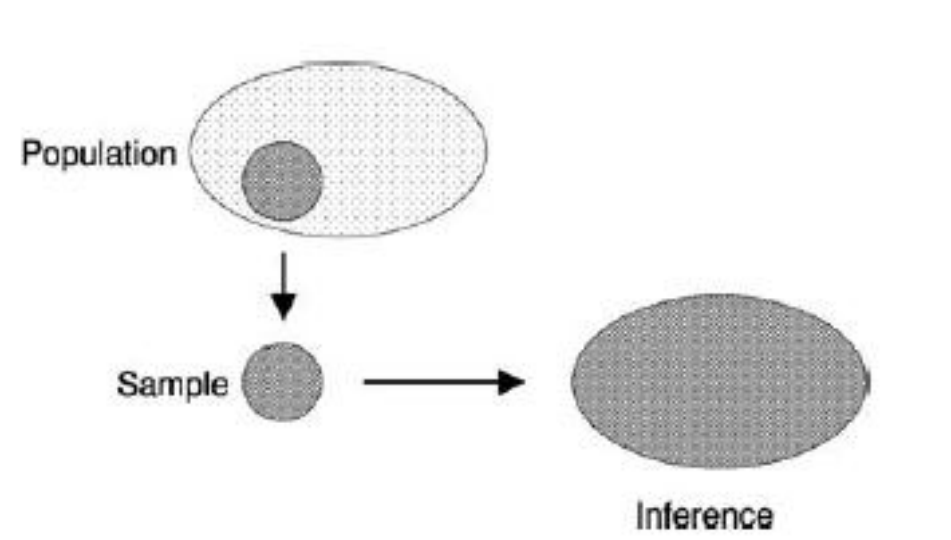
ANALIZA GRUPY GENÓW

ANALIZA SKUPIEŃ

Na poprzednim wykładzie … skrót

Które geny znajdujące się na mikromacierzy uległy *zróżnicowanej ekspresji*?

Cała populacja vs wybrane osobniki



- Nie możemy zbadać całej populacji chorych na raka
- Wybieramy reprezentację kilku (dziesięciu) osobników
- Uogólniamy wyniki na całą populację pacjentów chorych na raka

Hipoteza zerowa

H_0 zakłada że geny nie uległy zróżnicowanej ekspresji

Zbiór A – gen nie uległ zróżnicowanej ekspresji po chemii

Zbiór B – gen nie uległ zróżnicowanej ekspresji dla wszystkich pacjentów z białaczką typu ALL i AML

Jeśli H_0 jest prawdziwa, to znaczy że nie została stwierdzona istotna statystycznie zmiana w ekspresji.

Testy statystyczne

Każdy test hipotezy zerowej tworzy model, który wyznacza prawdopodobieństwo obserwowanej statystyki, np. średnie zróżnicowanie ekspresji genów, które jest co najmniej tak ekstremalne jak obserwowane statystyki w danych.

To prawdopodobieństwo to ***p*-value** . Im mniejsze tym mniej prawdopodobne, że obserwowane dane pojawiły się przypadkowo i tym bardziej pewne wyniki.

Zakładamy że dla genów z niską wartością p-value jest mało prawdopodobne, aby obserwowana zróżnicowana ekspresja pojawiła się przypadkowo, a zatem jest skutkiem biologicznego efektu, który testujemy.

p-value = 0.01 oznacza że mamy 1% szansy na obserwowanie zróżnicowanej ekspresji przez przypadek

	Odrzucenie hipotezy zerowej	**Nie-odrzucenie hipotezy zerowej**
Hipoteza zerowa (H_o) jest prawdziwa	False positive **Błąd I typu**	True positive **poprawne**
Hipoteza zerowa (H_o) jest fałszywa	True negative **poprawne**	False negative **Błąd II typu**

Swoistość - specificity

Moc testu - sensitivity

Testy parametryczne (dla danych z rozkładem normalnym)

test *t* sparowany

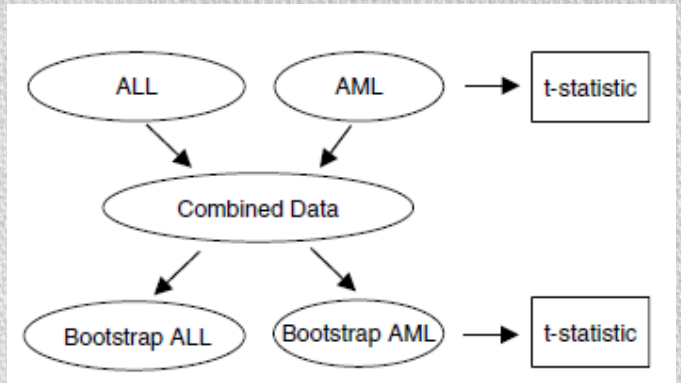
test *t* niesparowany

Testy nieparametryczne

Test Wilcoxona dla par obserwacji (Wilcoxon sign-rank test)

Test Manna-Whitneya = test sumy rang Wilcoxona (Wilcoxon rank-sum test)

Bootstrap – pozwala ominąć założenie o rozkładzie normalnym danych



(mulitifactor) Anova
gdy mamy więcej niż 2 warunki (2 rodzaje próbek) do porównania

eBayes – gdy jest zbyt mało powtórzeń i nie można wyznaczyć wariancji

model liniowy
oraz
złagodzona statystyka t

Wielokrotne testowanie - problem

- Z definicji p-value, każdy gen ma 1% szansy posiadania wartości $p<0.01$, czyli będzie znaczący przy poziomie istotności 1%
- Dla 10000 genów, oczekujemy że
  - 100 genów przejdzie próg $p<0.01$
  - 10 genów przejdzie próg $p<0.001$
  - 1 gen przejdzie próg $p<0.0001$
- Dla zestawu A mamy 9216 genów. Jeśli chemioterapia nie miałaby żadnego wpływu na zmianę ekspresji genów to i tak oczekiwalibyśmy, że dla 92 genów $p<0.01$

- Czy gen naprawdę uległ zróżnicowanej ekspresji, czy jest to wynik błędu I typu (false positive)?

Kontrolowanie false positives

- **Family-wise error rate** (**FWER**)
  - Prawdopodobieństwo co najmniej jednego błędu I typu pomiędzy genami wybranymi jako znaczące

$$FWER = \Pr(FP > 0)$$

Bonferroni

- **False discovery rate** (**FDR**)
  - Oczekiwana proporcja błędów I typu spośród odrzuconych

$$FDR = E(Q), \text{gdzie } Q = \begin{cases} \frac{FP}{R}, & \text{jeśli } R > 0 \\ 0, & \text{jeśli } R = 0 \end{cases}$$

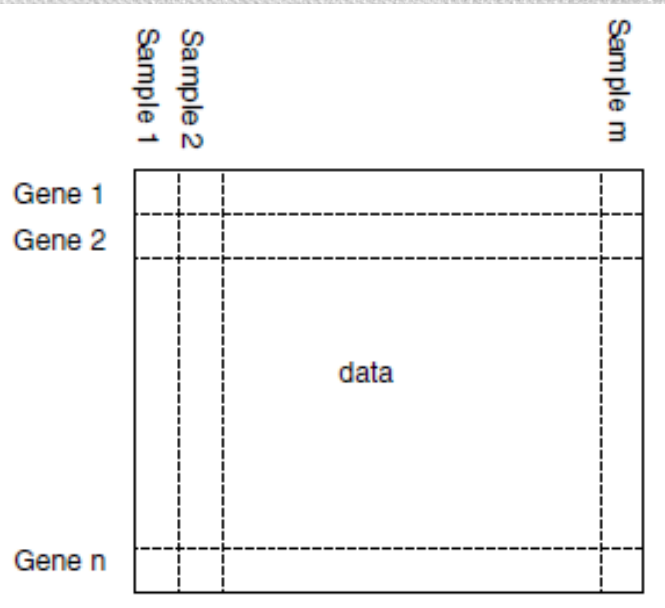
Benjamini Hochberg

R – to suma False Positive i True Negative (czyli wszystkich z odrzuconą hipotezą)

Na dzisiejszym wykładzie …

Będziemy **szukać związku pomiędzy grupami genów**, a nie badać różnicową ekspresję dla pojedynczego genu

Dwa sposoby patrzenia na dane



- Patrzymy na **związek między genami** wykorzystując ekspresję każdej próbki jako pomiar genu
- Patrzymy na **związek między próbkami** używając ekspresji każdego genu jako miarę dla każdej próbki

- z naukowego punktu widzenia jest to odrębna analiza
- z punktu widzenia metod to jest to samo

Klasyfikacja dla mikromacierzy

Klasyfikacja PRÓBEK

- Odróżnia pomiędzy znanymi typami komórek lub warunków np. pomiędzy tkanką rakową a zwykłą
- Identyfikuje różne a poprzednio nieznane typy komórek lub warunków, np. nieznane klasy istniejących chorób rakowych

Klasyfikacja GENÓW

- Przypisanie nieznanej sekwencji DNA do jednej ze znanych klas
- Podział grupy genów na nowe, nieznane funkcjonalne klasy na podstawie ich profili ekspresji dla pewnej liczby próbek

Klasyfikacja czy klastrowanie

Cancer classification	**Class discovery**	**Class prediction**
Machine learning	**Unsupervised learning**	**Supervised learning**
Statistics	**Cluster analysis**	**Discriminant analysis**

Klastrowanie

Klasyfikacja

Klasyfikacja i klastrowanie

Analiza dyskryminacyjna: KLASY ZNANE

Klastrowanie: KLASY NIEZNANE

Metody eksploracji danych

Metody obejmują

- **klastrowanie** (np. hierachiczne, podział, k-średnich)
- **projekcję** (principal component analysis – analiza głównych składowych, skalowanie wielowymiarowe)

Analiza klastrowa/ Analiza skupień

Cel analizy klastrowej

Grupowanie kolekcji obiektów w grupy klastrów, takich że obiekty wewnątrz klastra będą do siebie bardziej podobne niż obiekty powiązane do innych klastrów.

Dwa składniki potrzebne do tworzenia grup obiektów:

Miara odległości

W jaki sposób ocenić czy obiekty są do siebie podobne czy też nie?

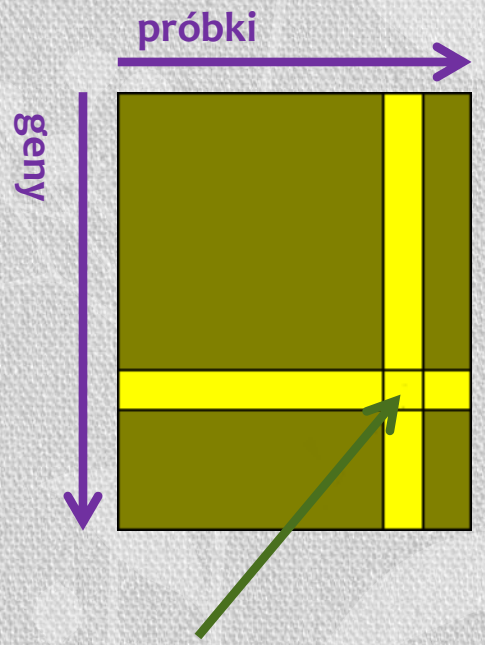
Algorytm do klastrowania

Procedura do minimalizacji odległości pomiędzy obiektami w klastrze/grupie lub/i maksymalizacji odległości pomiędzy grupami

Analiza klastrowa

- Klastrowanie kolumn: grupowanie podobnych próbek
- Klastrowanie wierszy: grupowanie genów z podobną trajektorią

Macierz ekspresji genów

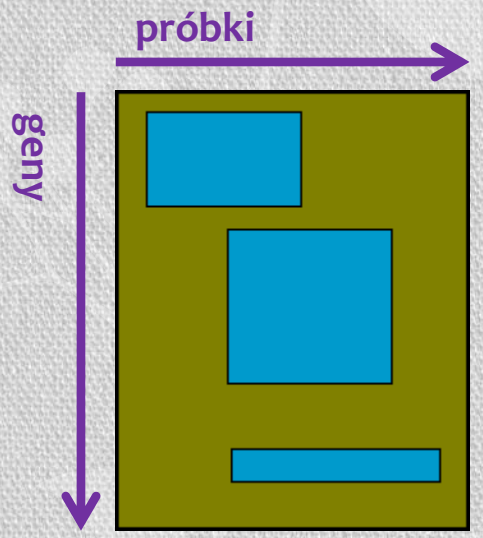


L_{ij} : poziom ekspresji genu *i*-tego w *j*-tej próbce

Analiza klastrowa

- Klastrowanie kolumn: grupowanie podobnych próbek
- Klastrowanie wierszy: grupowanie genów z podobną trajektorią
- Bi-klastrowanie: grupy genów mają podobną trajektorię dla podzbioru próbek

Macierz ekspresji genów

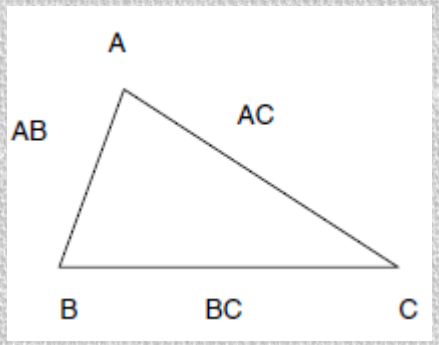


Cechy miary odległości (*distance measure*)

Odległość pomiędzy:

- dwoma profilami powinna być dodatnia
- profilem a nim samym = 0
- dwoma profilami jest 0 ⇔ profile są identyczne
- profilem A i B = odległości między profilem B i A

