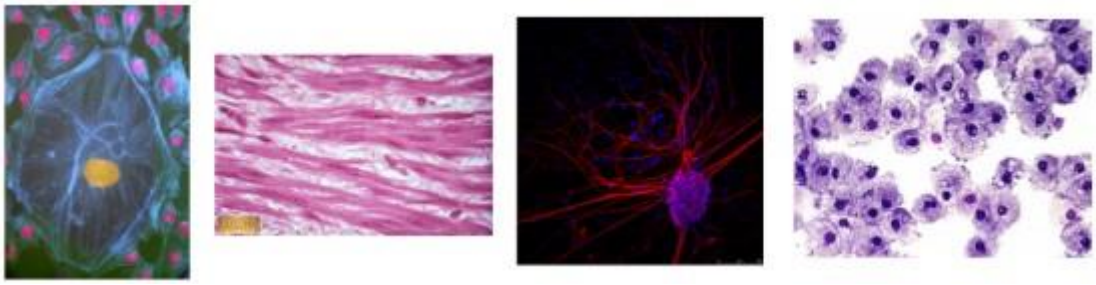
Instytut Informatyki i Matematyki Komputerowej UJ,
opracowanie: mgr Ewa Matczyńska, dr Jacek Śmietański

Mikromacierze

1. Mikromacierze – wprowadzenie

Mikromacierze to technologia pozwalająca na pomiar aktywności genów w komórce. Dla przypomnienia, na pierwszych zajęciach dowiedzieliśmy się, że to co różnicuje komórki to inny zestaw genów, ulegający ekspresji tzn. inny zestaw genów jest poddawany procesowi transkrypcji i translacji na białko.



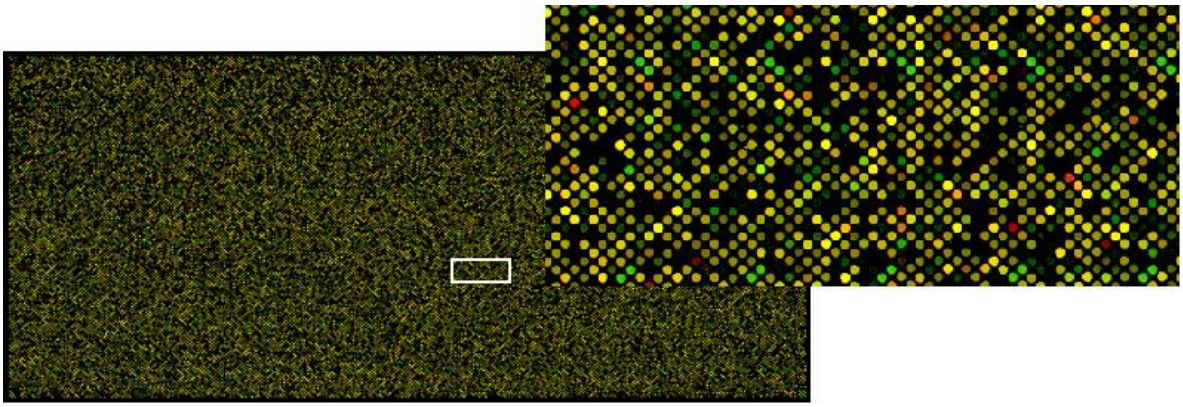
Rysunek 1: Różne typy komórek ludzkich.

Dzięki temu, mimo tego iż w każdej komórce jest dokładnie taki sam genom, komórki spełniają różne funkcje. Okazuje się, że zestaw genów, które ulegają ekspresji może się różnić także gdy mamy np. komórkę tego samego typu, ale w jednej zachodzą pewne zmiany nowotworowe, a druga jest zdrowa. Dlatego pomiar ekspresji genów wydaje się być obiecującą technologią dla wspomagania diagnozy, leczenia, medycyny personalizowanej.

