x)$ jest warunkową gestością, czyli:

$$f(y|x) = \frac{f(x,y)}{f_X(x)}$$

Wzór Bayesa. Niech zdarzenia $A_1, A_2, \ldots$ spełniają następujące warunki: $A_i \cap A_j = \emptyset$ dla $i \neq j$, $A_1 \cup A_2 \cup \ldots = \Omega$ oraz $P(A_i) > 0 \quad \forall i$. Wtedy dla dowolnego zdarzenia B, jeśli $P(B) > 0$, zachodzi:

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_k P(B|A_k)P(A_k)}$$



$$\boldsymbol{x} = (x_1, \ldots, x_p)$$

pochodząca z wielowymiarowego rozkładu normalnego ze średnią:

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$$

i macierzą kowariancji:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \ldots & \sigma_p^2 \end{bmatrix}$$

ma gęstość:

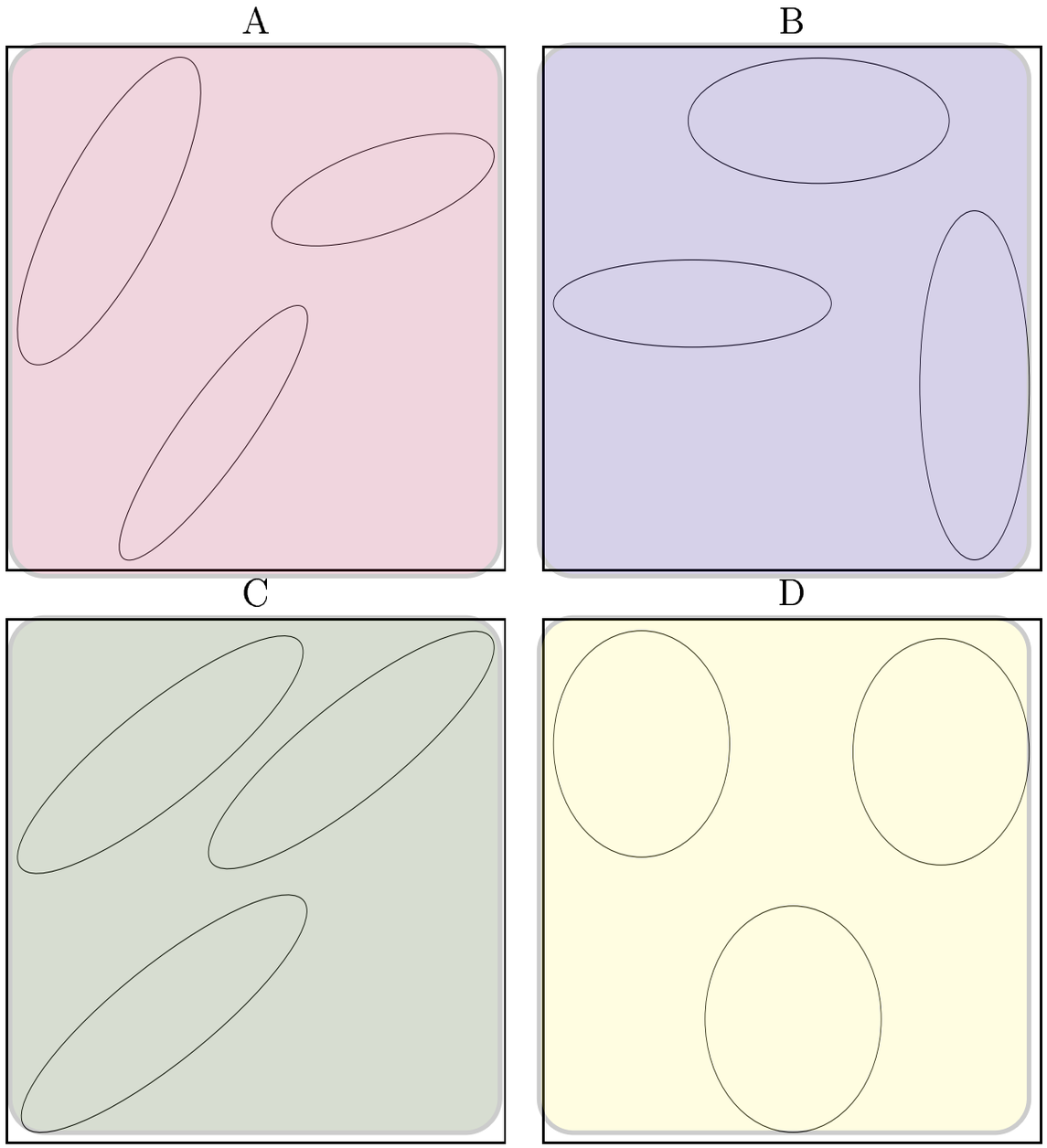
$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p \cdot |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

Powracając do zadania grupowania, zakładamy że gęstość obiektów w grupie k wynosi:

$$f_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p \cdot |\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}$$

Często przyjmuje się dodatkowe założenie, że wszystkie grupy mają tę samą orientację oraz rozrzut, czyli $\forall k \;\; \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$. Lub jeszcze bardziej restrykcyjnie, że wszystkie grupy są sferyczne, czyli $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$.

macierze kowariancji



Rysunek 1: A: różne macierze kowariancji; B: diagonalne macierze kowariancji; C: identyczne macierze kowariancji; D: identyczne, sferyczne macierze kowariancji

Gęstość dla wszystkich obiektów jest *mieszaniną* gęstości dla poszczególnych grup, czyli:

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\boldsymbol{x})$$

Załóżmy na chwilę, że znamy „zapomniane" etykiety grup S_i $(i = 1, \ldots k)$. Wtedy *wiarygodność* naszego modelu, czyli prawdopodobieństwo, że obserwacje $x_1, \ldots x_n$ pochodzą z odpowiednich populacji wynosi:

$$\prod_{i=1}^{n} \pi_{S_i} f_{\boldsymbol{\mu}_{S_i}, \boldsymbol{\Sigma}_{S_i}}(\boldsymbol{x}_i) \tag{1}$$

