POLITECHNIKA WARSZAWSKA
Wydział Elektroniki i Technik Informacyjnych
Instytut Radioelektroniki i Technik Multimedialnych

Rok akad. 2015/2016

PRACA DYPLOMOWA MAGISTERSKA

Lidia Maria Chrabąszcz

Analiza algorytmów tworzenia oraz walidacji sygnatur genetycznych z przykładami zastosowań do danych z The Cancer Genome Atlas

Kierownik pracy

prof. nzw. dr hab. inż. Przemysław Biecek

Ocena .....

.....
Podpis Przewodniczącego
Komisji Egzaminu Dyplomowego

Kierunek studiów: Inżynieria biomedyczna

Specjalność:

Nr albumu: 265968

Data urodzenia 16.02.1991

Data rozpoczęcia studiów 2.stopnia 02.2014

ŻYCIORYS

Urodzona w 1991 roku w Płocku. Ukończyła liceum im. Marsz. St. Małachowskiego w Płocku w klasie o profilu matematyczno – informatycznym. Uzyskała tytuł inżyniera na Akademii Górniczo – Hutniczej im. St. Staszica w Krakowie na kierunku inżynieria biomedyczna. W czasie studiów drugiego stopnia na Politechnice Warszawskiej ukończyła studia podyplomowe z zakresu zastosowań biostatystyki na Uniwersytecie Medycznym w Białymstoku.

.....
podpis

EGZAMIN DYPLOMOWY

Złożył(a) egzamin dyplomowy w dniu .....

z wynikiem .....

Ogólny wynik studiów .....

Dodatkowe uwagi i wnioski Komisji .....

.....

Streszczenie

W pracy poruszono tematykę związaną z analizą wzbogacenia funkcji biologicznych w zbiorze genów. Przedstawiono w niej algorytmy znajdowania grup podobnych genów oraz wyznaczania dla nich charakterystycznych funkcji biologicznych. Zastosowano podejścia wykorzystujące data mining oraz statystykę. Zaprezentowano działanie algorytmu na przykładzie danych z The Cancer Genome Atlas i opisano narzędzie stworzone w języku R umożliwiające przeprowadzenie przedstawionej w pracy analizy.

Słowa kluczowe: ontologia genów, analiza wzbogacenia, analiza danych, statystyka , data-mining

ANALYSIS OF ALGORITHMS OF CREATING AND VALIDATING GENE SIGNATURES WITH APPLICATIONS TO DATA FROM THE CANCER GENOME ATLAS

Summary

This thesis regards gene enrichment analysis. Algorithms of finding similar genes groups and determining the most characteristic biological functions for these groups are presented in the thesis. Two approaches are utilized: data mining and statistics. The implementation of the algorithm is presented based on the data from The Cancer Genome Atlas. The R package which enables the execution of the analysis presented in this thesis is also described.

Keywords: gene ontology, enrichment analysis, data analysis, statistics, data mining

Spis treści

Rozdział 1

Wprowadzenie

Obecnie ze względu na rozwój sprzętu oraz oprogramowania coraz powszechniej wykonywane są analizy dotyczące informacji powiązanych z genami [1]. Ze względu na koszty uzyskania tych informacji wiele z nich jest dostępnych *pro publico bono* w różnych źródłach internetowych. Repozytorium Bioconductor [2] zajmuje się zarządzaniem (gromadzeniem, udostępnianiem i kontrolą jakości) narzędzi stworzonych w języku R [3] dotyczących zagadnień bioinformatyki. Wspomniane repozytorium oraz środowisko R są darmowe. Jedną z głównych cech środowiska R jest możliwość tworzenia i udostępniania zespołów funkcji związanych z danym zagadnieniem (pakietów) przez dowolnego użytkownika.