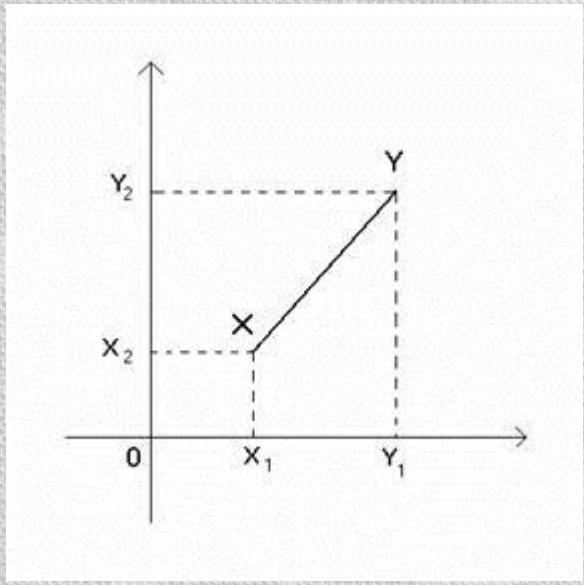
- Nierówność trójkąta



$$AB \leq AC + BC$$
$$AC \leq AB + BC$$
$$BC \leq AB + BC$$

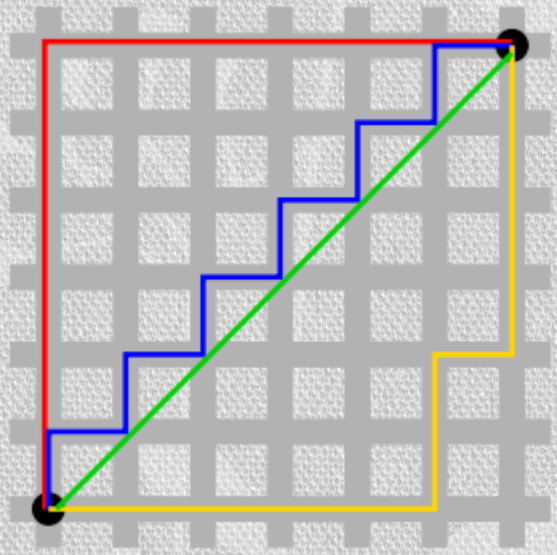
Odległość Euklidesowa

$$d(x,y) = \sqrt{\sum (x_i - y_i)^2}$$



Odległość miejska - Manhattan

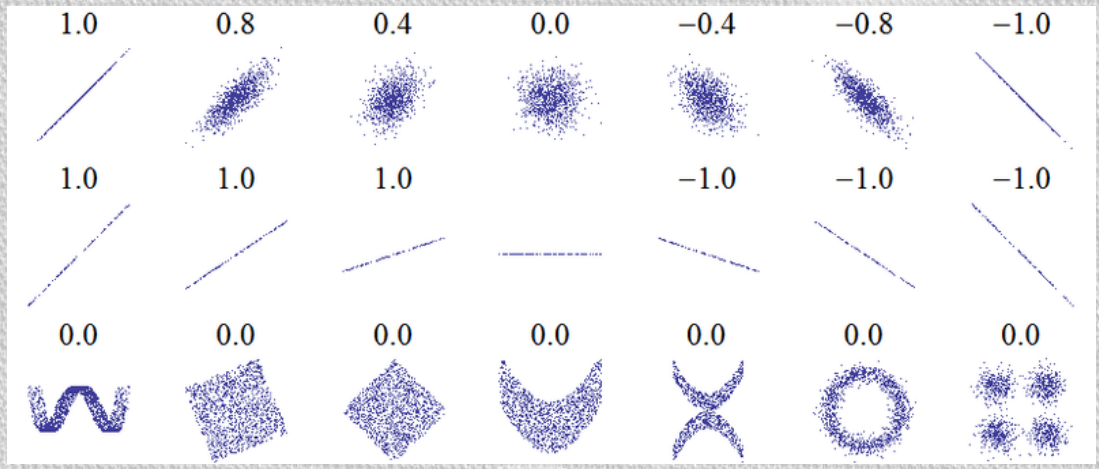
$$d(x, y) = \sum |x_i - y_i|$$



Korelacja

Współczynnik korelacji określa w jakim stopniu zmienne są współzależne

- Współczynniki są znormalizowane:
  - -1 zupełna korelacja ujemna
  - 0 brak korelacji
  - +1 zupełna korelacja dodatnia



Przykłady

- Korelacja Pearsona
- Korelacja Spearmana

- W przypadku, gdy rozkład zmiennej nie jest normalny lub są obserwacje odstające, współczynnik korelacji Pearsona może fałszywie wskazywać na nieistniejącą korelację.
- Wady tej nie mają współczynniki rangowe, które z kolei mają mniejszą efektywność dla rozkładów bliskich normalnemu

Korelacja Pearsona i Spearmana

Korelacja Pearsona:

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

Odchylenie standardowe dla y

Odchylenie standardowe dla x

Korelacja rangowa mierzy monotoniczną zależność między zmiennymi, a nie tylko liniową. Np. jeśli jedna ze zmiennych jest rosnącą funkcją drugiej, to korelacja rangowa przyjmuje wartość maksymalną.

Korelacja Spearmana

Dla każdej porównywanej zmiennej(próbki) dokonuje się rangowania:

1. Zaobserwowane wartości porządkowane są rosnąco
2. Każdej wartości x_i przypisywana jest ranga Rx_i równa pozycji w uporządkowaniu
3. Jeśli pewna wartość występuje kilkukrotnie to wszystkim pozycjom z tą wartością przypisujemy rangę równą średniej arytmetycznej pozycji (*ranga wiązana*)
4. Wracamy do pierwotnego porządku (zamiast wartości mamy teraz rangi)
5. Wyznaczamy korelacje wg wzoru na korelacje Pearsona – **dla rang zamiast wartości!**

Korelacja Pearsona i Spearmana

Korelacje Pearsona i Spearmana można przyrównać do testów parametrycznych (statystyka *t* sparowana, niesparowana) i nieparametrycznych (test Wilcoxona, Mann-Whitneya).

Te pierwsze wymagają rozkładu normalnego, natomiast rangowe są bardziej odporne na wartości odstające, ale za to mają mniejszą moc.

Korelacja, a miara odległości

$$r_{xy} \in [-1,1]$$

Odległość pomiędzy:

- dwoma profilami powinna być dodatnia
- profilem a nim samym = 0
- dwoma profilami jest 0 ⇔ profile są identyczne
- profilem A i B = odległości między profilem B i A
- Nierówność trójkąta

Konwersja korelacji na miarę odległości:

$$d(x,y) = 1 - r_{xy}$$

Przykład - porównanie miar

Biologia

Pomiar ekspresji genu badany dla 4 kolejnych dni.

Statystyka

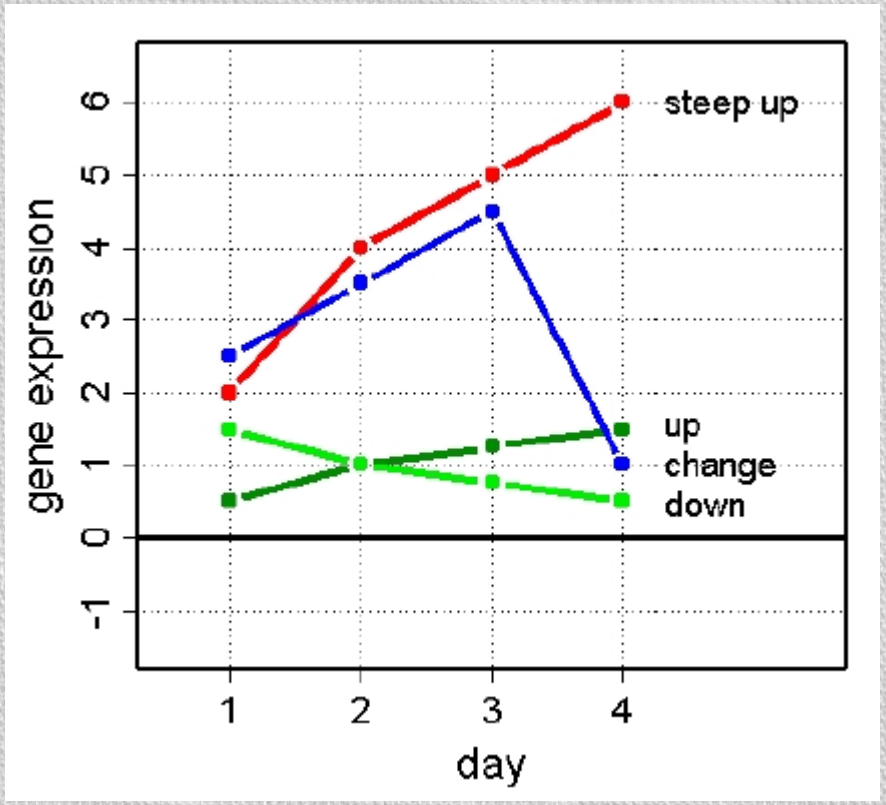
Każdy gen zakodowany jest przez wektor o długości 4.

steep up $x_1 = (2,4,5,6)$

up $x_2 = (2/4,4/4,5/4,6/4)$

down $x_3 = (6/4,4/4,3/4,2/4)$

change $x_4 = (2.5,3.5,4.5,1)$



Przykład - porównanie miar

Odległość Euklidesowa

Odległość pomiędzy dwoma wektorami to pierwiastek kwadratowy z sumy kwadratów różnic poszczególnych koordynat

$$d(x_1, x_2) = \sqrt{(2-2/4)^2+(4-4/4)^2+(5-5/4)^2+(6-6/4)^2} = \sqrt{729/16} = 6.75$$

steep up $x_1 = (2,4,5,6)$
up $x_2 = (2/4,4/4,5/4,6/4)$
down $x_3 = (6/4,4/4,3/4,2/4)$
change $x_4 = (2.5,3.5,4.5,1)$

0	6.75	7.59	5.07
6.75	0	1.5	4.59
7.59	1.5	0	4.64
5.07	4.59	4.64	0

Macierz odległości

Przykład - porównanie miar

Odległość Manhattan

Odległość pomiędzy dwoma wektorami to suma z wartości bezwzględnej z różnic poszczególnych koordynat

$$d(x_1, x_2) = |2 - 2/4| + |4 - 4/4| + |5 - 5/4| + |6 - 6/4| = 51/4 = 12.75$$

steep up $x_1 = (2,4,5,6)$
up $x_2 = (2/4,4/4,5/4,6/4)$
down $x_3 = (6/4,4/4,3/4,2/4)$
change $x_4 = (2.5,3.5,4.5,1)$

0	12.75	13.25	6.50
12.75	0	2.50	8.25
13.25	2.50	0	7.75
6.50	8.25	7.75	0

Macierz odległości

Przykład - porównanie miar

Odległość korelacji

Odległość pomiędzy dwoma wektorami jest równa $1-r_{x1x2}$, gdzie r jest współczynnikiem korelacji Pearsona

$$d(x_1, x_2) = 1 - r_{x_1 x_2} = 1 - \frac{\left(2-\frac{17}{4}\right)\left(\frac{2}{4}-\frac{17}{16}\right)+\left(4-\frac{17}{4}\right)\left(\frac{4}{4}-\frac{17}{16}\right)+\left(5-\frac{17}{4}\right)\left(\frac{5}{4}-\frac{17}{16}\right)+\left(6-\frac{17}{4}\right)\left(\frac{6}{4}-\frac{17}{16}\right)}{\sqrt{\left(2-\frac{17}{4}\right)^2+\left(4-\frac{17}{4}\right)^2+\left(5-\frac{17}{4}\right)^2+\left(6-\frac{17}{4}\right)^2}\sqrt{\left(\frac{2}{4}-\frac{17}{16}\right)^2+\left(\frac{4}{4}-\frac{17}{16}\right)^2+\left(\frac{5}{4}-\frac{17}{16}\right)^2+\left(\frac{6}{4}-\frac{17}{16}\right)^2}}$$

steep up $x_1 = (2,4,5,6)$
up $x_2 = (2/4,4/4,5/4,6/4)$
down $x_3 = (6/4,4/4,3/4,2/4)$
change $x_4 = (2.5,3.5,4.5,1)$

0	0	2	1.18
0	0	2	1.18
2	2	0	0.82
1.18	1.18	0.82	0

Macierz odległości

Przykład - porównanie miar

Euklidesowa

0	6.75	7.59	5.07
6.75	0	1.5	4.59
7.59	1.5	0	4.64
5.07	4.59	4.64	0

Manhattan

0	12.75	13.25	6.50
12.75	0	2.50	8.25
13.25	2.50	0	7.75
6.50	8.25	7.75	0

Korelacja

0	0	2	1.18
0	0	2	1.18
2	2	0	0.82
1.18	1.18	0.82	0

Porównanie:
Wszystkie dystanse zostały znormalizowane do przedziału [0,10] i zaokrąglone

	steep up	up	down	change
steep up	0 0 0	9 9 0	10 10 10	8 4 5
up	9 9 0	0 0 0	4 1 10	7 6 5
down	10 10 10	4 1 10	0 0 0	7 5 4
change	8 4 5	7 6 5	7 5 4	0 0 0

Porównanie miar

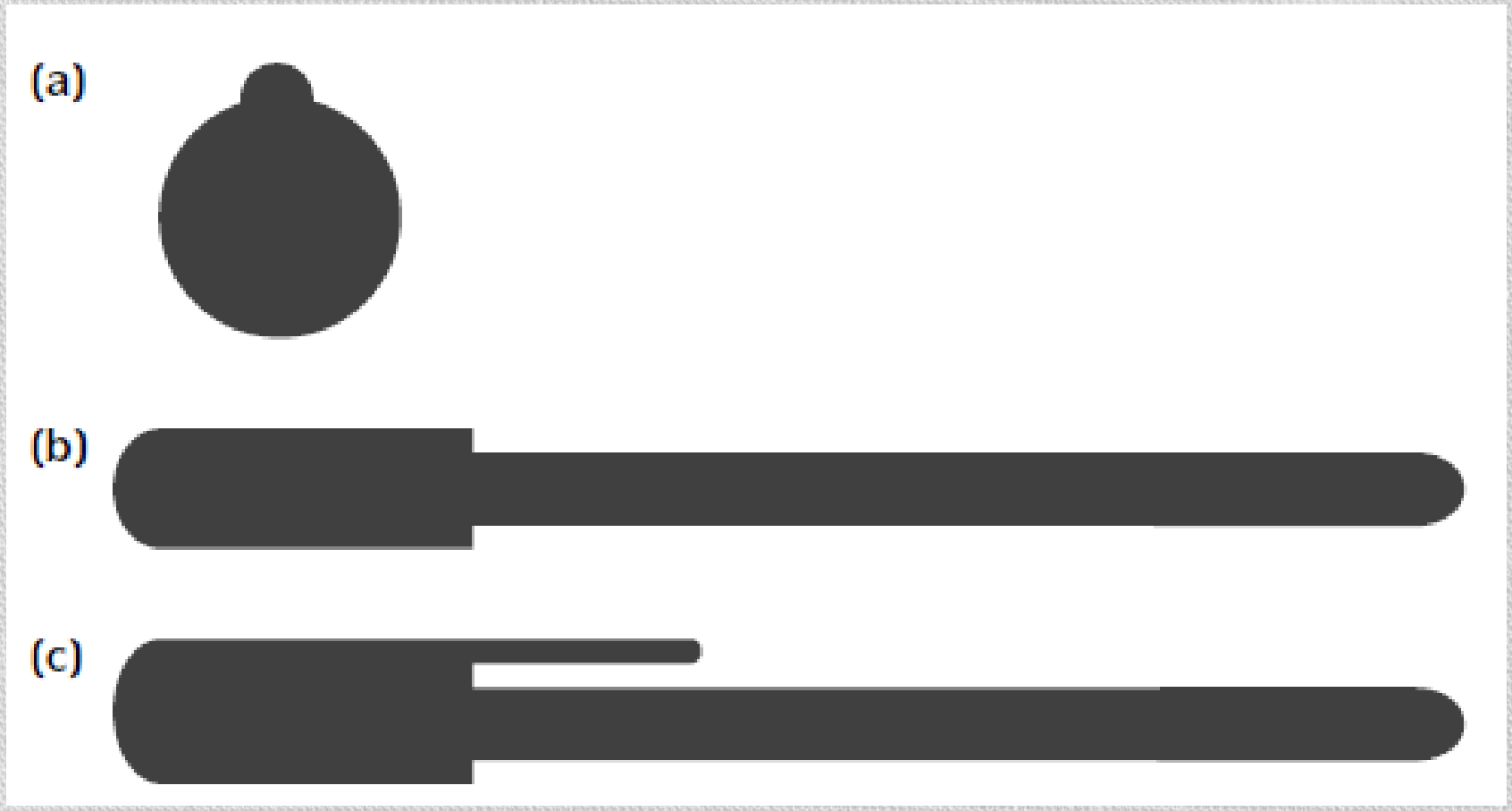
TABLE 8.2: Strengths and Weaknesses of Different Distance Measures