Expectation-Maximization

Oznaczmy przez θ wektor wszystkich parametrów w naszym zadaniu:

$$\theta = (\pi_1, \ldots \pi_K, \boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots \boldsymbol{\Sigma}_K)$$

Logarytmując wiarygodność zdefiniowaną równaniem (1) otrzymujemy:

$$L(\theta) = \sum_{i=1}^{n} \log \pi_{S_i} + \log f_{\boldsymbol{\mu}_{S_i}, \boldsymbol{\Sigma}_{S_i}}(\boldsymbol{x}_i)$$

$$\prod_{i=1}^{n} \pi_{S_i} f_{\boldsymbol{\mu}_{S_i}, \boldsymbol{\Sigma}_{S_i}}(\boldsymbol{x}_i)$$

Expectation-Maximization

Niestety w rzeczywistości nie znamy etykiet S_i, zaś algorytm EM próbuje zmaksymalizować warunkową wartość oczekiwaną $L(\theta)$. W kroku 'E' musimy policzyć warunkową wartość oczekiwaną dla $L(\theta)$ pod warunkiem obserwowanych danych używając parametrów $\hat{\theta}^{(h-1)}$ z poprzedniego kroku. Ze wzoru Bayesa liczymy rozkład warunkowy dla S_i:

$$w_{ik}^{(h)} := \mathcal{P}_{\hat{\theta}^{(h-1)}}(S_i = k) = \frac{\hat{\pi}_k^{(h-1)} f_{\boldsymbol{\mu}_k^{(h-1)}, \boldsymbol{\Sigma}_k^{(h-1)}}(\boldsymbol{x}_i)}{\sum_{\xi=1}^{K} \hat{\pi}_\xi^{(h-1)} f_{\boldsymbol{\mu}_\xi^{(h-1)}, \boldsymbol{\Sigma}_\xi^{(h-1)}}(\boldsymbol{x}_i)} \tag{2}$$

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_k P(B|A_k)P(A_k)}$$

Expectation-Maximization

Zatem warunkowa wartość oczekiwana wynosi:

$$\mathcal{E}_{\hat{\theta}^{(h-1)}}(L(\theta)|x_1,\ldots,x_n)) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(h)} \log \pi_k + w_{ik}^{(h)} \log f_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}) \quad (3)$$

W kroku 'M' naszym zadaniem jest policzenie $\hat{\theta}^{(h)}$, które maksymalizuje warunkową wartość oczekiwaną 3.

E: Estymuj warunkowy rozkład S_i pod warunkiem obserwowanych danych oraz parametrów rozkładów z poprzedniego kroku $\hat{\boldsymbol{\mu}}_k^{(h-1)}$ oraz $\hat{\boldsymbol{\Sigma}}_k^{(h-1)}$ dla $k = 1, 2, \ldots K$. Uzyskując wagi $w_{ik}^{(h)}$ ze wzoru (2).

M: Popraw oczacowania parametrów następująco:

$$\hat{\pi}_k^{(h)} := \frac{1}{n} \sum_{i=1}^{n} w_{ik}^{(h)},$$

$$\hat{\boldsymbol{\mu}}_k^{(h)} := \frac{1}{\sum_{i=1}^{n} w_{ik}^{(h)}} \sum_{i=1}^{n} w_{ik}^{(h)} \boldsymbol{x}_i$$

$$\hat{\boldsymbol{\Sigma}}_k^{(h)} := \frac{1}{\sum_{i=1}^{n} w_{ik}^{(h)}} \sum_{i=1}^{n} w_{ik}^{(h)} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(h)})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k^{(h)})^T.$$

nie zakładamy, że obserwacje pochodzą z różnych grup (w których dane są modelowane rozkładem normalnym ze średnią $\boldsymbol{\mu}_k$ oraz kowariancją $\boldsymbol{\Sigma}_k$ ^{2}). Różnica polega na tym, że szukany podział (grupowanie): $\Omega = \Omega_1 \cup \ldots \cup \Omega_K$ nie jest modelowany poprzez ukryte zmienne („zapomniane" etykiety grup) tylko jest uważany za kolejny parametr modelu. Jeśli założymy, że wszystkie grupy posiadają identyczne i sferyczne macierze kowariancji otrzymamy

algorytm k-średnich

W modelach z użyciem mieszanin rozkładów normalnych średnie dla grup $\boldsymbol{\mu}_k$ były poprawiane z pomocą ważonych średnich obserwacji x_i. Wagi w_{ik} odpowiadały warunkowym prawdopodobieństwom zdarzenia, że i-ta obserwacja pochodzi z grupy k. Te prawdopodobieństwa były wysokie dla obserwacji leżących w pobliżu środka grupy czyli wartości $\boldsymbol{\mu}_k$.

Na tym pomyśle opiera się algorytm k-średnich [4], który może być postrzegany jako pewne uproszczenie algorytmu EM. Zamiast wyliczać wyrafinowane wagi przypisujemy wszystkie obserwacje do najbliższej grupy, tzn takiej której aktualny środek $\hat{\boldsymbol{\mu}}_k$ położony jest najbliżej. [3] Następnie poprawiamy położenie środków licząc średnią arytmetyczną wszystkich obserwacji przypisanych do danej grupy.

