Laboratorium nr. 6

1. Wczytywanie bibliotek

source("http://bioconductor.org/biocLite.R")
biocLite("ALL")
library("ALL")
Podobnie wczytaj następujące moduły: genefilter, hgu95av2.db, bioDist, RColorBrewer, cluster

2. Zestaw danych ALL

Zestaw danych ALL składa się ze 128 mikromacierzy dla różnych osobników chorych na białaczkę (acute lymphoblastic leukemia). Jest 95 próbek z białaczką typu B-cell i 33 z białaczką typu T-cell. Ponieważ to są różne tkanki i różne choroby zazwyczaj do analizy wybierany jest tylko jeden typ choroby

3. Wybór próbek i genów (*zrzut sesji*)

Wybierz z danych tylko próbki typu B, dla których fenotyp biologii molekularnej jest równy BCR/ABL lub NEG. Aby dalej zawęzić listę genów do analizy, wybierz te, które są czynnikami transkrypcyjnymi. Pomoc w wyborze próbek i genów znajduje się w pomocy do ćwiczeń nr 2 (str. 138-139), a wytłumaczenie poszczególnych funkcji R w pomocy do ćwiczeń nr 1.

4. Wyznaczanie odległości pomiędzy genami (*3 wykresy + zrzut sesji + komentarz*)

Wycentruj i przeskaluj dane z macierzy ekspresji, a następnie wyznacz macierz odległości za pomocą różnych metryk *manhattan, euclidean, minkowski* tworząc odpowiednio obiekty ***manDist***, ***eucDist***, ***minkDist***. Narysuj 3 heatmapy (dla każdej odległości jedną), które zobrazują wyznaczone odległości. Czy są zauważalne różnice pomiędzy wykresami? Jakie?
Skorzystaj z pomocy do ćwiczeń nr.2 (str 140/141)

5. Klastrowanie hierarchiczne (*3x3 wykresy + zrzut sesji +komentarz*)

Dla każdej z wyznaczonych miar zastosuj algorytm klastrowania hierarchicznego typu *bottom-up* (***hclust***) z metodą *complete linkage* oraz *single linkage* oraz typu *top-down* (***diana***). Powinno powstać 9 różnych obiektów (3 miary odległości * 3 metody klastrowania). Przedstaw dendrogramy, pogrupuj je funkcją *par(mfrow=c(3,1))* dla tej samej miary odległości. Czy są zauważalne różnice pomiędzy wykresami? Jakie?
Za pomocą funkcji *cutree* przetnij drzewo z hierarchicznego klastrowania typu bottom-up dla complete linkage, tak aby utworzyć 2,3,4,6 klastrów i utwórz obiekty ***hcut2***,***hcut3***,***hcut4***,***hcut6***.
Skorzystaj z pomocy do ćwiczeń nr.2 (str 144/145)

6. Klastrowanie kmeans i PAM (*zrzut sesji* + *komentarz*)

Wykonaj klastrowanie metodą kmeans (funkcja *kmeans*) podając jako zbiór wejściowy zmienną ***gvals*** (utworzoną w punkcie 4, jeśli nie jest ona utworzona wróć do pomocy, str 140-1), k=2,3,4 i 6, a liczbę restartów=15 (algorytm kmeans zwraca często różny wynik, dlatego ważne jest powtórzenie obliczeń). Utwórz obiekty o nazwach ***km2***, ***km3***, ***km4***, ***km6***.
Wykonaj klastrowanie PAM dla dwóch różnych miar manDist i eucDist, k podobnie jak wcześniej. Utwórz obiekty o nazwach ***pamMk*** i ***pamEk***, np. ***pamM2.***

Porównaj klastrowania dla k=3: hcut3 vs km3, km3 vs pamE3, pamE3 vs pamM3, metodą ***table***, np. table(hcut3,pamM3$cluster).
Czy różne metody klastrowania zwracają takie same klastry, czy różne? Które z metod dają jako wynik najbardziej zbliżone klastry? (str 146-7)

7. Ocena liczby klastrów – silhouette (*4 wykesy* + *zrzut sesji* + *komentarz*)

Aby ocenić czy dobrze została wybrana liczba klastrów można wykorzystać szerokość silhouette (funkcja *silhouette*). Dla każdego z utworzonych obiektów klastrowania ***pamMk***, ***pamEk***, ***hcutk***, ***kmk*** wywołaj tę funkcję. Obiekty typu *pam* można podać jako jedyny parametr do funkcji *silhouette*, gdyż obiekt tego typu posiada wyznaczoną miarę odległości, natomiast dla obiektów ***hcutk*** i ***kmk*** należy podać jako drugi parametr wyznaczoną macierz odległości ***eucDist***. Obiekty ***kmk*** wymagają wyciągnięcia tabeli z klastrowaniem, np. ***km2$cluster***. Wyniki przedstaw na 4 wykresach, pogrupowanych wg typu klastrowania (na każdym wykresie będą 4 wykresy dla różnej liczby klastrów, ale jednej metody klastrowania).
O czy mówi szerokość silhouette? Jaka liczba klastrów jest najlepsza? Czy w zależności od metody klastrowania silhouette wskazuje na różną liczbę klastrów?