Sztandarowym sposobem analizy ekspresji genów jest wyodrębnienie z pewnego zbioru (np. wszystkich genów organizmu) genów, których aktywność w istotny sposób różnicuje co najmniej dwa organizmy (np. chory i zdrowy człowiek). Kolejne etapy tej analizy to zgrupowanie wybranych genów i znalezienie dla nich charakterystycznych funkcji biologicznych. Metody użyte w każdym z kroków tego algorytmu są determinowane przez osobę przeprowadzającą dane doświadczenie. Wykorzystywane jest podejście statystyczne oraz data mining-owe. Istnieją wypracowane metody przeprowadzania kolejnych etapów jednak ich mnogość pozwala na tworzenie wielu sposobów postępowania (szczególną uwagę należy zwrócić na rozkłady prawdopodobieństwa cech). W ramach tej pracy tworzony jest algorytm przeprowadzania wyżej przedstawionej analizy oraz jego implementacja w środowisku R. Ponadto poruszane są zagadnienia dotyczące genów oraz stosowanych metod statystycznych, w tym data mining. W szczególności, spośród metod najczęściej mających zastosowanie w analizie genów, zostały użyte: analiza wariancji, klasteryzacja hierarchiczna oraz test Fishera. Dodatkowo wykorzystano metodę porównań wielokrotnych za pomocą testu Tukey'a, co nie jest powszechnie stosowane do podziału genów na grupy. Działanie algorytmu zostało przetestowane na danych z The Cancer Genome Atlas [4].

1.1. Cel pracy

Celem pracy jest zbadanie metody identyfikacji interesujących funkcji biologicznych charakteryzujących wybrany zbiór pacjentów i genów. W analizie tej metody wykorzystano zarówno istniejące narzędzia omawiane w artykułach naukowych powiązanych z analiza genów, jak i te niemające dotychczas zastosowania w omawianym kontekście.

Realizując cele pracy dodatkowo:

1. Zaimplementowano wspomniane algorytmy przy użyciu języka R.
2. Na podstawie wykonanych narzędzi stworzono pakiet *GOpro* w języku R. Pakiet ten służy do zastosowania omawianych w pracy metod związanych z wyznaczaniem specyficznych funkcji biologicznych dla danego zbioru ludzkich genów.
3. Opublikowano pakiet *GOpro* w repozytorium Bioconductor.
4. Zastosowano wypracowane podejście do analizy danych z The Cancer Genome Atlas.

Rozdział 2

Analiza funkcji biologicznej genów

Gen to odcinek DNA, w którym zapisane są informacje niezbędne do syntezy RNA lub białek. Odczyt fragmentu DNA i użycie rozszyfrowanej informacji do syntezy danej cząstki nazywane jest ekspresją genu. Kontrola ekspresji genu jest niezbędna do tego, aby komórka mogła wytwarzać określone produkty w momencie, gdy są one potrzebne. Dzięki temu komórka może się adaptować do zmieniających się warunków, czynników zewnętrznych i innych stanów. Za pomocą macierzy ekspresyjnych [5] możliwe jest badanie poziomów ekspresji wybranej populacji komórek. Sygnatura genetyczna [6] jest zbiorem genów komórki, którego wzorzec ekspresji jednoznacznie określa fenotyp lub stan chorobowy. Na podstawie fenotypów możliwe jest m.in. rozróżnienie między podtypami danej choroby lub predykcja przeżywalności chorego pacjenta[7].