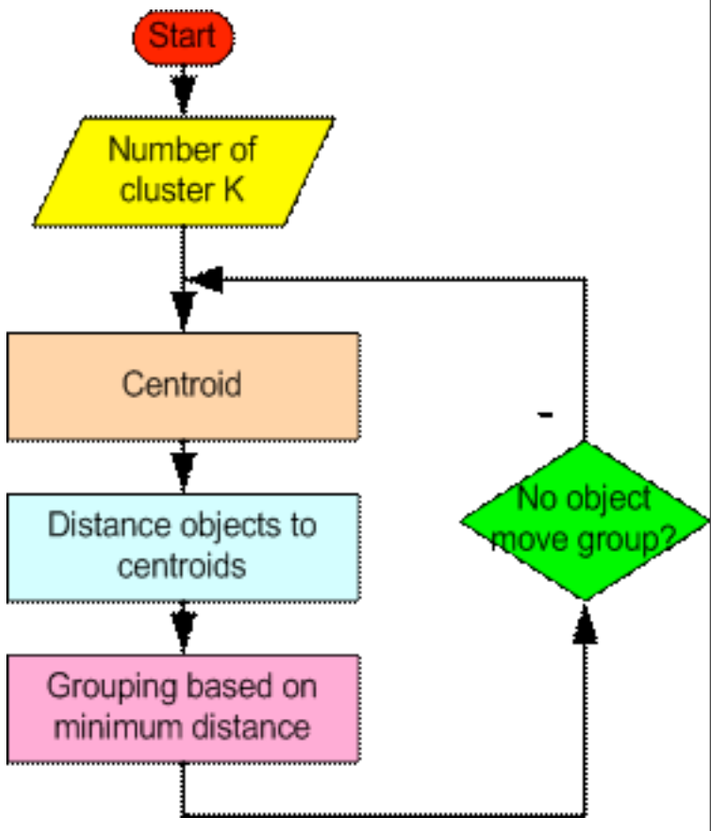
algorytm k-średnich

Dla odpowiednio dobranych wartości początkowych $\hat{\boldsymbol{\mu}}_1^{(0)}, \ldots, \hat{\boldsymbol{\mu}}_K^{(0)}$ algorytm k-średnich powtarza następujące dwa kroki dla $h = 1, 2, \ldots$:

- Dla każdej obserwacji $\boldsymbol{x}_i$ znajdź najbliższy środek $\hat{\boldsymbol{\mu}}_k^{(h-1)}$ i przypisz tę obserwację do grupy o tym środku.

- Popraw środki grup przez policzenie średniej w każdej grupie:

$$\hat{\boldsymbol{\mu}}_k^{(h)} := \frac{1}{\text{rozmiar grupy } k} \sum_{\text{obserwacje } \boldsymbol{x}_i \text{ należące do grupy } k} \boldsymbol{x}_i$$



Algorytm k-średnich używa odległości euklidesowej. Jako jego uogólnienie zaproponowano algorytm k-median, który używa dowolnej miary odległości. [4] Często przyjmujemy, że $d_{ij=|\boldsymbol{x}_i-\boldsymbol{x}_j|}$ oraz wymagamy, żeby środki grup były tożsame z pewną obserwacją należącą do danej grupy, czyli $\hat{\boldsymbol{\mu}}_k = \boldsymbol{x}_{c_k}$ dla pewnego c_k.

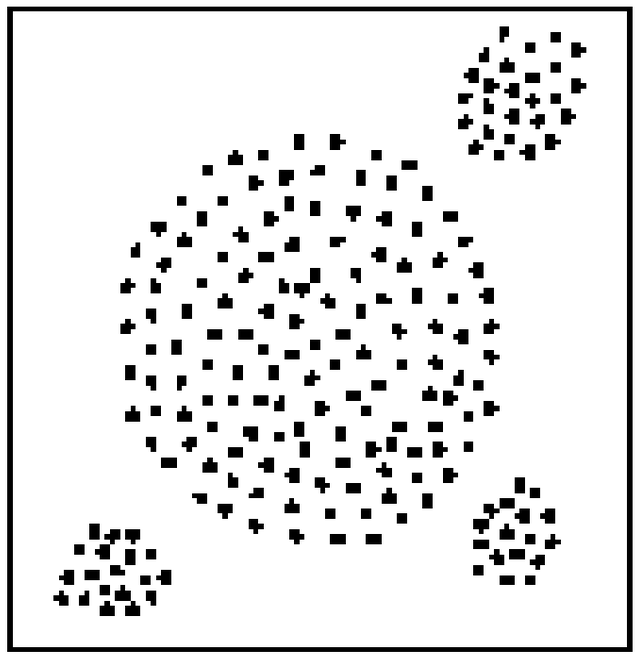
algorytm k-median

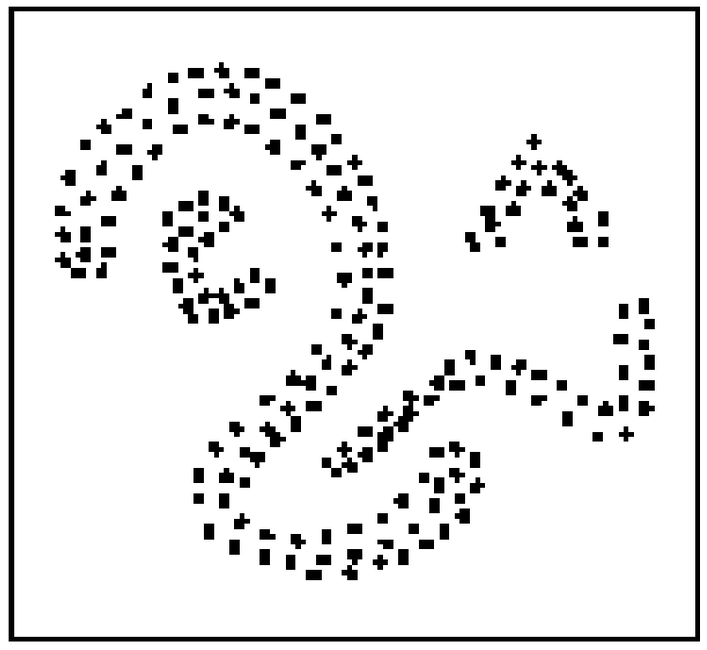
Algorytm powtarza następujące dwa kroki, aż grupowanie przestanie się zmieniać:

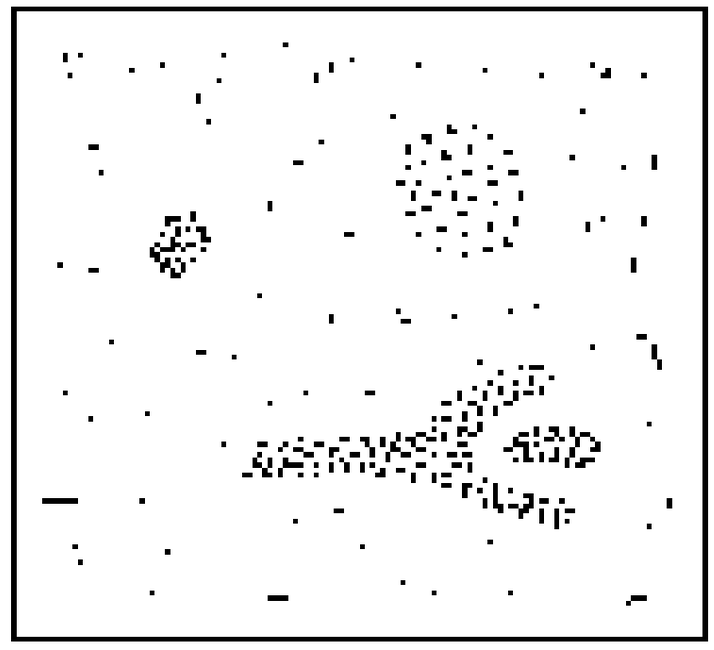
- Każdą obserwację $\boldsymbol{x}_i$ przypisz do grupy k, której środek $\hat{\boldsymbol{\mu}}_k = \boldsymbol{x}_{c_k}$ minimalizuje d_{ic_k}.

- Dla każdej grupy k (gdzie $k = 1, \ldots K$) znajdź centroid, czyli obserwację $\boldsymbol{x}_{c_k}$, która minimalizuje:

$$\sum_{\text{obserwacje } \boldsymbol{x}_i \text{ należące do grupy } k} d_{ic_k}$$