Pearson Correlation	Spearman Correlation	Euclidean Distance
✓ Powerful	✓ Robust to outliers	✓ Geometric interpretation
✓ Spots positive and negative correlations	✓ Spots positive and negative correlations	✓ Can retain up- or down-regulation information with appropriate scaling
✓ Scale invariant on centred data	✓ Completely scale invariant: no scaling or centering required	✓ Can detect magnitude of changes if used without scaling
× Assumes linearity	× Less powerful	× Not scale invariant: results depend on scaling used
× Susceptible to outliers	× Ignores pattern of up- or down-regulation in time series	× Cannot detect negative correlations

Analiza głównych składowych (*principal component analysis*)

- Przed użyciem skomplikowanych algorytmów chcemy obejrzeć dane – **ale jak obejrzeć dane wielowymiarowe na 2D**?
- PCA wizualizuje dane wielowymiarowe (liczba genów * liczba próbek) patrząc „***pod pewnym kątem***” na te dane
- **„*pod pewnym kątem*” oznacza obrócenie osi układu współrzędnych tak, aby uchwycić jak największą różnorodność w danych** – przedstawiamy na 2D (kartce/ekranie), a następnie ignorujemy pozostałe wymiary

PCA - przykład



PCA dla mikromacierzy

- Dla eksperymentu, gdzie mamy 10 000 genów tworzymy **macierz kowariancji** (10 000 x 10 000)

Macierz kowariancji

Macierz pokazuje różnorodność każdej zmiennej, oraz jej korelacji ze wszystkimi innymi zmiennymi

Wariancja dla przypadku wielowymiarowego

$$\begin{bmatrix} \sigma_1^2 & cov_{12} & \cdots & cov_{1n} \\ cov_{21} & \sigma_2^2 & \cdots & cov_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ cov_{n1} & cov_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

- cov_{ij} to kowariancja między x_i a x_j

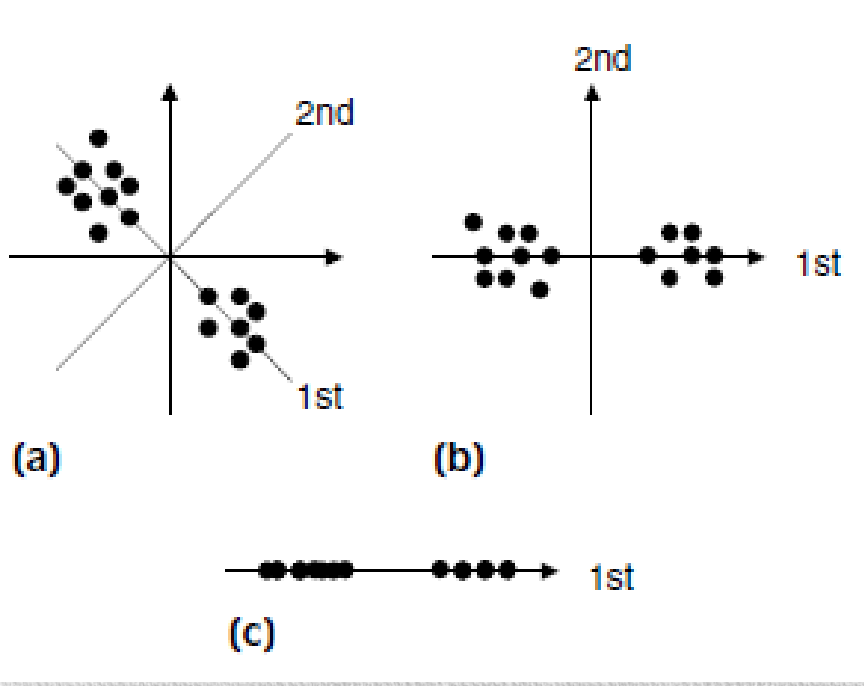
$$cov_{ij} = r_{ij} * \sigma_i * \sigma_j$$

- $cov_{ii} = \sigma_i^2$ to wariancja zmiennej x_i

- Kowariancja jest = 0 jeśli między zmiennymi nie istnieje żadna zauważalna **korelacja liniowa**
- *Może natomiast istnieć korelacja nieliniowa*

PCA dla mikromacierzy

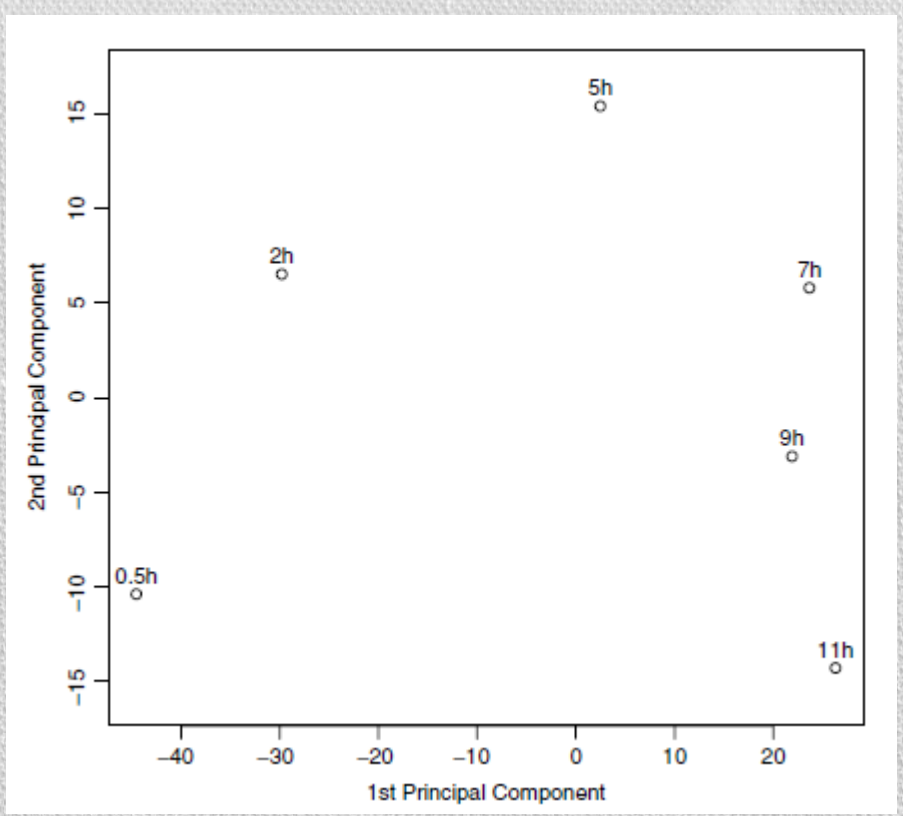
- Dla eksperymentu, gdzie mamy 10 000 genów tworzymy **macierz kowariancji** (10 000 x 10 000)
- Szukamy takiej *liniowej kombinacji genów*, która będzie miała największą wariancję – **pierwsza główna składowa**
- Szukamy kolejnej liniowej kombinacji o największej wariancji, która będzie prostopadła do pierwszej – **druga główna składowa**
- Krok powtarzamy dopóki nie uzyskamy wymaganej liczby głównych składowych. Każda składowa jest liniową kombinacją wszystkich oryginalnych genów



PCA – przykład dla mikromacierzy

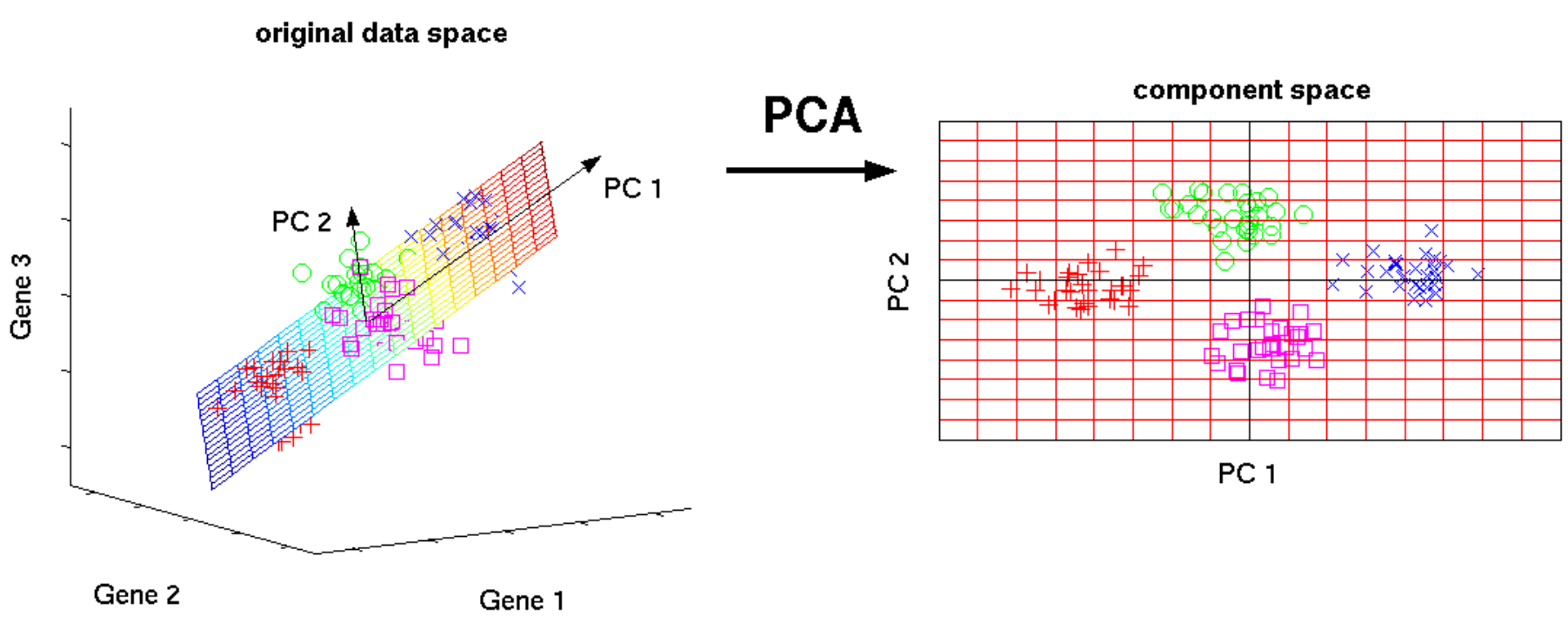
- Badanie drożdży, którym zapewniono warunki do rozmnażania bezpłciowego
- Próbki drożdży zostały pobrane w 6 różnych punktach czasowych.
- Chcemy zidentyfikować grupy genów, które zachowywały się w podobny sposób w tym samym punkcie czasowym
- Z 6000 genów zostało wybranych 1000, o największej wariancji

- Próbki tworzą jasny wzorzec na wykresie PCA
- Jeśli byśmy patrzyli tylko na pierwszą składową (oś X) to punkty czasowe 7, 9 i 11 byłyby nierozróżnialne
- Trzy ostatnie punkty czasowe wskazują że jest pewien proces, który zostaje zatrzymany po 5h i następuje powrót do stanu początkowego



Principal Component	1st	2nd	3rd	4th	5th
Standard deviation	30.3	11.3	9.1	5.9	5.3
Proportion of variance	77%	11%	7%	3%	2%
Cumulative proportion	77%	88%	95%	98%	100%

Pierwsze 5 głównych składowych wyjaśnia całkowitą różnorodność pośród 1000 genów, a 88% różnorodności jest wyjaśnione przez 2 główne składowe na rysunku powyżej


original data space
PC 2
PC 1
Gene 3
Gene 2
Gene 1
PCA
component space
PC 2
PC 1

Multidimensional Scaling plot (MDS)

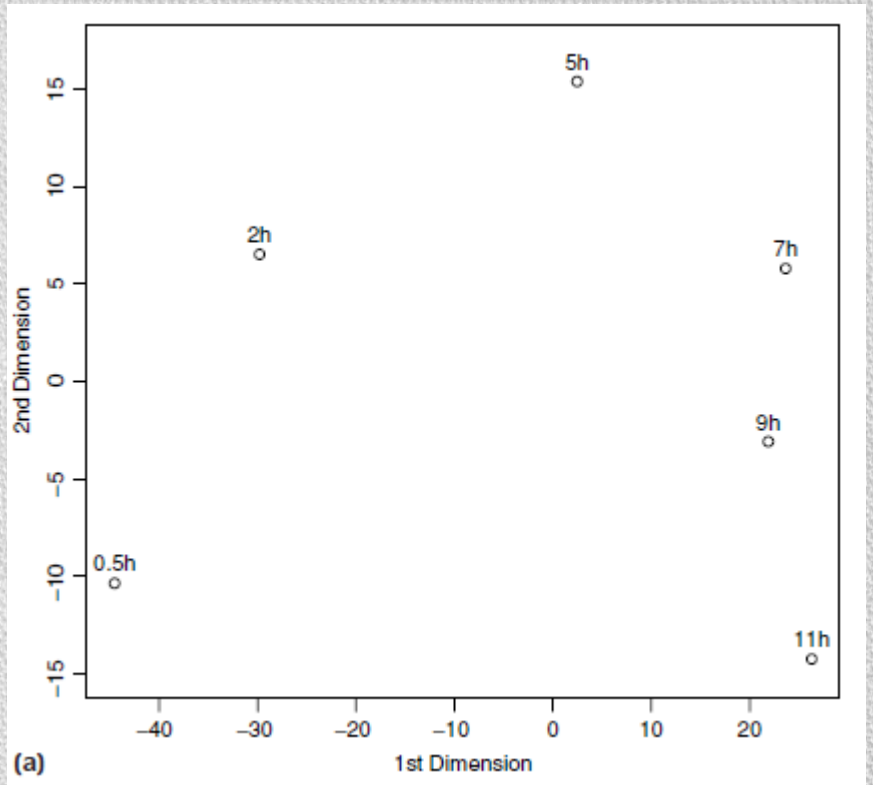
Redukcja wymiarów zaczyna się od wyznaczenia odległości pomiędzy próbkami lub profilami czasowymi (w *time-series analysis*)

Następnie próbki/profile są umiejscawiane w przestrzeni 2D lub 3D tak, aby bliskie sobie próbki znalazły się blisko siebie, a niepodobne z dala od siebie.