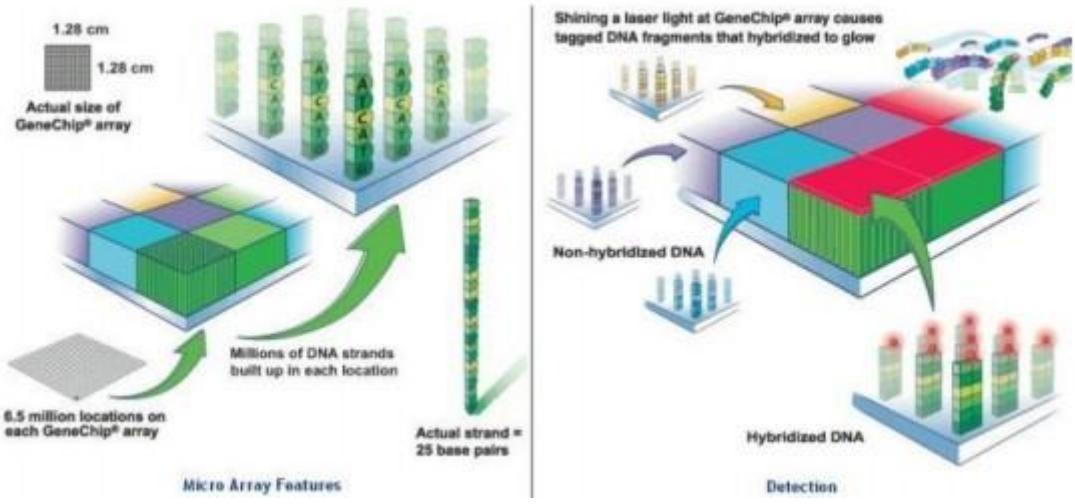
Rysunek 2: Mikromacierz.

Zasada działania mikromacierzy opiera się o pomiar ilości mRNA danego genu w komórce. Jak pamiętamy, mRNA jest matrycą na której powstaje białko. W komórce dość szybko ulega ono degradacji, dlatego pomiar jego ilości pośrednio przenosi się na pomiar aktywności danego genu w komórce. Mikromacierz konstruowana jest jako bardzo mały chip, podzielony

na wiele tysięcy pól, tzw. spotów, gdzie przyczepione są nici komplementarne do pewnych charakterystycznych sekwencji genowych. Każde pole jest związane z określonym genem. Z komórki, którą chcemy zbadać wyizolowywane jest mRNA (przepisywane jest do cDNA - complementary DNA), do którego przyczepiane są znaczniki fluorescencyjne. Po nałożeniu tak przygotowanej próbki na płytkę nici, które będą komplementarne połączą się (zhybrydyzują), wtedy określone pole będzie wykazywało fluorescencję. Im więcej nici przyłączy się do danego pola, tym poziom fluorescencji będzie większy. Wzbudzoną światłem płytkę skanuje się i uzyskuje obraz, który następnie przetwarza się do danych liczbowych, mówiących o aktywności genów.



Rysunek 3: Obraz uzyskany po zeskanowaniu mikromacierzy.



Rysunek 4: Schemat działania mikromacierzy.

Z racji technologii pomiaru dane z mikromacierzy mogą być obciążone dużymi błędami i szumami wynikającymi zarówno z przebiegu eksperymentu jak i obróbki obrazu ze skanera. Nie mniej analiza danych ekspresji genów stanowi ważną gałąź bioinformatyki.

W dziedzinę badania ekspresji genów powoli wchodzą sekwencjonery nowej generacji. Są one lepsze od mikromacierzy, ponieważ są w stanie zsekwencjonować nawet te geny, o istnieniu których nie wiemy, więc nie wbudujemy ich sekwencji komplementarnej w mikromacierz. Nowa technika, nazwana RNA Seq, szybko zyskuje popularność i można przypuszczać, że w niedługim czasie niemal całkowicie zastąpi eksperymenty mikromacierzowe.