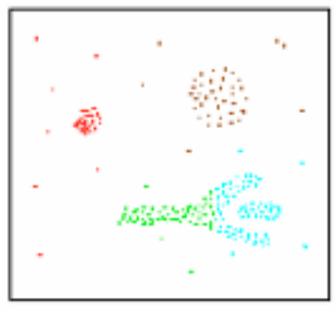




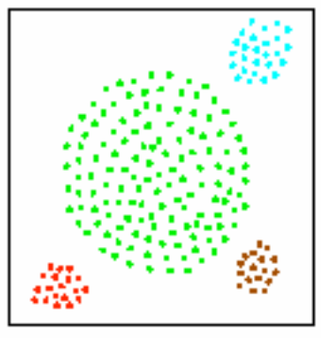


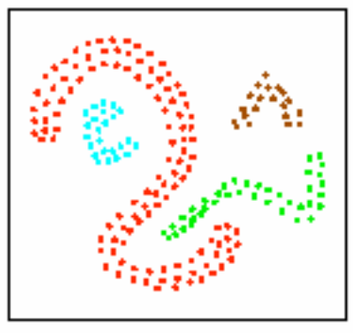
database 1

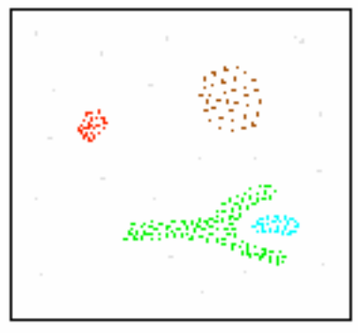
database 2



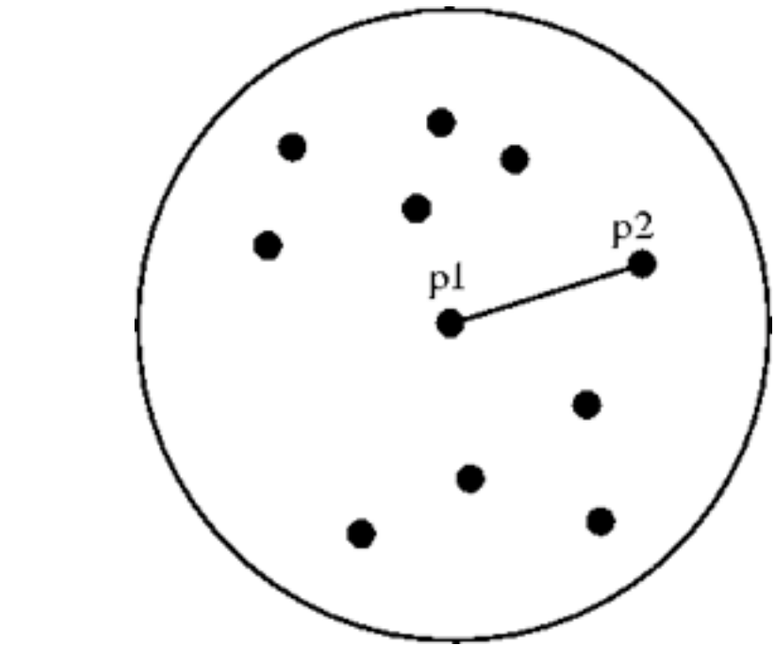
database 3

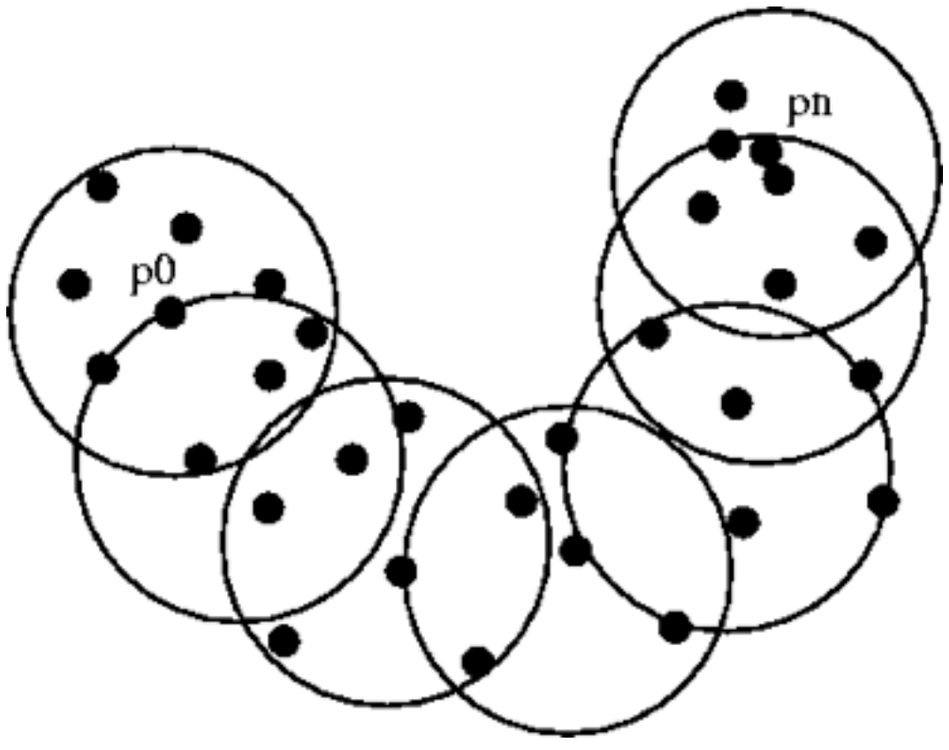






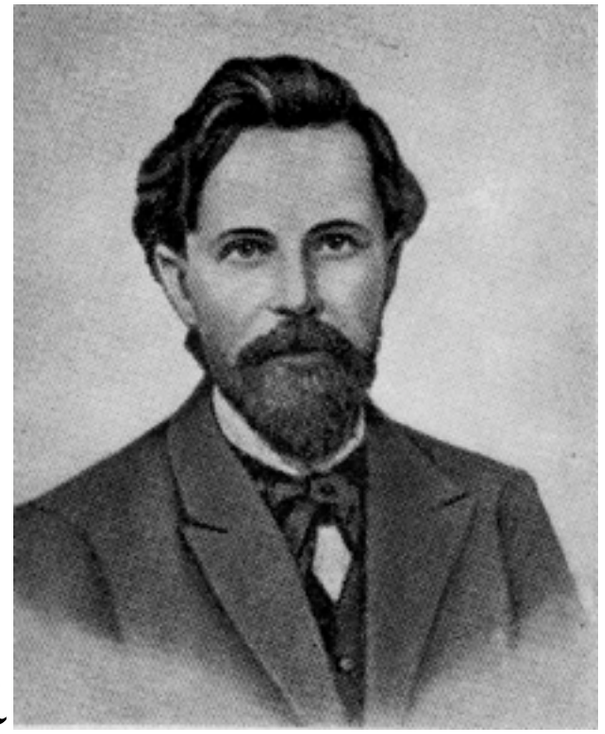
DBSCAN






```
DBSCAN(D, eps, MinPts)
   C = 0
   for each unvisited point P in dataset
D
      mark P as visited
      N = getNeighbors (P, eps)
      if sizeof(N) < MinPts
         mark P as NOISE
      else
         C = next cluster
         expandCluster(P, N, C, eps,
MinPts)

expandCluster(P, N, C, eps, MinPts)
   add P to cluster C
   for each point P' in N
      if P' is not visited
         mark P' as visited
         N' = getNeighbors(P', eps)
         if sizeof(N') >= MinPts
            N = N joined with N'
      if P' is not yet member of any
cluster
         add P' to cluster C
```