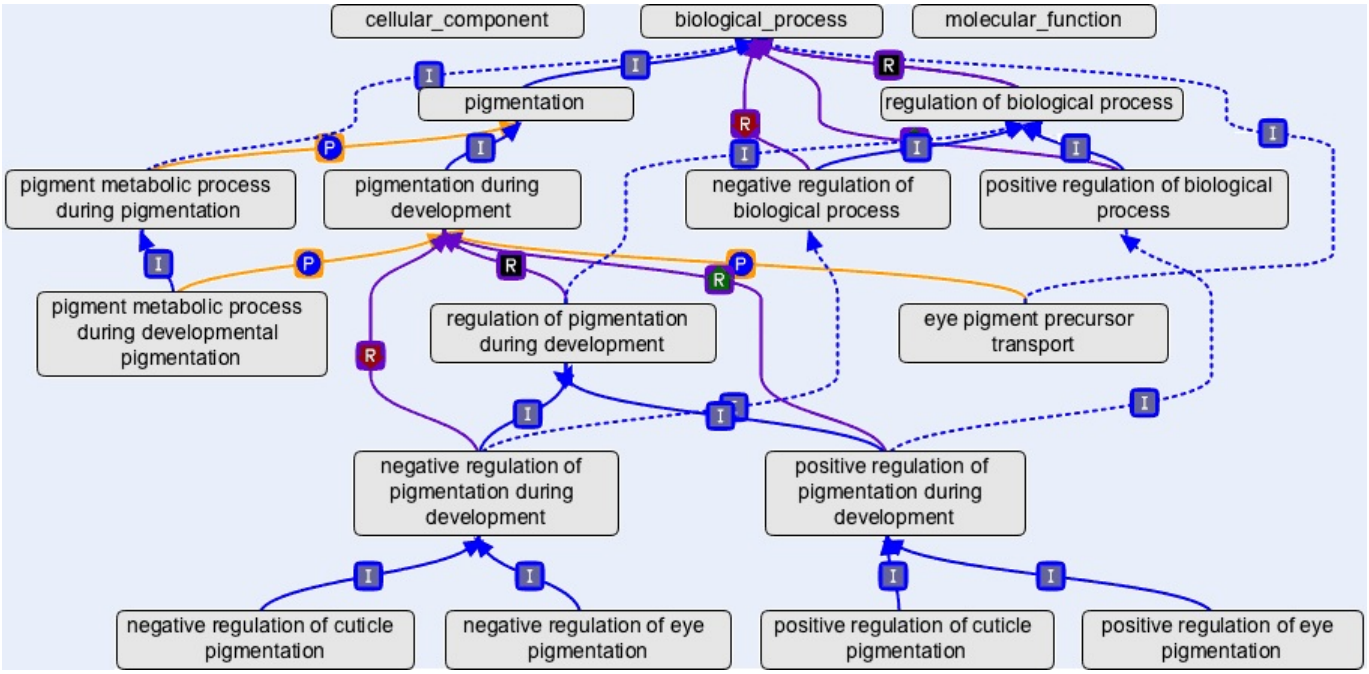
2.1. Ontologia genów

Ontologia jest nauką zajmującą się tworzeniem słownika pojęć i powiązań między tymi pojęciami w danej dziedzinie. Ontologia genów zajmuje się konstruowaniem pojęć dotyczących produktów genów (ale nie produktów *per se*) oraz powiązań między atrybutami produktów na poziomie komórkowym. Na podstawie informacji o ontologii można dokonać analizy funkcjonalnej genów, czyli powiązać produkty danego genu (a tym samym geny) z określonymi obiektami komórkowymi i procesami biologicznymi. Informacje dotyczące ontologii genów zgromadzone są w bazie *Gene Ontology* (GO) [8].

2.1.1. Struktura ontologii genów

Ontologia genów składa się z kategorii GO *ang. GO terms* i relacji między nimi. Kategorie GO można podzielić na trzy subontologie:

- proces biologiczny (biological proces, oznaczany jako BP) - opisuje zestaw zdarzeń realizowanych przez co najmniej jedną grupę funkcji produktu ekspresji genu. Przykładami kategorii dla tej subontologii są, np.: fizjologiczny proces komórkowy, przewodzenie sygnału, przemiany metaboliczne pirymidyn,



Rysunek 2.1: Przykład schematu ontologii genów. Źródło: [10]