2. GEO

NCBI udostępnia bazę danych przechowującą dane z eksperymentów mikromacierzowych. Baza GEO – *Gene Expression Omnibus* udostępnia również pewne narzędzia do analizy mikromacierzy.
Podstawowymi typami rekordów w GEO są:

- GPL - platforma mikromacierzowa na której został wykonany dany eksperyment, posiada dokładny opis jakie geny związane są z danym polem;
- GSM - *Sample* - wynik jednego eksperymentu mikromacierzowego na określonej platformie GPL, zawiera poziomy ekspresji określonych przez GPL genów uzyskane w eksperymencie;
- GDS - *GEO Dataset* – zbiór danych złożony z wielu eksperymentów, pozwala analizować wiele eksperymentów na raz;
- GEO Profiles - profile ekspresji genów stworzone na postawie np. danego GDS

Zadanie: GEO (2 pkt)

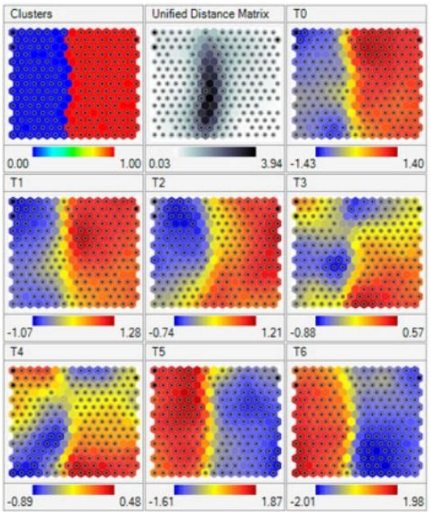
Zadaniem będzie przeanalizowanie konkretnego eksperymentu mikromacierzowego przy wykorzystaniu bazy GEO (*Gene Expression Omnibus*) Datasets.

1. Wejdź na NCBI, w wyszukiwarce wpisz *lung cancer* i wybierz bazę ***GEO Datasets*** (GDS).
2. Znajdź rekord o nazwie **GDS3309**, czego dotyczyło to badanie?
3. Obejrzyj link z platformą na której zostało wykonane badanie. Ekspresję ilu genów zmierzono w tym badaniu? Z jakiego organizmu pochodziły sekwencje? Jak odnaleźć te geny w NCBI?
4. Wróćmy do rekordu GDS, pliki z danymi można ściągnąć z GEO w różnym formacie:
   SOFT – plik *plain text* z danymi o ekspresji,
   MINiML - format z opisem eksperymentu w *xml*, oraz z każdym testem w oddzielnym pliku z danymi. Zobacz pliki po prawej stronie w sekcji *Download*.
5. U góry wybierz zakładkę ***Sample Subsets***. Ile mamy eksperymentów (GSM) w tym przykładzie? Czego dotyczą podzbiory?
6. Wybierz którykolwiek z linków do GSM, zobacz fragment tabeli z wynikami pomiaru ekspresji.
7. Wróćmy do rekordu GDS, u góry wybierz zakładkę ***Data Analysis Tools***.
   a) Wybierz analizę porównawczą 2 podzbiorów (compare 2 sets of samples) i wybierz *Two-tailed t-test* (test t-studenta bada hipotezę zerową: „średnie w obu grupach są równe”, w związku z tym wybierze nam te geny dla których średnie ekspresje w obu grupach były istotnie różne.
   b) Przypisz poszczególne podzbiory eksperymentów do grupy A i B (Step 2)
   c) Wykonaj test (Step 3)
   Jako wynik uzyskano profile genów, dla których hipoteza zerowa o średnich równych w obu grupach została odrzucona. Obejrzyj profile genów dla dwóch z nich, które wydają Ci się najbardziej obiecujące – ich ekspresja różni się znacząco w obydwu grupach. Sprawdź jaka jest ich funkcja w bazie *Gene*.
8. Wróćmy do rekordu GDS i zakładki ***Data Analysis Tools***. Wybierz sekcję ***Cluster heatmaps***.