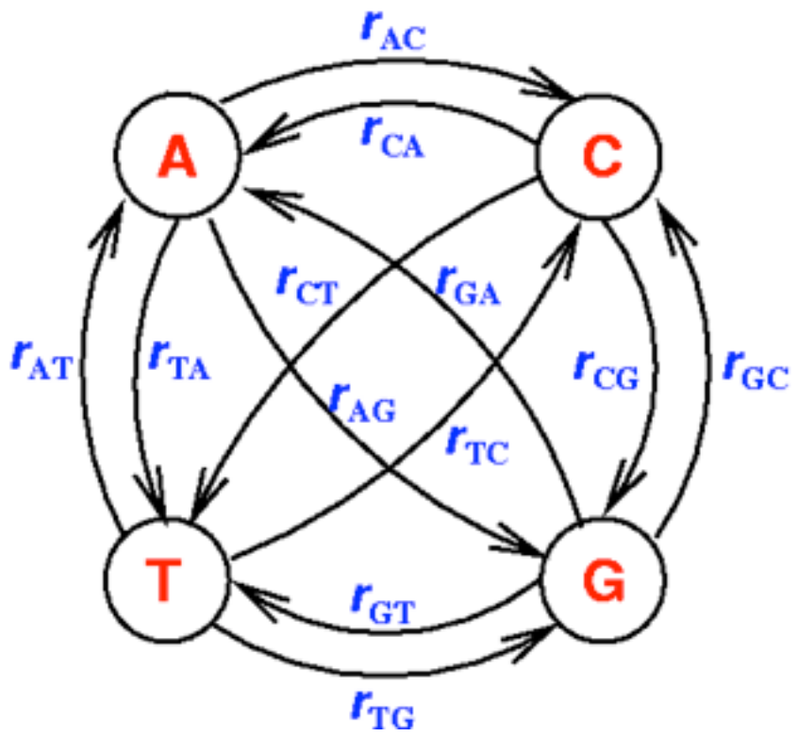
łańcuch Markowa

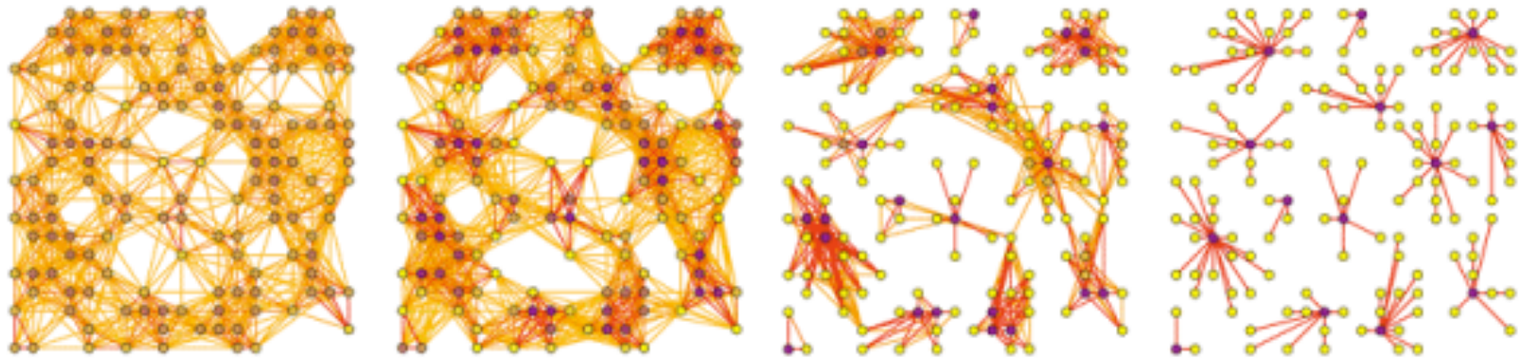


$Q \neq \emptyset,$ zbiór stanów

$M = (p_{k,l})_{k,l \in Q},$ stochastyczna macierz przejść

$$\sum_{l \in Q} p_{k,l} = 1.$$

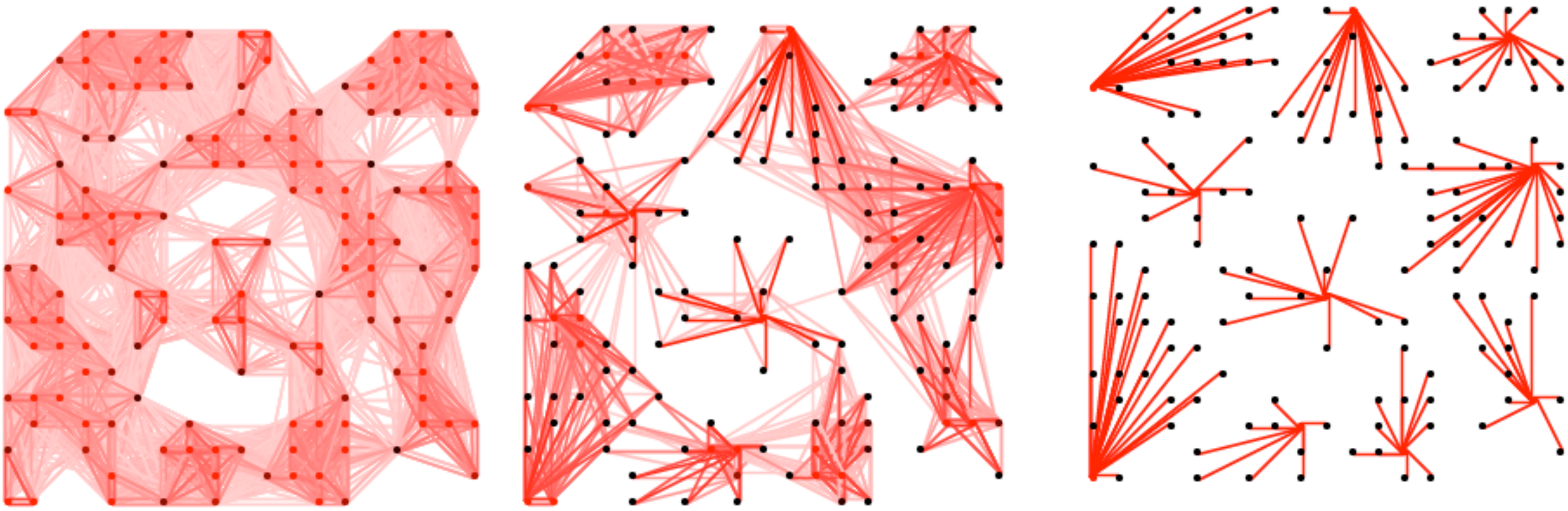




G is a graph
add loops to **G**
set Γ to some value
M_1 to be the matrix of random walks on **G**
while (change)
{ **M_2** = **M_1** * **M_1** # expansion
M_1 = Γ(**M_2**) # inflation
change = difference(**M_1**, **M_2**) }
set CLUSTERING as the components of **M_1**

$$\Gamma_r(M_{ij}) = M_{ij}^r / \Sigma_{r,j}(M)$$

Markov cluster algorithm



http://www.micans.org/mcl/

porównujemy

$$\bigcup_{k=1}^{K} C_k = V, \quad \bigcup_{k=1}^{K'} C'_k = V.$$

N_{11} liczba par elementów, które w obu grupowaniach znajdują się w jednej grupie;

N_{00} liczba par elementów, które w obu grupowaniach znajdują się w różnych grupach;

N_{10} liczba par elementów, które w grupowaniu C znajdują się w tej samej grupie, natomiast w C' w różnych;

N_{01} liczba par elementów, które w grupowaniu C' znajdują się w tej samej grupie, natomiast w C w różnych.