- lokalizację komórkową (cellular component, oznaczaną jako CC) - te kategorie przedstawiają składnik komórki, który jest częścią większego elementu. Może to być struktura anatomiczna (np. jądro komórkowe) lub grupa produktów genów (np. rybosom),

- funkcję produktu ekspresji genu (molecular function, oznaczane jako MF) - opisuje ona działania na poziomie komórkowym, np. aktywność katalityczna, aktywność wiążąca lub aktywność cyklazy adenylanowej. Kategorie GO funkcji produktu ekspresji genu reprezentują aktywność a nie element, który tę aktywność wykonuje i nie określają gdzie, kiedy oraz w jakim kontekście dana aktywność się odbywa. Subontologia MF zwykle odnosi się do aktywności, która może być wykonana przez pojedyncze produkty genu jednak niektóre działania są przeprowadzane przez kompleksy produktów genów [8][9].

Powiązania w obrębie każdej z subontologii są przedstawiane w postaci skierowanego grafu acyklicznego. Przykładową strukturę związków między kategoriami GO przedstawiono na rysunku 2.1.1. Każda z przedstawionych subontologii w istocie reprezentuje oddzielną ontologię, ponieważ trzy węzły tego grafu przedstawione przez BP, CC oraz MF są jednocześnie jego korzeniami. Wynika to z tego, że nie występują między nimi powiązania oraz nie posiadają one wspólnego rodzica.

W ramach danej subontologii między węzłami (kategoriami GO) istnieją następujące rodzaje powiązań: „jest", „jest częścią", „reguluje", „ma część". Przedstawione subontologie GO są rozłączne, ponieważ między żadnymi węzłami różnych subontologii nie występuje powiązanie „jest". Możliwe jest występowanie między kateoriami zależności „jest częścią", „reguluje" lub „ma część". Kategorie GO są tworzone w sposób uniwersalny, tak aby można było odnosić je do dowolnego organizmu. Skrócona struktura pojedynczej kategorii GO przedstawia się następująco:

- id: GO:0016049 - identyfikator kategorii GO,

- name: cell growth - nazwa przypisana danej kategorii GO,
- namespace: biological_process - subontologia, w której dana kategoria się zawiera,
- def: "The process in which a cell irreversibly increases in size over time by accretion and biosynthetic production of matter similar to that already present." - opis kategorii
- synonym: "cell expansion" RELATED - synonim do nazwy danej kategorii,
- synonym: "cellular growth" EXACT - synonim do nazwy danej kategorii,
- synonym: "growth of cell" EXACT - synonim do nazwy danej kategorii,
- is_a: GO:0009987 ! cellular process - dana kategoria GO jest częścią kategorii GO:0009987,
- is_a: GO:0040007 ! growth - dana kategoria GO jest częścią kategorii GO:0040007,
- relationship: part_of GO:0008361 ! regulation of cell size - powiązania: dana kategoria jest częścią kategorii GO:0008361, która reguluje rozmiar komórki [10].

Do każdej z kategorii może być przypisane wiele genów, również każdy gen może być przypisany do wielu kategorii. Gen lub jego produkt mogą nie zostać przypisane do żadnej kategorii GO. Czasem w analizie funkcji biologicznych w celu ograniczenia liczby rozważanych kategorii GO wykorzystywane jest pojęcie poziomu. Jako poziom określana jest liczba krawędzi łączących korzeń danej subontologii z daną kategorią GO. Jeśli kategorie GO występują na danym poziomie, to minimalna liczba krawędzi łączących korzeń danej subontologii z każdą z wybranych kategorii (najkrótsza ścieżka między kategorią a korzeniem) jest taka sama.Należy zauważyć, że poziomy między różnymi subontologiami nie są między sobą porównywalne.