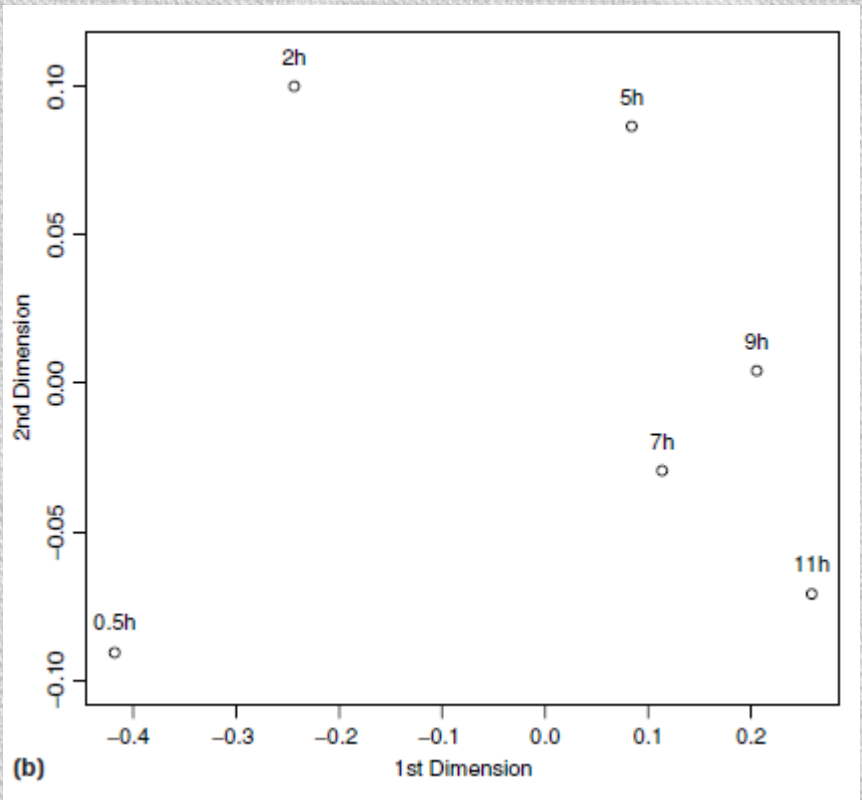
Różnica w porównaniu do PCA polega na możliwości zastosowanie różnych miar podobieństwa/odległości

Multidimensional Scaling plot (MDS)

Odległość Euklidesowa



Odległość – korelacja Pearsona



Algorytmy klastrowania

- **Popularne algorytmy** do klastrowania danych mikromacierzowych:
  - Hierarchiczne klastrowanie
  - K-means
  - Paritioning Around Medoids (PAM)
  - Self-Organizing Maps (SOM)
- K-means i SOM biorą **oryginalne dane jako wejście**: zakłada się że atrybuty leża w przestrzeni Euklidesowej
- Hierarchiczne klastrowanie i PAM pozwala na **wybór macierzy podobieństwa** d, która przypisuje dla każdej pary obiektów x_i oraz x_j wartości $d(x_i, x_j)$ jako odległości

Hierarchiczne klastrowanie

Hierarchiczne klastrowanie było pierwszym algorytmem użytym w eksperymencie mikromacierzowym do klastrowania genów (Eisen et al., 1998)

Algorytm

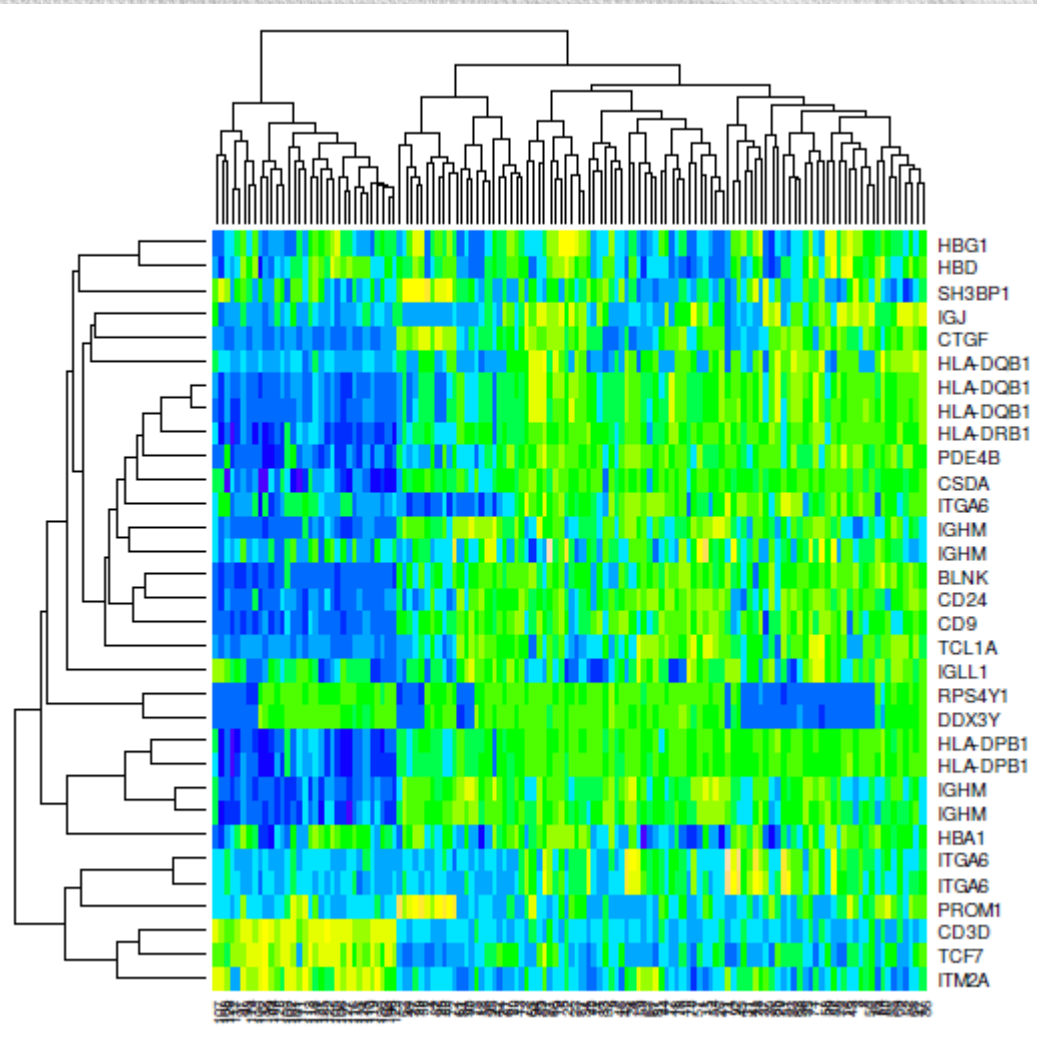
1. Każdy obiekt tworzy swój własny klaster
2. Iteracyjnie:
   - dwa najbardziej podobne klastry są łączone i zastępowane nowym wierzchołkiem w drzewie klastrów. Nowy wierzchołek jest wyznaczony jako średnia wszystkich obiektów w połączonym klastrze
   - macierz podobieństwa jest uaktualniana nowym wierzchołkiem, który zastąpił dwa połączone klastry
3. Powtarzaj krok nr 2 dopóki nie zostanie jeden klaster

Hierarchiczne klastrowanie

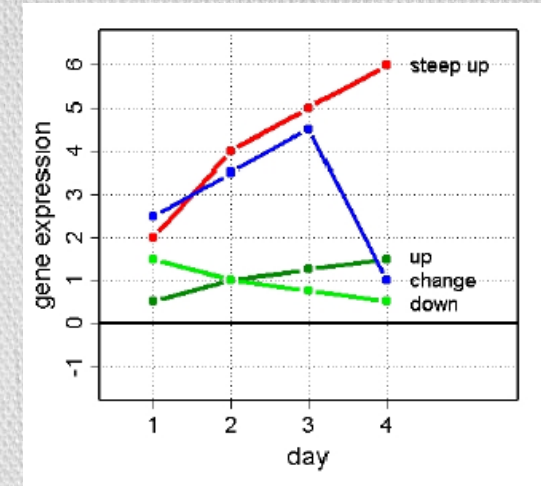
- Wyznaczenie odległości $d(\boldsymbol{G},\boldsymbol{H})$ pomiędzy klastrami $\boldsymbol{G}$ i $\boldsymbol{H}$ opiera się na podobieństwie obiektów pomiędzy obiektami z dwóch klastrów:
  - Single linkage używa najmniejszej odległości $d_d(G,H) = \min_{i\in G, j\in H} d_{ij}$
  - Complete linkage używa największej odległości $d_c(G,H) = \max_{i\in G, j\in H} d_{ij}$
  - Average linkage używa średniej odległości $d_a(G,H) = \frac{1}{N_G N_H} \sum_{i\in G} \sum_{j\in H} d_{ij}$
- Dwa sposoby budowania drzewa: bottom-up, top-down

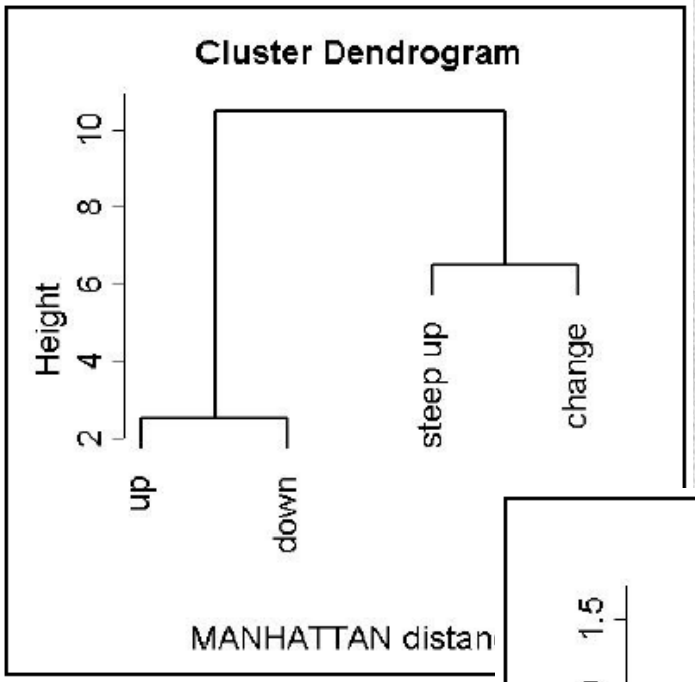
Hierarchiczne klastrowanie

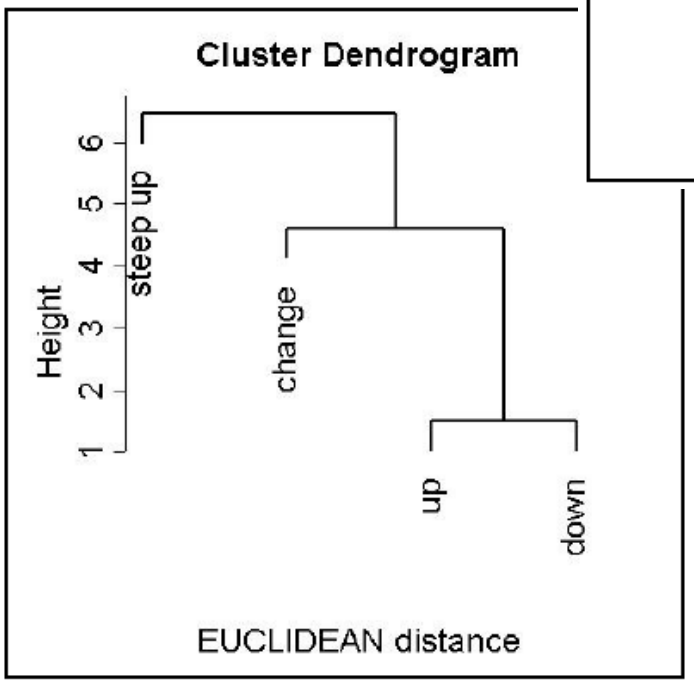
Heatmap i dendrogram

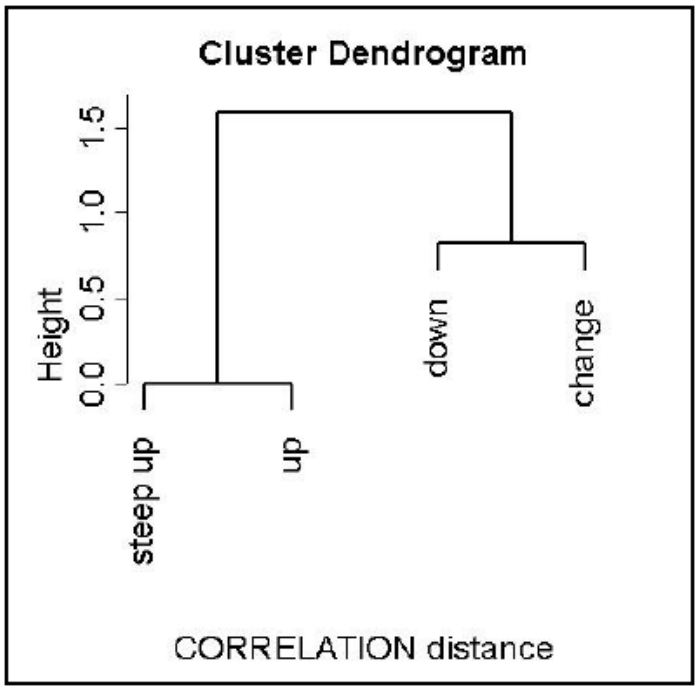


Hierarchiczne klastrowanie









K-means

Cechy charakterystyczne dla k-means w odróżnieniu od klastrowania hierarchicznego:

- Liczba klastrów musi być zdefiniowana na początku
- Nie ma hierarchii ani powiązań między klastrami
- Algorytm *k*-means startuje od losowo utworzonych klastrów, więc każde uruchomienie algorytmu może utworzyć inne klastry

K-means

Algorytm:

1. Wybierz liczbę klastrów *k*
2. Losowo przypisz każdy profil ekspresji genu do jednego z klastrów
3. Wyznacz centroid dla każdego klastra
4. Dla każdego profilu po kolei wyznacz jego odległość od centroidów każdego z klastrów
5. Przypisz profil do klastra, którego centroid jest najbliższy. Jeśli profil zmienił przypisanie do klastra, wyznacz centroidy dla obu klastrów (poprzedniego i obecnego)
6. Wróć do kroku 4 dopóki żaden z profili nie zmieni klastra

Partitioning around medoids (PAM)

Algorytm:

1. Wybierz losowo *k* punktów jako **punkty centralne klastrów** (medoidy)
2. Połącz każdy punkt z najbliższym punktem centralnym używając miary podobieństwa
3. Dla każdego punktu centralnego:
   - zamień go z każdym innym punktem nie będącym medoidem i wyznacz koszt takiej konfiguracji
4. Wybierz konfigurację z najniższym kosztem
5. Powtarzaj kroki 2-4 dopóki nie ma żadnych zmian w medoidach

Różnica między medoidem a centroidem jest taka, że medoid jest jednym z punktów, a centroid niekoniecznie

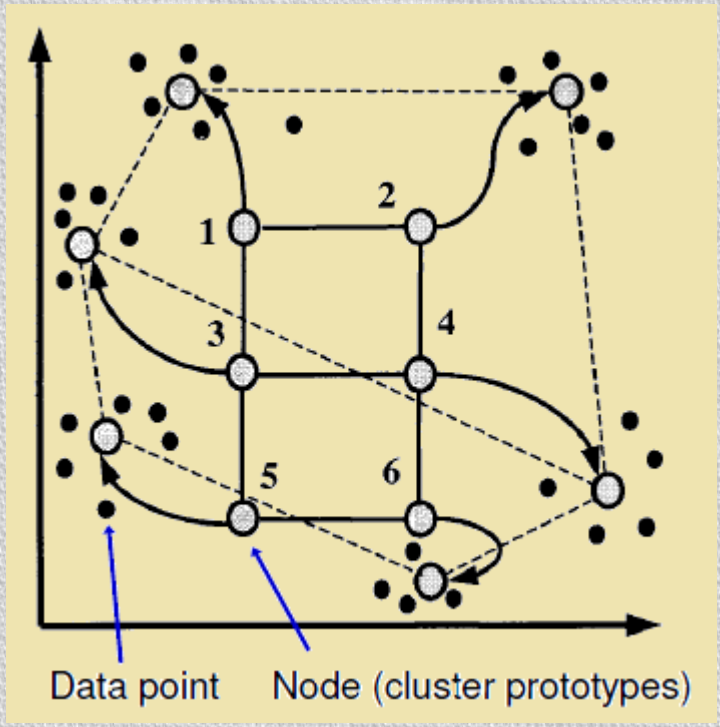
Self-Organizing Maps (SOM)

SOM służy do wizualizacji danych wielowymiarowych na przestrzeni o mniejszej liczbie wymiarów, np.2D. SOM zwane są także **sieciami Kohonena**.

Tworzona jest sieć wierzchołków zwanych także neuronami lub prototypami. Sieć neuronów dąży do wzorca zgodnego ze strukturą analizowanych danych

Algorytm:

1. Wybierz losowo punkt *P*
2. Przesuń wszystkie prototypy w kierunku punktu; leżące najbliżej przesuń najwięcej, a odległe – najmniej
3. Powtarzaj punkty 1-2 dopóki nie wykona się zadana liczba iteracji



Jak wybrać liczbę klastrów *k*?

- Wskaźniki wewnętrzne
  - Statystyki wyznaczane są na podstawie macierzy odległości wewnątrz i pomiędzy – klastrowych.
  - Estymowana jest liczba klastrów, która minimalizuje/maksymalizuje wskaźniki
- Gap statistics
- Average silhouette width

Jak wybrać liczbę klastrów *k*?

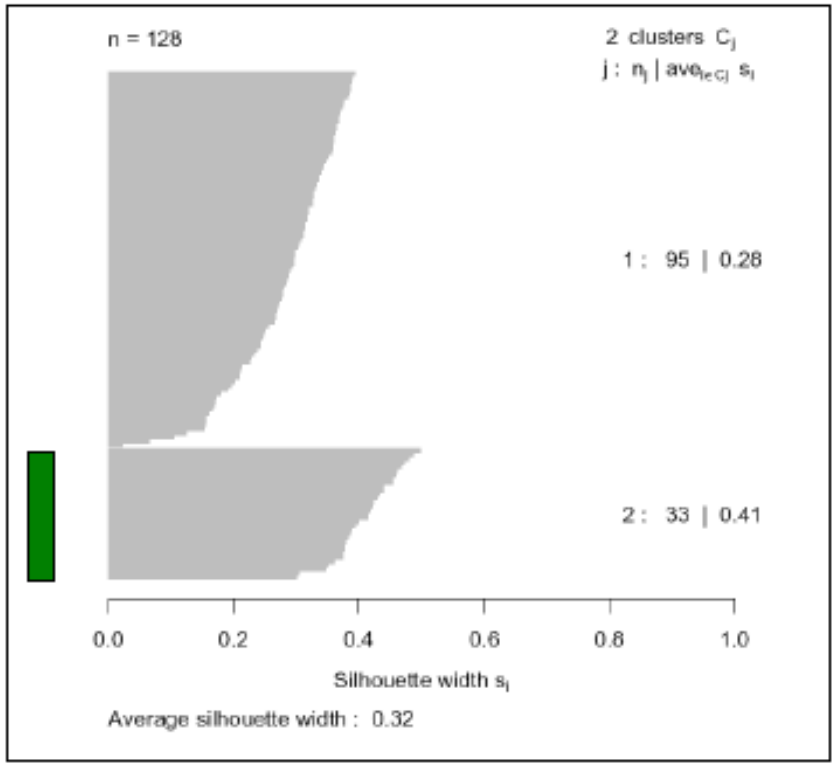
Average silhouette width – podejście heurystyczne

- Dla każdej obserwacji *i*, definiuje się szerokość silhouette ***s(i)***
  - ***a(i)*** = średnia odległość pomiędzy *i* oraz wszystkimi punktami z jego klastra
  - dla *wszystkich innych* klastrów C, niech ***d(i,C)*** = średnia odległość od *i* do wszystkich obserwacji w C. Niech $b(i) = \min_C d(i,C)$
  - Szerokość silhouette $s(i) = (b(i)-a(i)) / \max(a(i),b(i))$
- Maksymalna **średnia szerokość silhouette** dla wszystkich obserwacji może być użyta do wyboru liczby klastrów
- Obserwacje z *s(i)* bliskim 1 mogą być rozważane jako dobrze klastrowane, a $s(i)<0$ jako złe.
- Optymalna liczba klastrów nie może być zdeterminowana w ogólności, ponieważ jakość wyników klastrowania zależy od koncepcji klastrowania (wybranej metody, miary odległości,…)

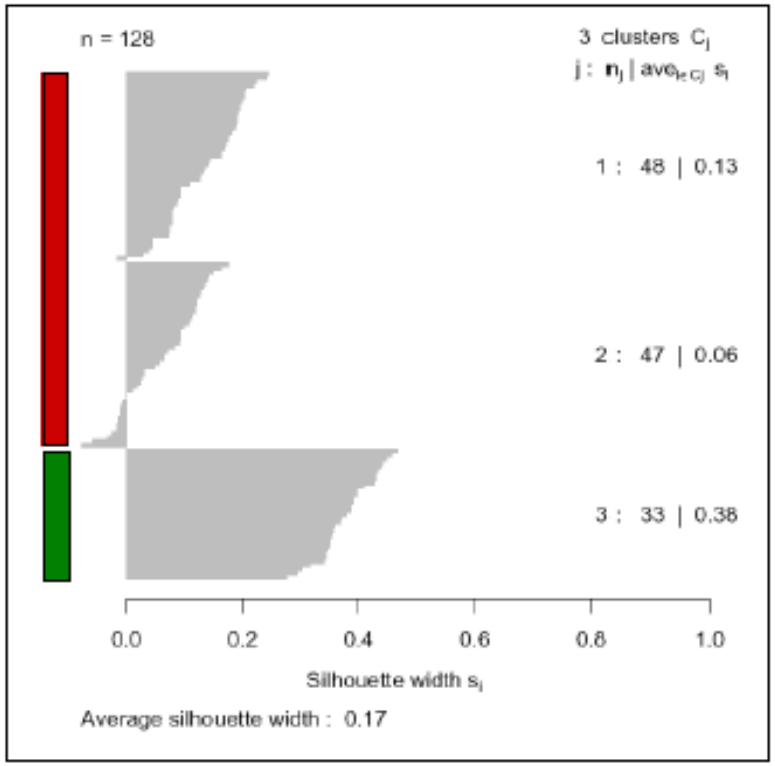
Jak wybrać liczbę klastrów *k*?

- Silhouette plots for clustering Leukemia patients (Chiaretti et al., 2004)

K=2 clusters



K=3 clusters



Green: Well separated cluster

Red: No clear cluster structure

W jaki sposób ocenić jakość klastrowania?

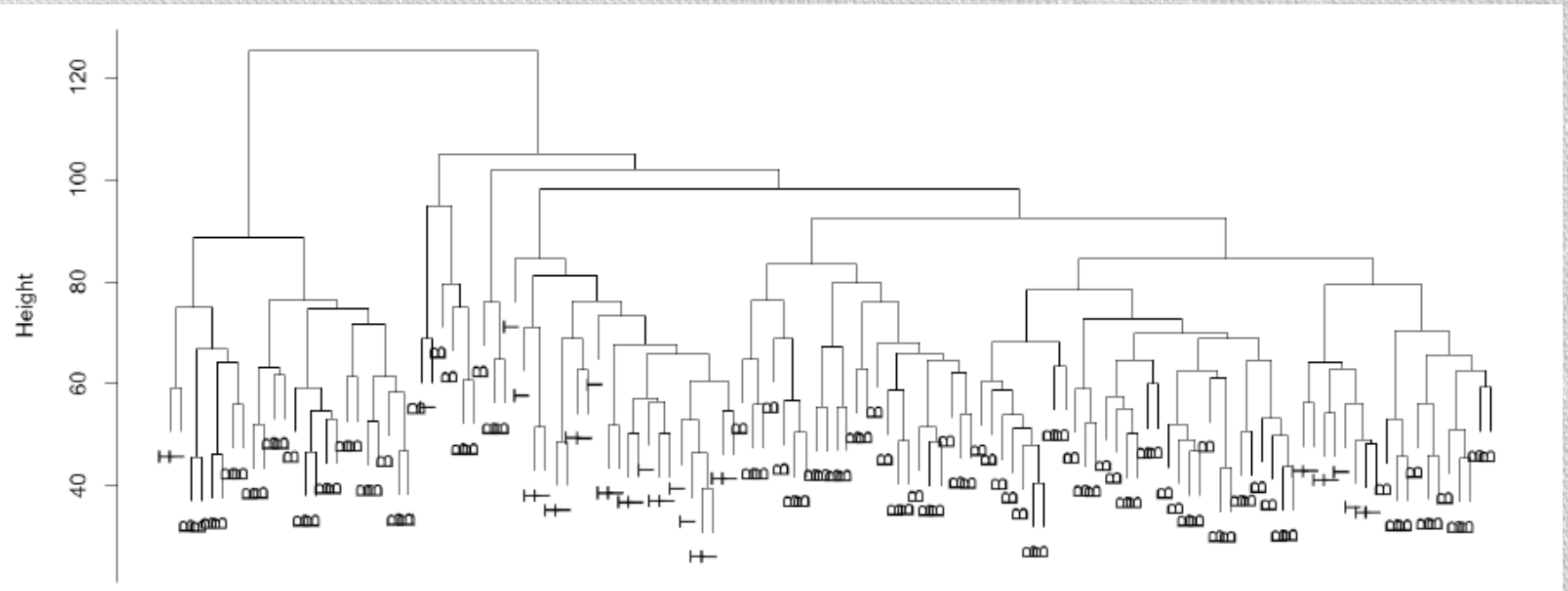
- Wizualnie – obejrzeć czy profile ekspresji genów w klastrach są podobne. Trudne do weryfikacji, gdy jest wiele genów i próbek
- Powiązanie biologiczne – sprawdzenie czy geny w jednym klastrze mają podobną funkcjonalność
- K-means – *sprawdzenie czy dla kilkukrotnego uruchomienia z tą samą liczbą klastrów otrzymujemy takie same klastry* (jeśli nie, być może liczba klastrów jest niepoprawnie dobrana)
- Metody bootstrapowe – tworzonych jest wiele zestawów wejściowych i każdy z nich jest klastrowany, następnie klastry są porównywane;
  - najpewniejsze są klastry, które powtarzają się we wszystkich badanych zbiorach;
  - klastry, które pojawiają się rzadko nie są odporne na błędy eksperymentalne i nie można z nich wyciągać pewnych wniosków

Nawet proste algorytmy klastrowania działają lepiej w przypadku gdy dołożymy chociaż trochę więcej wiedzy

Wybór genów dla klastrowania

Jest wiele metod wyboru grupy genów. Do klastrowania wybiera się zazwyczaj **grupę genów z największą wariancją - top100** .