a) Dla korelacji Pearsona porównaj metodę *Average Linkage* i *Single Linkage* jeśli chodzi o klastrowanie próbek - eksperymentów oraz klastrowanie genów.
b) Kliknij na metodę *k-średnich* i dla korelacji Pearsona poeksperymentuj z dzieleniem genów na różne liczby klastrów, zaobserwuj *patterny* dla klastrów.
c) Kliknij na klastrowanie pod względem lokalizacji na chromosomie, czy widzisz jakąś zależność między lokalizacją na danym chromosomie, a ekspresją?

3. Przykładowa analiza - na podstawie ekspresji genów drożdży

W pliku yeast_expression.txt mamy dane o ekspresji genów w komórkach drożdży podczas procesu oddychania. Dane jest 7 chwil czasowych w których został dokonany pomiar: T0, ..., T6. Drożdże oddychają w sposób beztlenowy - fermentacja alkoholowa. Są dwa etapy procesu oddychania drożdży:
- rozkład glukozy na kwas pirogronowy
- przemiana kwasu pirogronowego do alkoholu
Te dwa etapy są kontrolowane przez 2 klasy genów odpowiedzialnych za każdy z tych procesów. Można je dobrze wyraźnie zaobserwować na poniższych heatmapach utworzonych za pomocą tzw. samoorganizującej się mapy (self-organizing map - SOM).



Rysunek 5: Heatmapy SOM dla kolejnych chwil czasowych.

SOM jest modelem obliczeniowym przydatnym w wielu dziedzinach. Charakteryzuje się mapowaniem danych z wyżej wymiarowych przestrzeni w najczęściej 2 wymiary, zachowując strukturę tych danych. Ponadto jest odporny na szumy.

Na rysunku obrazującym oddychanie drożdży na SOM, pokazano tzw. heatmapy - kolor pokazuje nam odpowiednio wartość danej zmiennej na mapie w określonym wymiarze przestrzeni wejściowej. W tym przypadku klastrowaliśmy geny. Każdy gen był określony przez wektor 7 liczb, odpowiadającym poziomowi ekspresji w danej chwili czasowej. Wybranym wymiarem dla każdej heatmapy jest chwila czasowa, a kolor oznacza intensywność ekspresji.

Zadanie 2: klasteryzacja genów (2 pkt)

W pliku *yeast_expression.txt* znajdują się dane o ekpresji genów w komórkach drożdży podczas procesu oddychania. Pomiar został dokonany w 7 chwilach czasowych T0-T6.

Napisz skrypt który za pomocą metody K-średnich poklastruje chwile czasowe na dwa etapy oddychania drożdży. Można użyć gotowej metody *Bio.Cluster.kcluster*

Dokumentacja:
http://www.biopython.org/DIST/docs/api/Bio.Cluster.cluster-module.html

Przydatne metody do utworzenia macierzy ekspresji:

```
yeastMat = zeros((0,7))                         #deklaracja macierzy
yeastRow = asarray(mat(list))                   #lista do wektora
yeastMat = concatenate((yeastMat,yeastRow))     #konkatenacja macierzy
```

Rozwiązania obu zadań prześlij mailem do wtorku, **7.01.2020** włącznie, na adres:
jacek.smietanski@ii.uj.edu.pl
Temat wiadomości proszę opatrzyć przedrostkiem **[Bio] Lab 11**. Rozwiązaniem ma składać się z **dwóch plików** – zad 1: dokument pdf; zad.2: skrypt lub notebook zgodny z Pythonem w wersji 3.x**,** zawierający wszystkie niezbędne funkcje oraz procedurę wykonawczą. Proszę o nazwanie plików wg schematu**: Imie.Nazwisko.11.pdf, Imie.Nazwisko.11.py** (lub .ipynb).