BIOINFORMATYKA 2 – KURS MAŁY

Adrian Kania

PubMed przez E-Utilities

Baza danych PubMed zawiera informacje dotyczące artykułów z zakresu nauk biologicznych i medycznych. Wyszukiwanie może odbywać się poprzez przeglądarkę internetową. W tym ćwiczeniu jednak posłużymy się skryptami napisanymi w języku Python, a korzystającymi z funkcjonalności ESearch oraz EFetch, dostępnych w NCBI E-Utilities.

Zadanie1 Odszukaj pracę o identyfikatorze *14697198* w bazie danych *PubMed,* a następnie odpowiedz na pytania:

- jaki jest tytuł tej pracy?

- w jakim wydawnictwie została wydana ta praca?

- ilu autorów odpowiada za tę prace?

- podaj przykładowe MeSH termy z tej pracy.

Zadanie2 Zaproponuj hasło do wyszukania prac posiadających w tytule *mRNA*, opublikowanych w *2017* roku w czasopiśmie *BMC Genomics*. Ile jest takich prac?

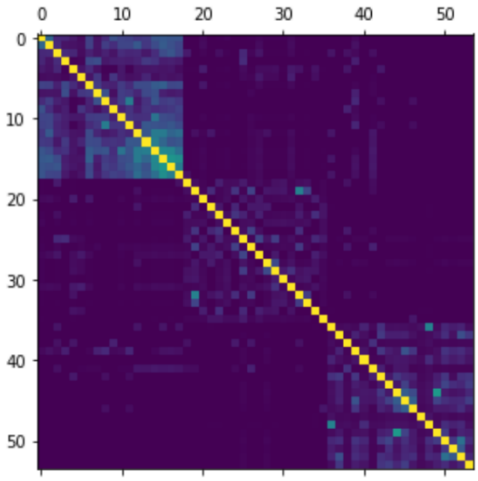
Zadanie3 Pobrano *MeSH* termy dla 54 prac dotyczących trzech kategorii:

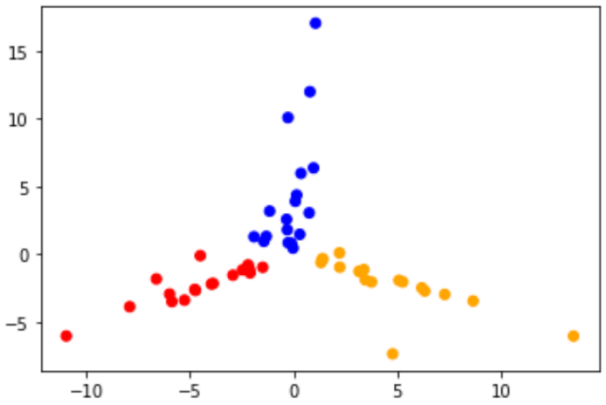
- pierwsze 18 prac dotyczyło *Arabidopsis*,

- kolejne 18 prac dotyczyło *modelowania molekularnego*,

- ostatnie 18 prac dotyczyło *Helicobacter Pylori.*

Następnie zastosowano metodę TF-IDF celem wektorowej reprezentacji każdej z pracy. Poniżej zamieszczono macierz podobieństw kosinusowych (im jaśniejszy kolor tym większe podobieństwo) oraz wynik działania algorytmu PCA dla 2 komponentów (tutaj kolory odpowiadają kategoriom).





Która grupa wydaje się być najbardziej jednorodna na podstawie macierzy podobieństwa?

Z czego mogą wynikać jaśniejsze pola występujące poza grupami?

Na podstawie powyższych wykresów stwierdź czy rozważane kategorie są separowalne.

Dane genetyczne

Zadanie 4 Wyznaczono poziom ekspresji genów dla 6 kolejnych chwil. Wyznacz współczynnik korelacji pomiędzy poziomem ekspresji genu A oraz genu B (możesz skorzystać z portalu: Correlation coefficient calculator (statskingdom.com)). Zinterpretuj uzyskany wynik.

Gen 1	0.7	0.74	0.86	0.83	1.2	1.31
Gen 2	1.5	1.32	1.11	0.84	0.72	0.60

GeneMania

Serwis ten pozwala na zbiorczą analizę grupy genów. Tworzona jest sieć połączeń ze względu na takie cechy jak: genetyczne interakcje, fizyczne interakcje, współdzielone domeny, kolokalizację czy koekspresję.

Zadanie 5 Wejdź na http://genemania.org/ a następnie przeanalizuj zestaw potencjalnych genów markerowych dla prognozy raka piersi.

EFNA1

EGFR

ERBB2

GATA3

GZMB

MST1

MYB

MYBL2

MYC

PLAT

SOX4

SOX9

SRF

XBP1

Który gen jest połączony z *GATA3* jeżeli chodzi o kolokalizację?

Który gen jest połączony z *XBP1* jeżeli chodzi o genetyczną interakcję?

Jaka występuje najbardziej znacząca funkcja w tej grupie genów (tzn. taka która ma najmniejszy FDR).