Skraca to czas obliczeń i redukuje szum w danych.

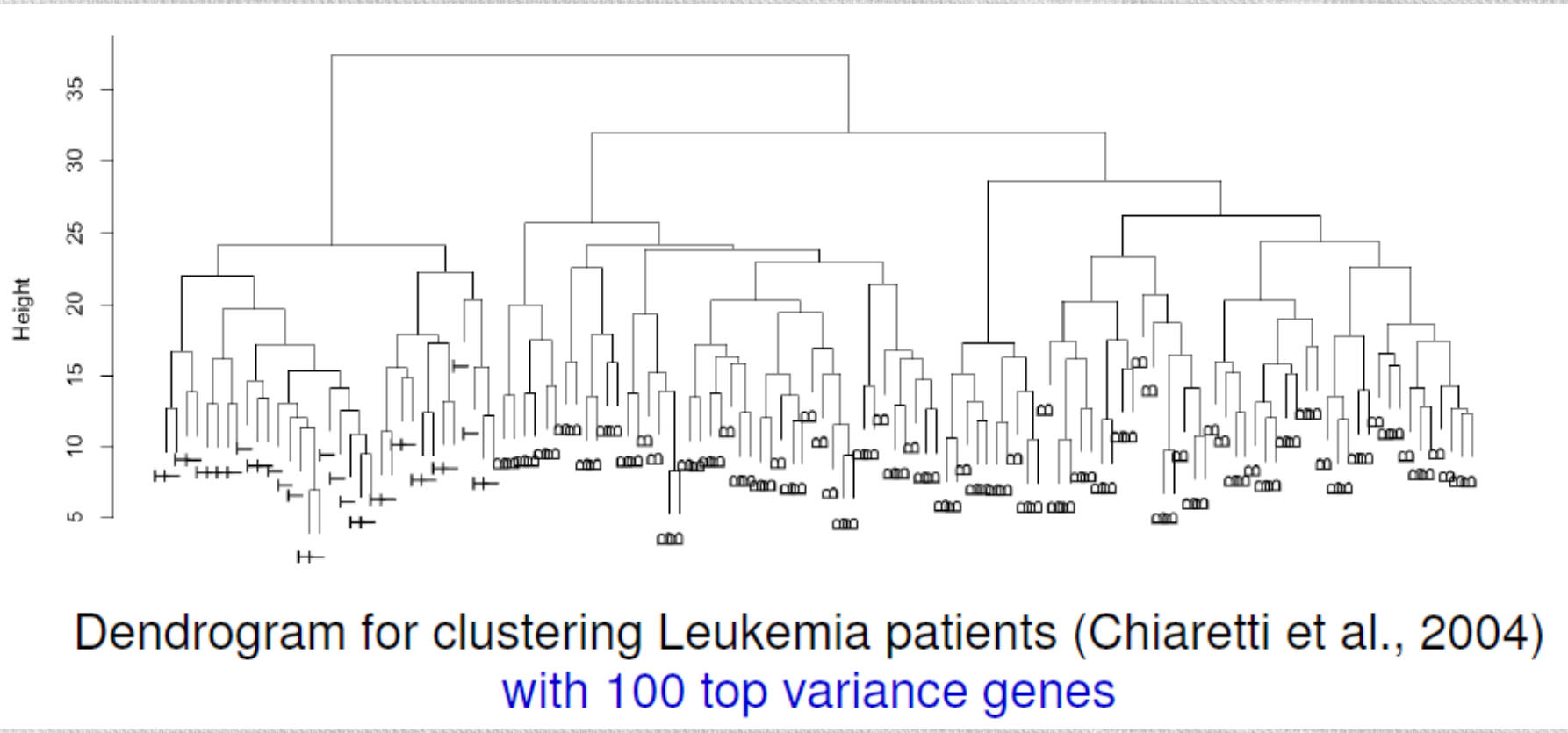


Dendrogram for clustering Leukemia patients (Chiaretti et al., 2004)
without gene selection

Wybór genów dla klastrowania

Jest wiele metod wyboru grupy genów. Do klastrowania wybiera się zazwyczaj **grupę genów z największą wariancją - top100** .

Skraca to czas obliczeń i redukuje szum w danych.



Dendrogram for clustering Leukemia patients (Chiaretti et al., 2004)
with 100 top variance genes

Analiza eksploracji danych - podsumowanie

- Poszukiwanie genów o podobnym profilu ekspresji dla kilku próbek
- Geny, które uległy podobnej ekspresji (co-expression) z dużym prawdopodobieństwem są regulowane przez te same czynniki lub też maja podobną funkcję
- Metody eksploracji danych znajdą jakieś wzorce w danych, nie ważne, czy będą one znaczące dla nas czy też nie
- Metody obejmują klastrowanie (np. hierachiczne, podział, k-średnich) oraz projekcję (principal component analysis, skalowanie wielowymiarowe)
- Taki rodzaj analizy powinien być używany tylko w przypadku, gdy nie mamy żadnej wiedzy, którą moglibyśmy wykorzystać

Klasyfikacja

Analiza dyskryminacyjna: KLASY ZNANE

Klastrowanie: KLASY NIEZNANE

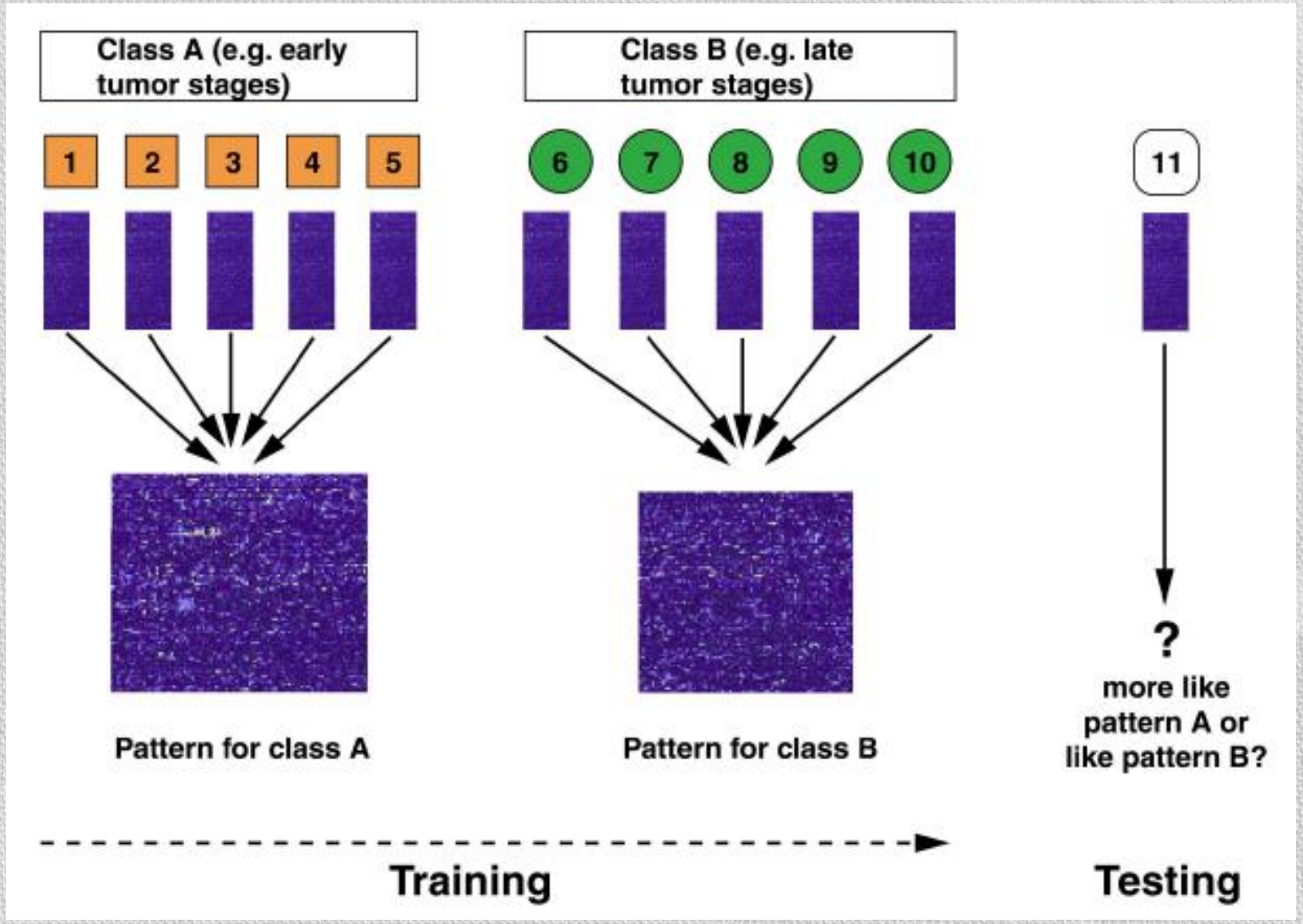
Analiza dyskryminacyjna

Mamy: po kilka próbek z badanych klas (np. rodzaj choroby pacjenta)

Chcemy : znaleźć wzorce klas z podzbioru genów, za pomocą których będziemy przydzielać nowe próbki do zadanych klas.

Metody klasyfikacji mogą być użyte do diagnostyki, np. aby stwierdzić na jaką chorobę cierpi pacjent, lub aby przewidzieć sukces lub porażkę zastosowanej terapii.

Analiza dyskryminacyjna



Ocena klasyfikatorów

Dla klasyfikacji binarnej – jedna z klas posiada szczególne znaczenie, np. zdiagnozowanie poważnej choroby

Oryginalne klasy	Przewidywane klasy decyzyjne	
	Pozytywna	Negatywna
Pozytywna	*TP*	*FN*
Negatywna	*FP*	*TN*

TP – liczba poprawnie sklasyfikowanych przykładów z wybranej klasy

FN – liczba błędnie sklasyfikowanych przykładów z tej klasy, tj. decyzja negatywna, podczas gdy w rzeczywistość przykład należy do zadanej klasy

TN – liczba przykładów poprawnie nie przydzielonych do wybranej klasy

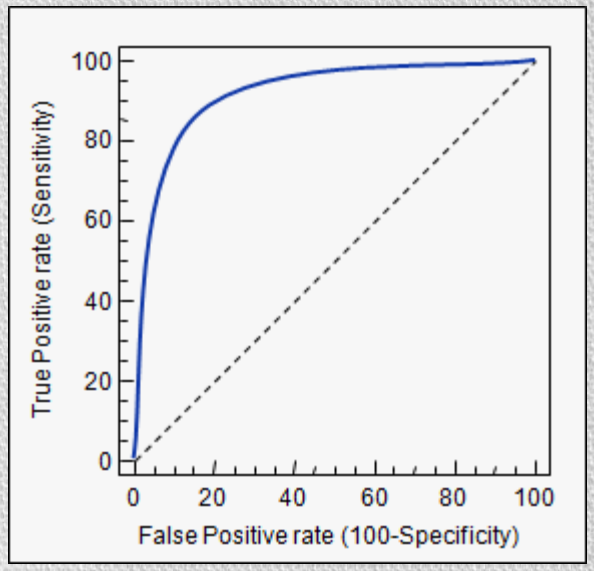
FP – liczba przykładów błędnie przydzielonych do zadanej klasy, podczas gdy do niej nie należą

Receiver Operating Characteristic (ROC) curve

Wykres true positive rate (*sensitivity*) **TP/(TP+FN)** w funkcji false positive rate (*100-Specificity*) **FP/(FP+TN)**

Sensitivity – czułość, prawdopodobieństwo że rezultat testu będzie pozytywny jeśli faktycznie powinien być

Specificity – swoistość, prawdopodobieństwo, że test będzie odrzucony jeśli faktycznie powinien być odrzucony



Każdy punkt przedstawia parę czułość/swoistość

Test z idealną dyskryminacją plasowałby się w lewym górnym narożniku

Sposoby na tworzenie klasyfikatorów

- Technika podziału **hold-out** *dla dużych zbiorów danych*
  - podział na dwa niezależne zbiory: uczący (2/3) i testowy (1/3)
- Ocena krzyżowa **cross-validation** *dla średnich zbiorów danych*
  - podział losowy danych na *k* pozdbiorów
  - użycie *k*-1 zbiorów jako części uczącej i jednej jako testującej
  - uśrednienie wyników z *k* iteracji
- **Bootstrapping** i **leaving-one-out** *dla małych zbiorów danych*

- Problem przeuczenia (*overfiting*) – nadmiernie dopasowane do specyfiki danych uczących się, bez dobrego uogólnienia

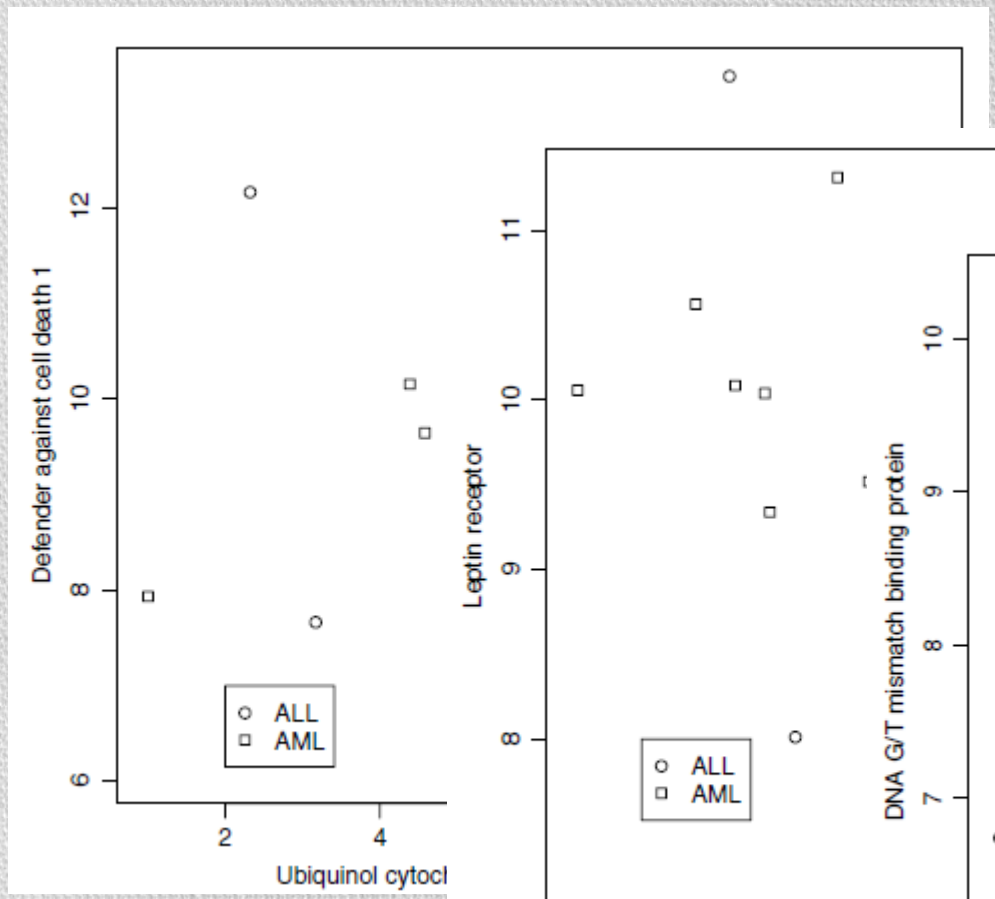
Które geny wybrać do tworzenia wzorca?