Oczywiście suma tych czterech liczb jest liczbą wszystkich par elementów w zbiorze V.

kryterium Wallace'a

$n_k = |C_k|$, $n'_k = |C'_k|$ oraz $n = |V|$

$$\mathcal{W}_I(C, C') = \frac{N_{11}}{\sum_k n_k(n_k - 1)/2},$$

$$\mathcal{W}_{II}(C, C') = \frac{N_{11}}{\sum_{k'} n_{k'}(n_{k'} - 1)/2}.$$

$$\mathcal{F}(C, C') = \sqrt{\mathcal{W}_I(C, C')\mathcal{W}_{II}(C, C')}.$$

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k), \qquad P(k) = n_k/n$$

Niech $P(k,k') = |C_k \cap C'_{k'}|/n$ oznacza prawdopodobieństwo, że element ze zbioru V należy do grupy C_k w grupowaniu C oraz do grupy $C'_{k'}$ w grupowaniu C'. Wzajemna informacja grupowań definiowana jest jako wzajemna informacja (miara zależności) zmiennych losowych indukowanych przez grupowania C i C':

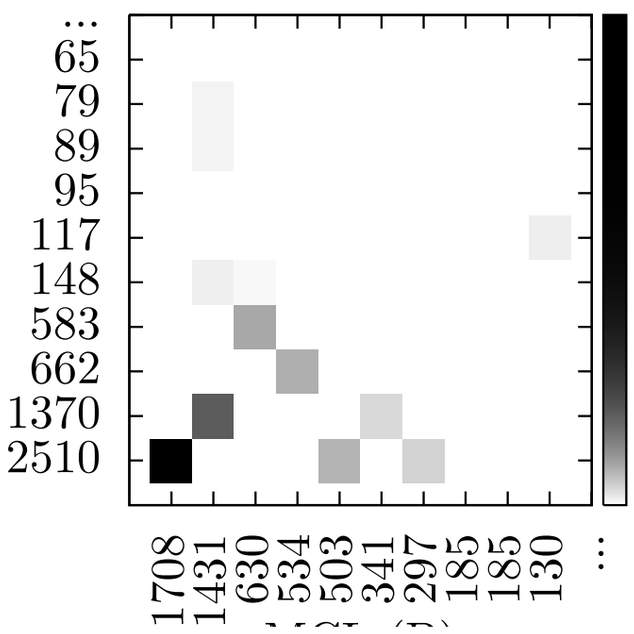
$$I(C,C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k,k') \log \frac{P(k,k')}{P(k)P(k')}.$$

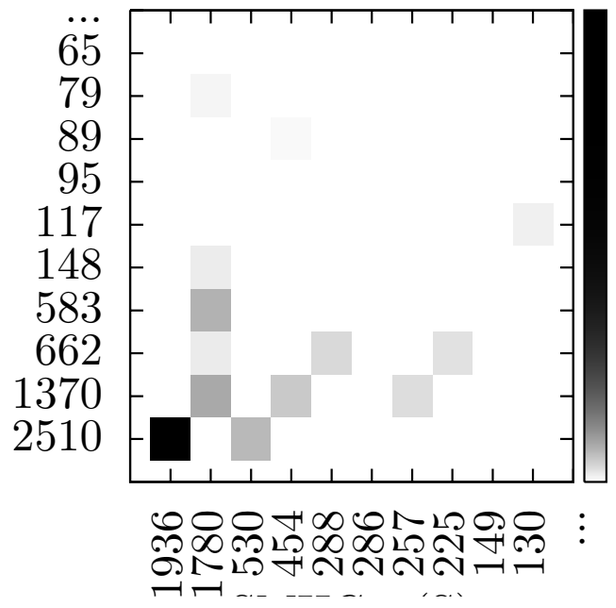
Wariancja informacji pomiędzy dwoma grupowaniami jest definiowana jako:

$$VI(C, C') = H(C) + H(C') - 2I(C, C').$$

Zdefiniujmy macierz koincydencji (*ang. contingency table*) grupowań C i C'. Element $m_{kk'}$ macierzy koincydencji definiujemy jako liczbę wspólnych elementów w zbiorach C_k i $C'_{k'}$

$$m_{kk'} = |C_k \cap C'_{k'}|.$$





- Porównywanie optymalizowanej funkcji celu
- Porównywanie wyników algorytmów

Odległość $D(F_1, F_2)$ pomiędzy algorytmami klastrowania na zbiorze danych X można przybliżać jako odległość $d(\cdot\,,\cdot)$ pomiędzy podziałami P_1^X oraz P_2^X zbioru X na klastry.

Odległość dwóch podziałów zbioru X będziemy liczyć ze wzoru (jest to tak zwany *Rand index*):

$$d(P_1^X, P_2^X) = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

gdzie:

- a- liczba par elementów X, które należą do tego samego klastra dla obu podziałów
- b- liczba par elementów X, które należą do różnych klastrów w podziale P_1^X oraz P_2^X
- c- liczba par elementów X, które należą do tego samego klastra w P_1^X, ale do różnych w P_2^X
- d- liczba par elementów X, które należą do różnych klastrów w P_1^X, ale do tego samego klastra w P_2^X
- n- liczba elementów X

klastrujemy klastrowania

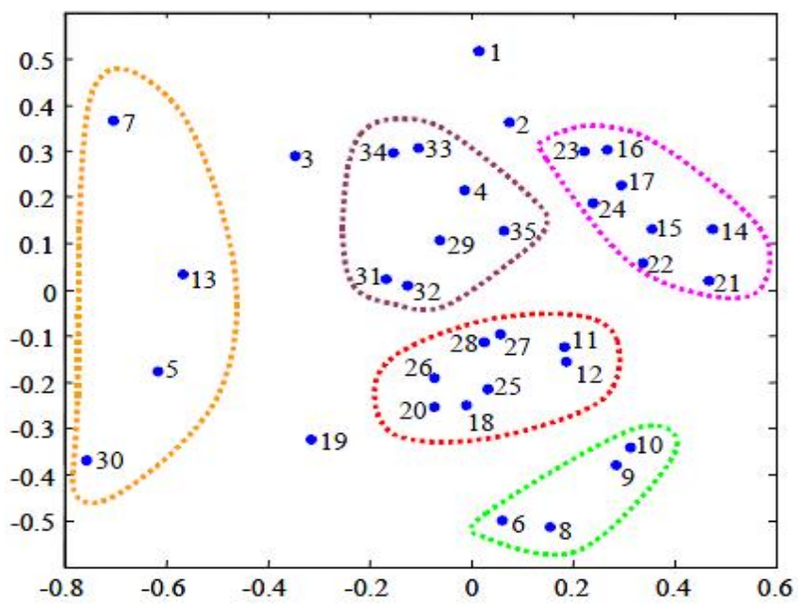
Zbadano 35 algorytmów klastrowania, numerowanych liczbami od 1..35. Niektóre z nich:

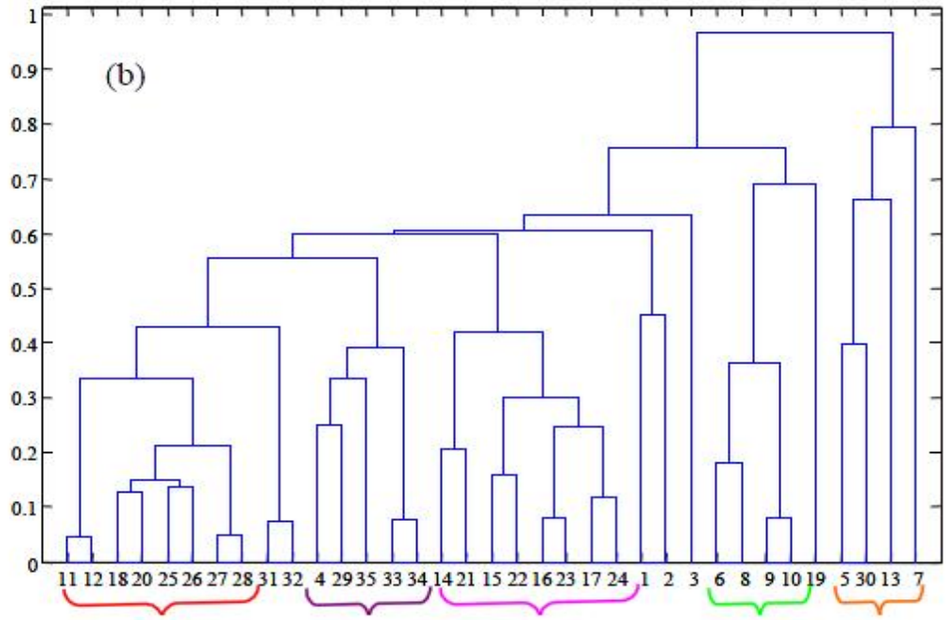
- Algorytm k-średnich (29)
- Algorytmy klastrowania hierarchicznego z użytymi metodami: SL(30), AL(5), CL(13) oraz Ward(35).
- Dwie wersje klastrowania spektralnego z dwoma różnymi parametrami odpowiedzialnymi za współczynniki skalowania

klastrujemy klastrowania

1. Wyznaczenie macierzy odległości 35x35 pomiędzy algorytmami uśrednionej z 12 macierzy dla różnych danych
2. Skalowanie Sammona (stress value=0.0587)
3. Dendogram algorytmów metodą complete-link

klastrujemy klastrowania





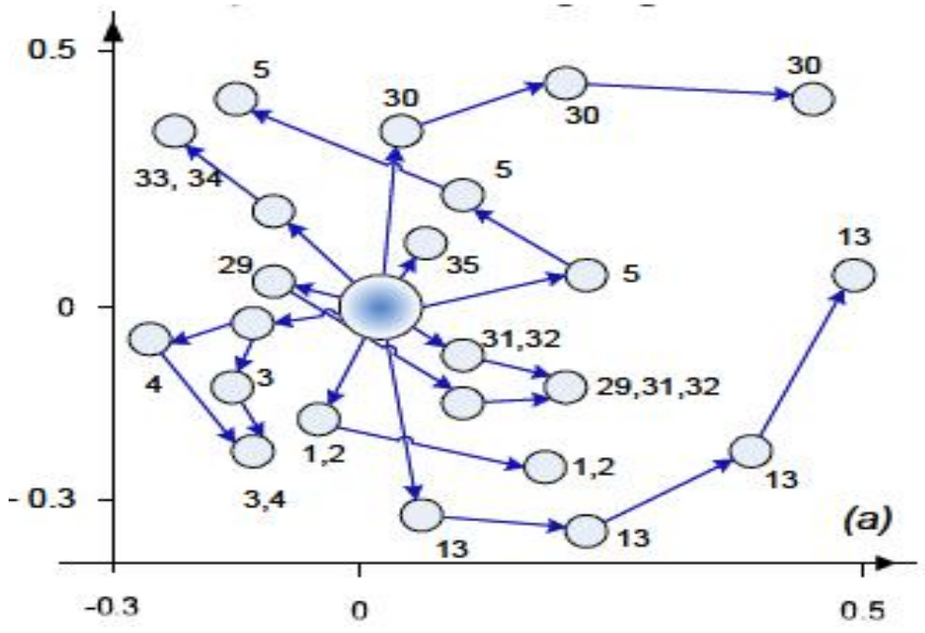
Obserwacje:

- algorytm K-średnich(29) w centrum przestrzeni
- algorytmy typu kameleon blisko siebie (6-12)
- algorytmy spektralne blisko siebie (31-34)

klastrujemy klastrowania

1. Przygotowano 12 sztucznych zestawów danych zawierających po 3 klastry wygenerowane z 2-wymiarowych rozkładów normalnych
2. Kolejne zbiory danych różnią się poziomem separowalności klastrów
3. W wyniku zmiejszania się separowalności klastrów obserwowano przemieszczanie się algorytmów w przestrzeni

klastrujemy klastrowania



Rysunek: Przestrzeń algorytmów klastrowania, ścieżki odpowiadają zmianom położenia algorytmów w wyniku zmiejszania odległości pomiedzy trzema klastrami

podziękowania za współpracę/obrazki:

- Bogusław Kluge (MIM UW)
- Piotr Skibiński (IChF PAN)
- Michał Woźniak (MIM UW)
- Marta Łuksza (MPI Berlin)