NCBI (GEO DataSets) udostępnia dane z eksperymentów mikromacierzowych a także pozwala na ich analizę online. W tym kroku poddamy analizie wybrany taki eksperyment.

Zadanie 6 Przypadek do analizy z NCBI GEO DataSets

- Wejdź w link poniżej: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS810
- Czego dotyczyło badanie?
- Gdzie badano ekspresje genów? (jaki materiał/tkanka)
- Ile było wszystkich próbek i na ile grup były one podzielone ***(Experiment design and value distribution oraz Sample Subsets)***
- Według jakich kryteriów podzielono próbki na grupy? Opisz wykorzystane parametry. ***(Experiment design and value distribution oraz Sample Subsets)***
- Jak wygląda przebieg ekspresji dla genów *SPARC, VSNL1* oraz *COL5A2* w kolejnych grupach? ***(Expression Profiles)*** Za co odpowiadają te geny? Czy obserwujesz jakieś tendencje zmiany poziomu ich ekspresji w kolejnych grupach? Poszukaj w źródłach zewnętrznych informacji na temat ich związku z chorobą Alzheimera.
- Czym są *housekeeping genes?* Jaką pełnią rolę w eksperymencie mikromacierzowym? Wybierz trzy przykładowe geny tej kategorii i sprawdź ich ekspresje w kolejnych próbkach.

- W wyniku analizy podanej mikromacierzy otrzymano zestaw następujących genów różnicujących. Przeanalizuj ich funkcje z użyciem *GeneMania*.

HSPA1L

UCHL1

GJA1

SNAP25

MMD

VSNL1

HPCA

CERS1

HBB

GNAS

ACTG1

SPARC

ABR

MT1F

RGS1

GAPDH

UQCR10

COL5A2

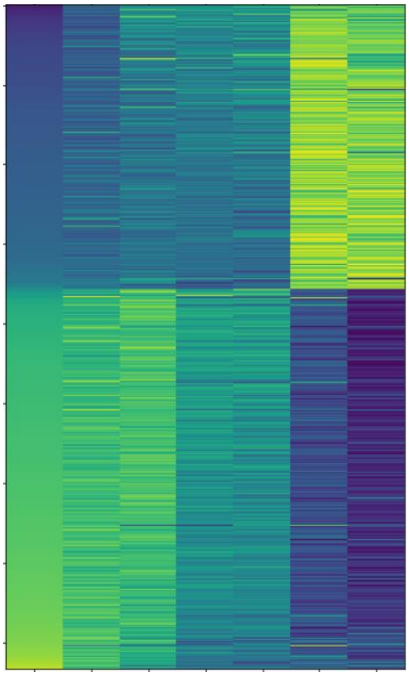
Zadanie 7 Poniżej zamieszczono dane odnośnie ekspresji genów w komórkach drożdży podczas procesu oddychania – fermentacji alkoholowej. Wyróżniamy dwa główne etapy tego procesu:

- rozkład glukozy do kwasu pirogronowego,

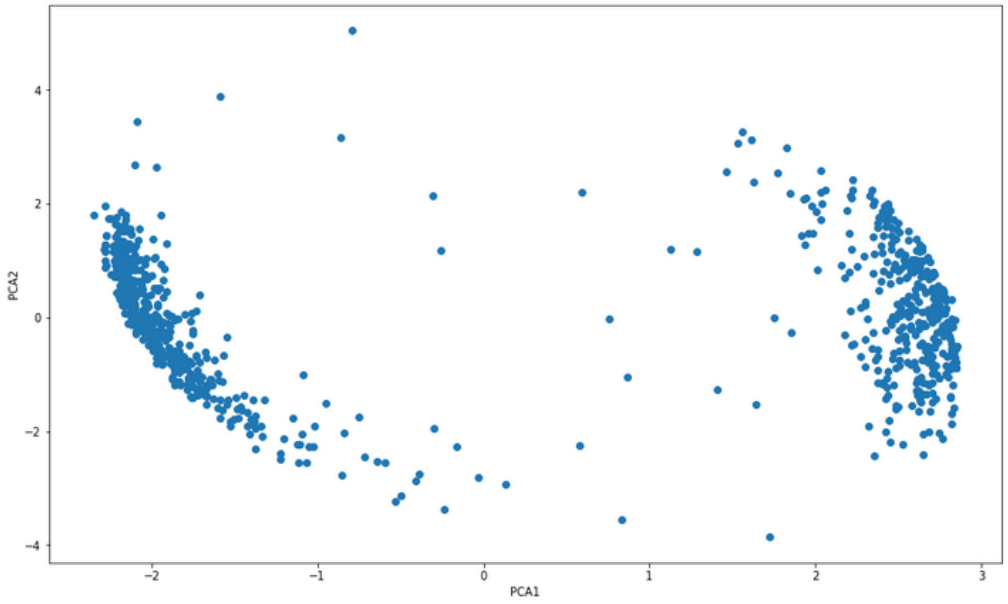
- przemianę kwasu pirogronowego do alkoholu.