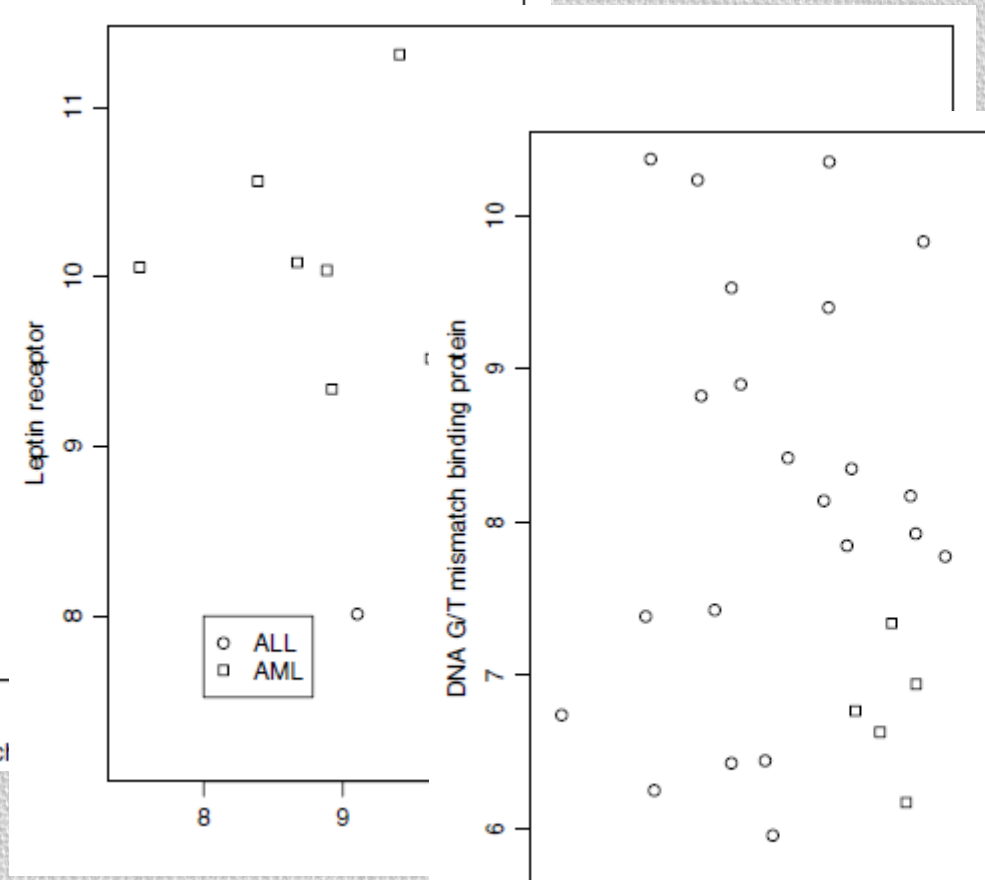
- Wybrać **najbardziej dyskryminujące geny**
- Geny mogą mieć znacząco zróżnicowaną ekspresję, ale mogą nie być przydatne do klasyfikacji

Heurystyczny wybór grupy genów

- Filtrowanie
  - Rangowanie genów zgodnie z ich zdolnością do dyskryminacji (statystyka t, Wilcoxon)
  - Używanie pierwszych *k* do klasyfikacji
- Shrinkage (kurczenie się)
  - Cały zbiór genów zmniejszamy o te geny, które nie mają wpływu na klasyfikację – kontunuujemy, dopóki nie osiągniemy satysfakcjonującego wyniku
  - Np. *Nearest Shrunken Centroids*

Czy dane można rozdzielić?





Separowalne nieliniowo

Czy dane można rozdzielić?

- Dane można rozdzielić liniowo czy nieliniowo?

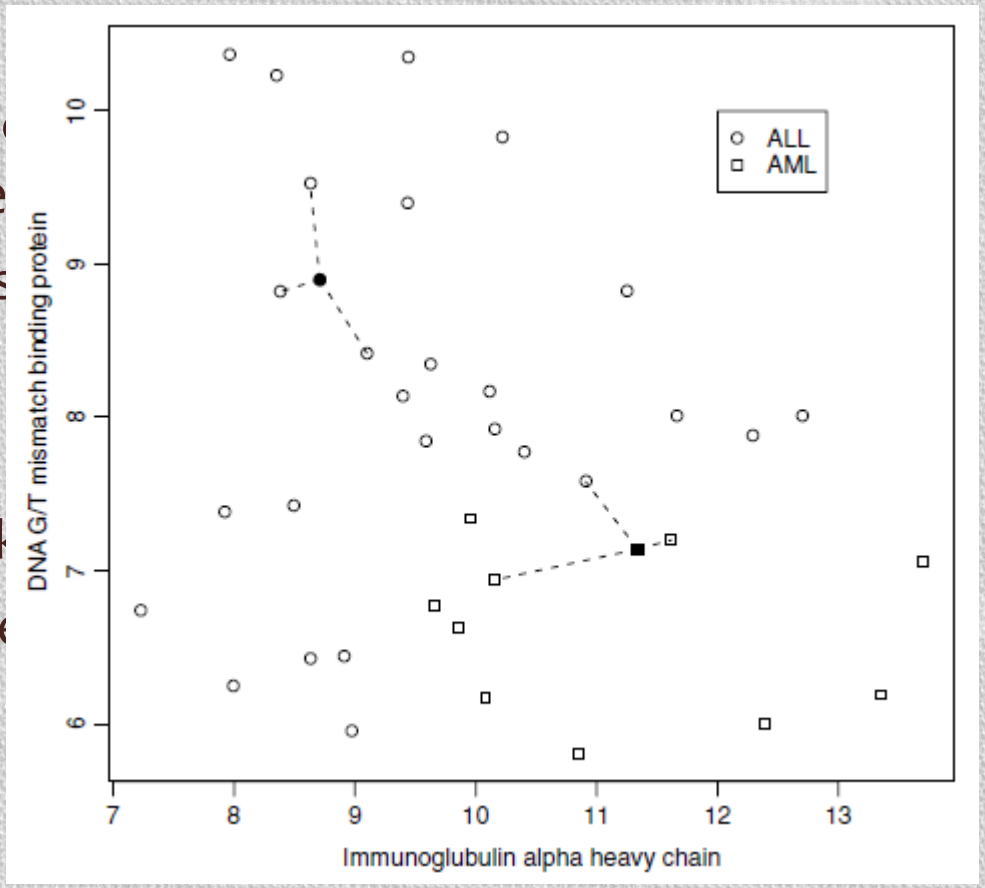
- Ile klas można jednocześnie ze sobą rozdzielić? Niektóre algorytmy potrafią rozdzielić tylko dwie klasy, inne mogą działać dla większej liczby klas

K- nearest neighbors

- Patrzymy na pomiar ekspresji genów dla próbki, którą chcemy sklasyfikować
- Szukamy najbliższe ze znany
  miarą odległości (np. Euklide
- Klasa próbki jest taka jak klas

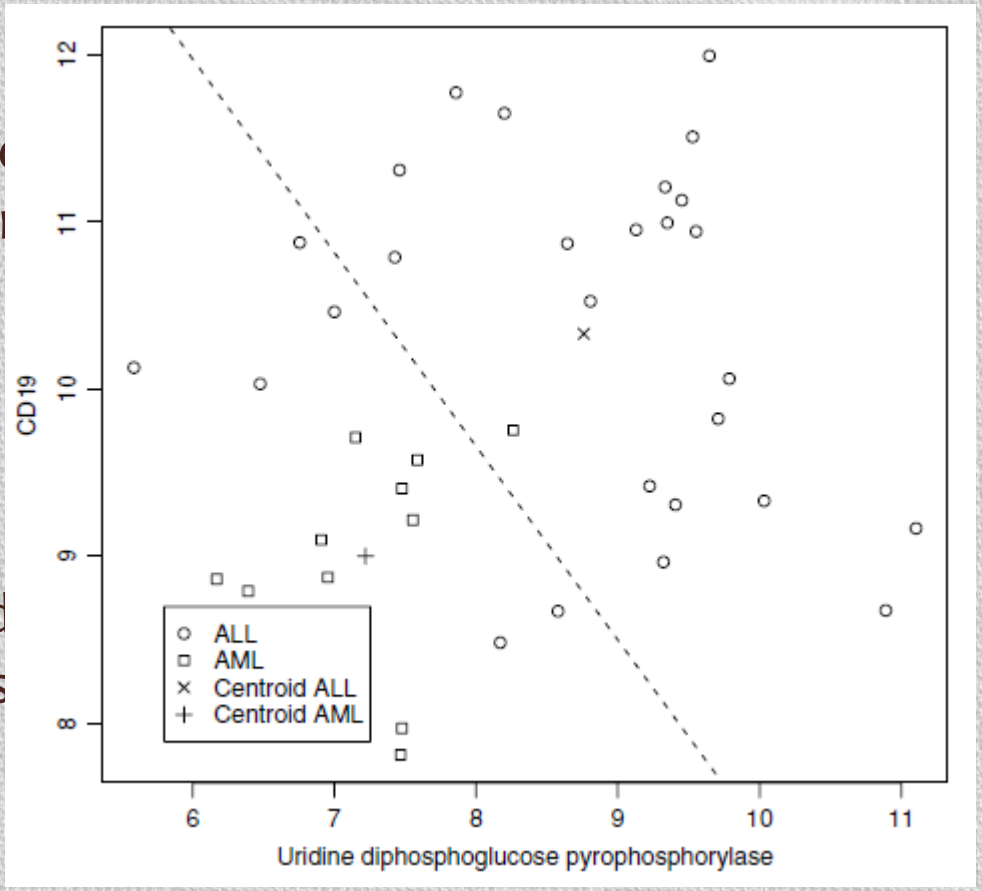
Parametry

- *k* – liczba najbliższych próbek
- *l* – najmniejszy margines, dzi



Centroid classification

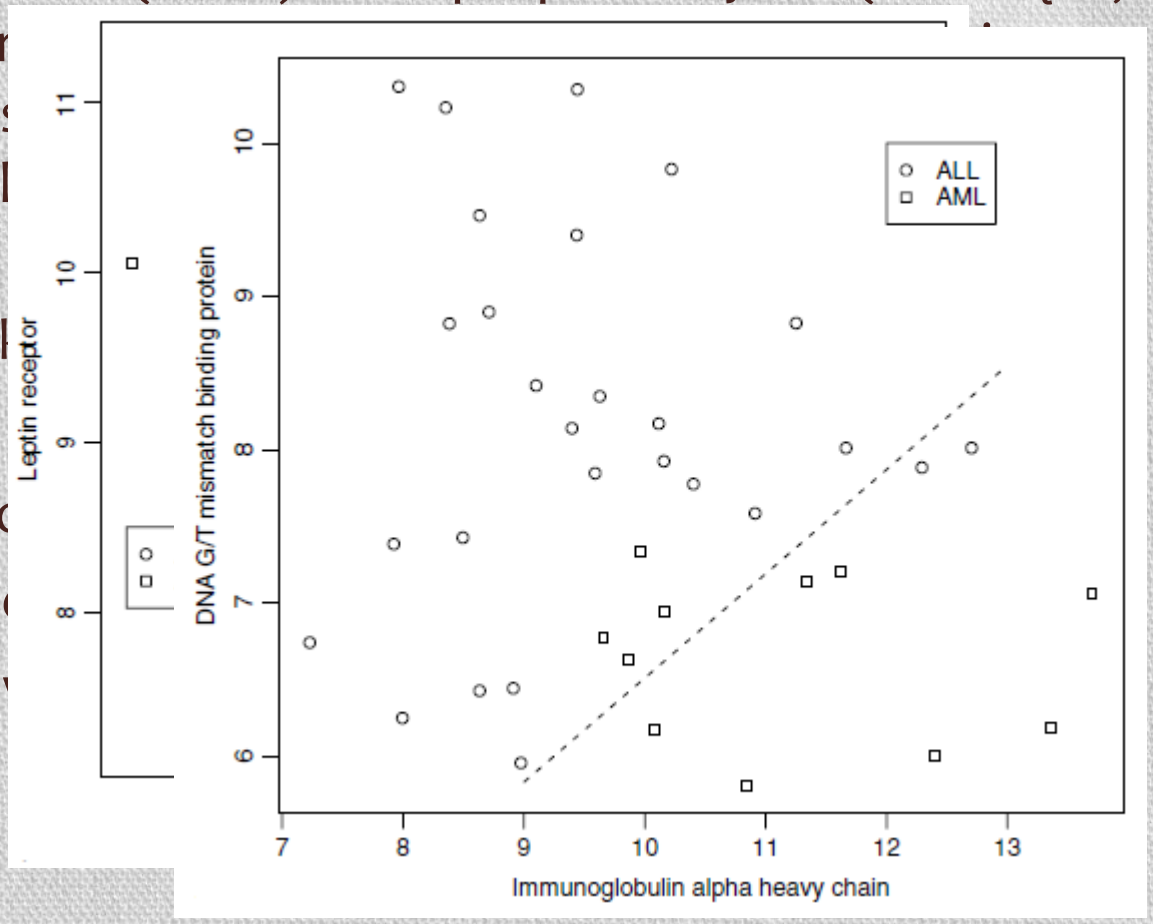
- Dla każdej klasy wyznaczamy środek ciężkości (centroid) ze wszystkich punktów z klasy
- Wyznaczamy odległość pomię
  centroidów (odpowiednia mia
- Przypisujemy próbkę do klasy,



Algorytm jest prosty, może rozd
kilkoma klasami, ale nie nadaje s
nieseparowalnych liniowo.

Liniowa analiza dyskryminacyjna (LDA)

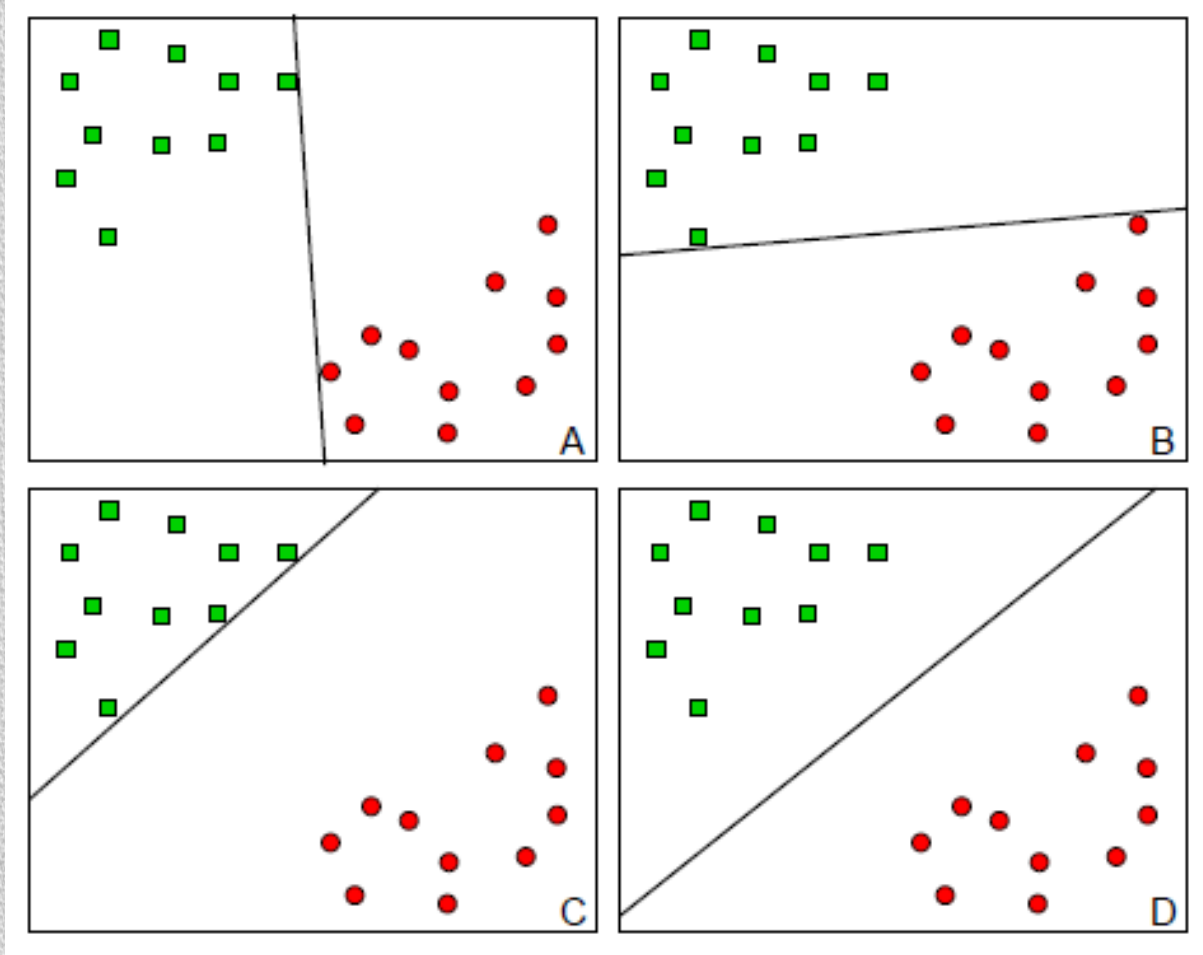
- Wyznaczana jest linia prosta (w 2D) lub hiperpłaszczyzna (dla więcej niż 2 wymiarów), która r
  wariancję wewnątrz klas
  dzielącej oraz maksymal
- Nowa, nieznana próbka
  zależności od tego, po k
- LDA jest mocno rozbudo
- Można rozdzielać tylko
- Klasy muszą być separo



Support vector machines (SVM)

SVM podobnie jak LDA rozdzielają liniowo przestrzeń na dwie części

Który podział jest najlepszy?

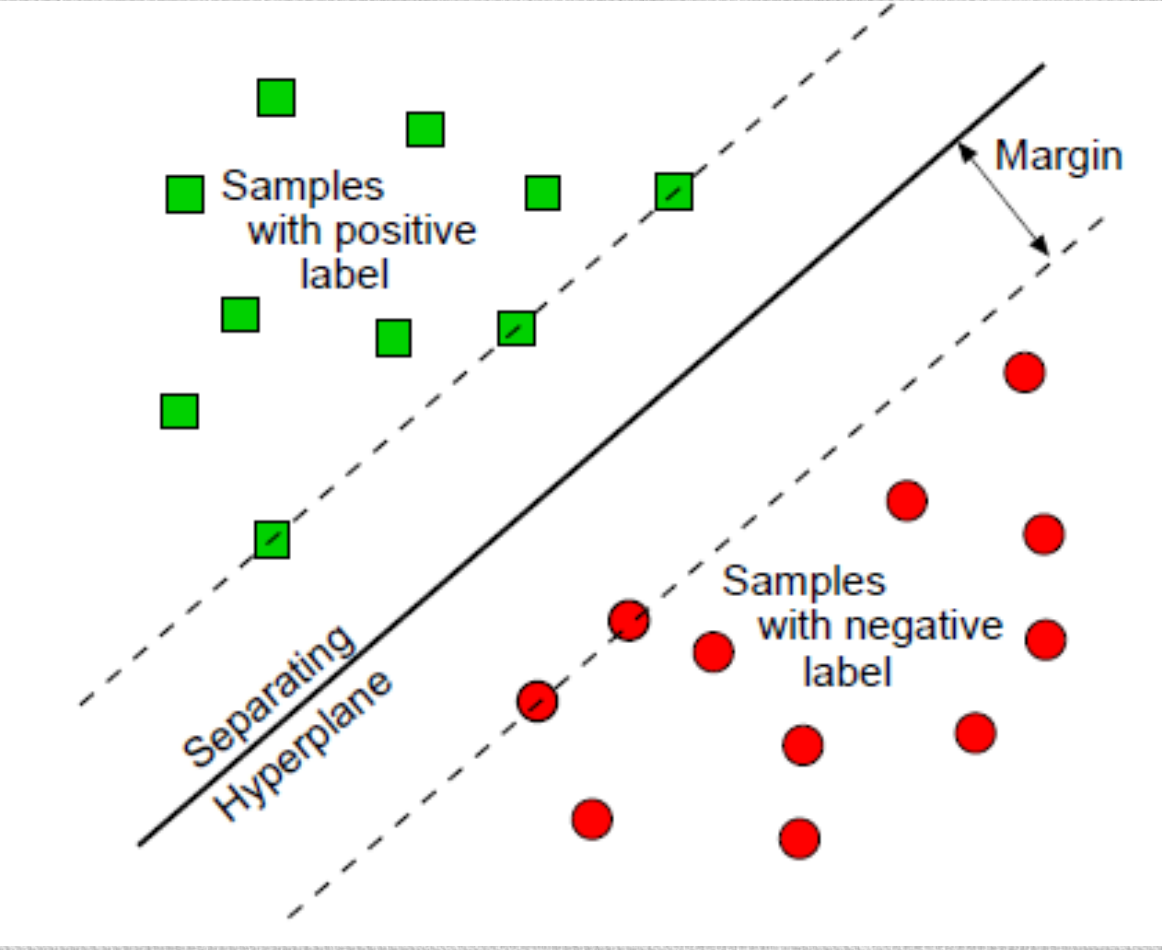


Support vector machines (SVM)

Najlepszy podział ma największy margines

Punkty leżące najbliżej hiperprzestrzeni nazywają się *support vectors*.

Pozostałe punkty nie mają wpływu na pozycję hiperprzestrzeni



Support vector machines (SVM)

Najlepszy podział ma największy margines

Maksymalizuj L(α)

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

Przy ograniczeniach

$$\alpha_i \geq 0, \ \forall i \ \sum_{i=1}^{N} \alpha_i y_i = 0$$

Rozwiązanie (α>0 dla i$\in$SV) ; b – odpowiednio uśredniane

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$

Hiperpłaszczyzna decyzyjna

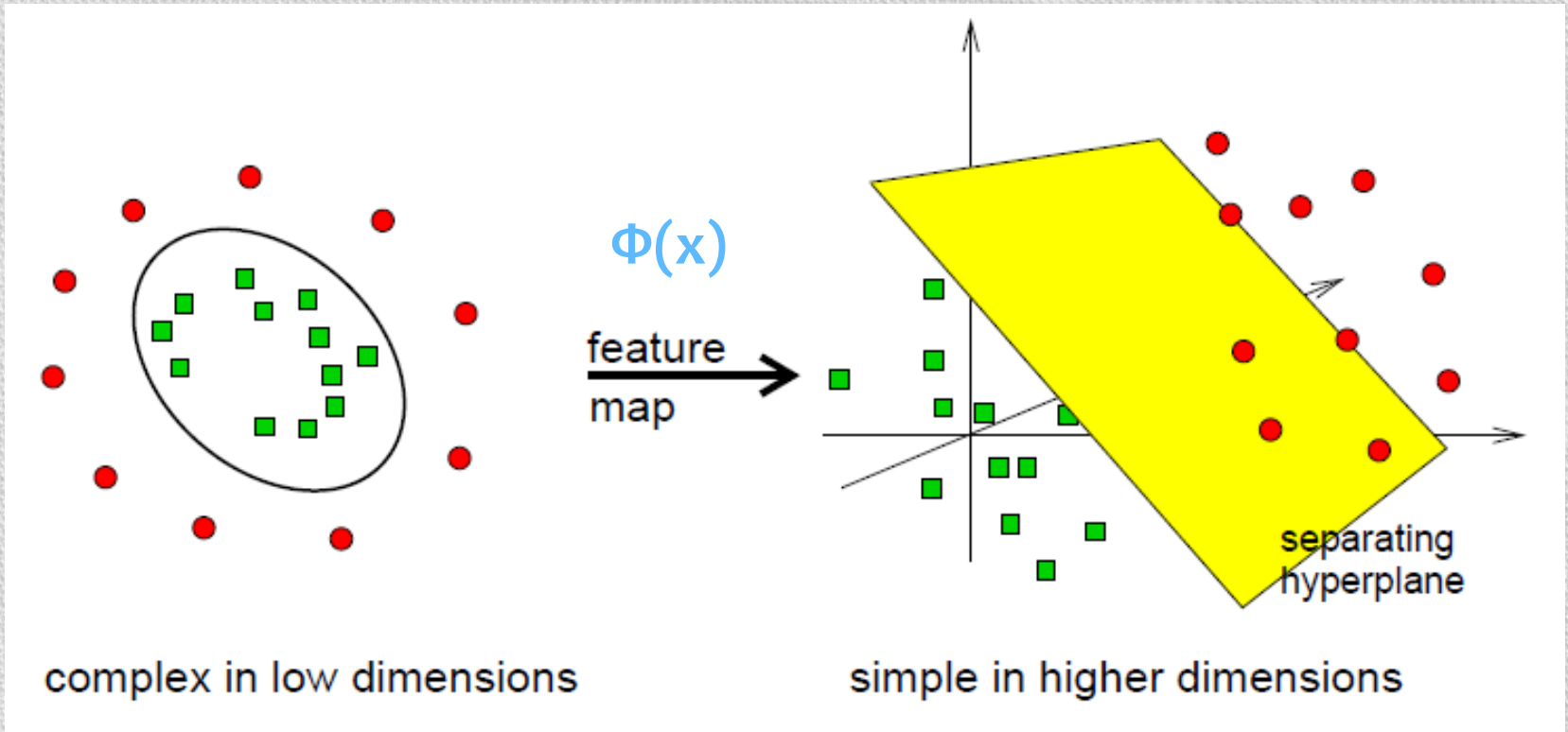
$$\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b = 0$$

Klasyfikacja nowego wektora x poprzez funkcję decyzyjną

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b)$$

Support vector machines (SVM)

SVM jednak potrafią rozdzielić dane, których nie można liniowo odseparować dzięki projekcji danych w przestrzeń o większym wymiarze (*kernel function*)



Support vector machines (SVM)

Załóżmy, że funkcja jądra K ma postać

$$K(x,y) = \Phi(x)*\Phi(y)$$

Dopuszczalne funkcja jądra (*kernel functions*)

Normalne (Gaussowskie)	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\frac{(\mathbf{x}_i - \mathbf{x}_j)}{2\sigma^2}\}$
Wielomianowe	$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + d)^p$
sigmoidalne	$K(\mathbf{x}_i, \mathbf{x}_j) = tgh(\kappa \mathbf{x}_i \cdot \mathbf{x}_j - \delta)$

Support vector machines (SVM)

Nie potrzebujemy wiedzieć jak wygląda przestrzeń zmiennych przekształconych (*feature space*), ani nie potrzebujemy znać funkcji Φ(x), wystarczy że znamy funkcję jądra, jako miarę podobieństwa.

Funkcję jądra K(x,y), którą możemy podstawić bezpośrednio do wcześniejszych wzorów

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{N} \alpha_i y_i \Phi(\mathbf{x}_i)\Phi(\mathbf{x}) + b\right)$$
$$sign\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right)$$

The kernel trick:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \bullet \mathbf{x}_j)^2 = \left(x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2\right) \bullet \left(x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2\right)$$
$$= \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

Original optimization function:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i \cdot \mathbf{x}_j)$$

Materiały:

- Jőrg Rahnenfűhrer „NGFN – Courses in practical DNA microarray analysis”, 2005
- Dov Stekel „Microarray Bioinformatics”, Cambridge University Press, 2003
- Florian Markowetz, „NGFN – Courses in practical DNA microarray analysis”, 2005
- Jerzy Stefanowski, prof. PP, przedmiot „Uczenie maszynowe”