Każdy z etapów kontrolowany jest przez 2 klasy genów odpowiedzialnych za te procesy. Dane pochodzą z 7 chwil czasowych (kolejne kolumny). Skomentuj i porównaj poniższe wyniki w kontekście powyższych informacji. W jaki sposób podzieliłbyś rozważane chwile czasowe?

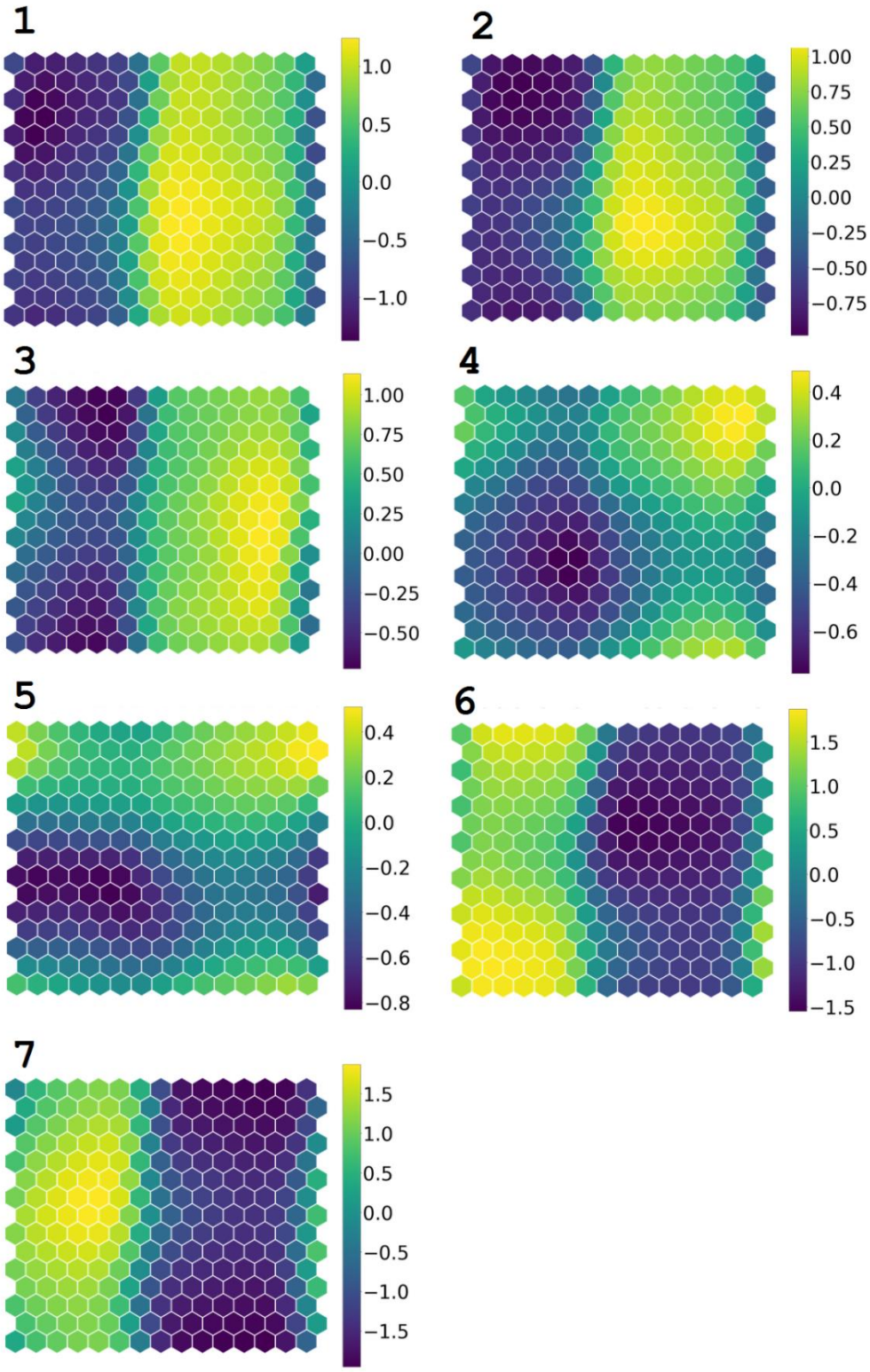
- rozważana mikromacierz (geny zostały posortowane względem pierwszej chwili czasowej):



- po zastosowaniu algorytmu PCA, gdzie jako kolejne obserwacje wybrano wiersze (geny) z powyższej macierzy:



- mapy Kohonena dla kolejnych chwil czasowych. Kolor wskazuje na poziom ekspresji określonej grupy genów:



Zadanie 8 Na koniec zobaczymy jak na ekspresję genów istotny wpływ może mieć epigenetyka. Wejdź na stronę [WashU EpiGenome Browser (wustl.edu)] zawierającą dane ATAC-seq. Piki wskazują na dostępność chromatyny. Wybierz *Corces_scATAC_BroadCellTypes*. Zlokalizuj pozycję chr19:6771658-6774320. Jaki gen znajduje się w tym zakresie? Jakie są jego funkcje? W których komórkach mózgowych wydaje się szczególnie ulegać ekspresji?