2.1.2. Baza danych Gene Ontology

Baza danych GO jest relacyjną bazą danych składającą się z ontologii genów oraz adnotacji genów i produktów genów do kategorii GO. Adnotacja jest procesem przypisywania kategorii GO do produktów genu. Adnotacje GO składają się z kategorii GO powiązanych z odnośnikiem do opisu pracy bądź analizy, na podstawie której powiązano kategorię GO z danym produktem genu. Zebranie danych dotyczących ontologii oraz adnotacji w jednej bazie danych umożliwia uzyskiwanie informacji dotyczących szczegółowych opisów kategorii GO już przypisanych do konkretnego produktu genu za pomocą zapytań do jednej bazy danych.

Liczba kategorii GO w bazie nie jest stała, ponieważ cały czas gromadzone są nowe rekordy a jednocześnie weryfikowane są rekordy już wcześniej w bazie zawarte. Dla przykładu, 27 lutego 2016 roku baza danych GO zawierała 29271 rekordów kategorii z subontologii BP, 10952 rekordów kategorii z subontologii MF oraz 4068 rekordów kategorii z ontologii CC. w porównaniu do 27 października 2015 roku łączna liczba przedstawionych rekordów zwiększyła się o 463 rekordy [11].

2.2. Analiza wzbogacenia genów

Analiza wzbogacenia genów ma na celu przypisanie funkcji biologicznej do pewnej grupy genów [12]. Grupa ta może zostać wyodrębniona przy użyciu analizy ekspresji, wiązania za pomocą tego samego czynnika transkrypcyjnego lub na podstawie wiedzy *a priori*. Identyfikacja wzorca dla danej grupy polega na poszukiwaniu wzbogacenia - ocenie czy dany zbiór genów wykazuje istotną nadprezentację wskazanej funkcji biologicznej.

Do przeprowadzenia analizy wzbogacenia genów niezbędna jest informacja o wszystkich genach, których ekspresje (lub inne cechy na podstawie, których zostały wyodrębniona grupa) zostały oznaczone w danym eksperymencie. Do każdego z genów przypisywane są wszystkie jego kategorie GO. Zakładając, że rozkład cechy ma rozkład hipergeometryczny testowana jest istotność każdej funkcji dla danej grupy genów. Dla każdej grupy i każdej funkcji określa się:

- **n** genów, z których **m** (zbiór **A**) ma adnotację z pewną funkcją,
- **m'** (n-m) genów (zbiór **T**), z których **k** ma tę funkcję,

wówczas $P(|A \bigcap T| = k) = HG(n, m, m', k)$, p-wartość wzbogacenia wynosi $P(|A \bigcap T| \geq k) = \sum_{j \geq k} HG(n, m, m', j)$, gdzie HG(n, m, m', k) oznacza rozkład hipergeometryczny z parametrami n, m, m', k.

Alternatywną metodą analizy wzbogacenia jest GSEA (*Gene Set Enrichment Analysis*)[13], która wykorzystuje zdefiniowany *a priori* zbiór genów, które zostały zgrupowane na podstawie udziału w tych samych szlakach biologicznych lub zbliżonej lokalizacji na chromosomie REF. Baza danych zawierająca zdefiniowane zbiory to *The Molecular Signatures Database (MSigDB)* REF. w analizie GSEA wykorzystywana jest informacja o ekspresjach genów, przy czym pod rozważana jest informacja o całych listach genów zamiast informacji o pojedynczych genach REF.

Rozdział 3

Aparat matematyczny

3.1. Analiza wariancji

Jednoczynnikowa analiza wariancji (ANOVA od ang. *analysis of variance*) wykorzystywana jest do badań porównawczych średnich w wielu grupach, w których zmienna objaśniana ma charakter ciągły [14]. Uwzględnia ona jeden czynnik, który występuje na kilku poziomach.

Za pomocą analizy wariancji testowana jest hipoteza

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \tag{3.1}$$

przeciwko hipotezie:

$$H_A : \text{istnieją } i \neq j \text{ takie, że } \mu_i \neq \mu_j, \tag{3.2}$$

gdzie k≥2 jest liczbą porównywanych grup.

Model analizy wariancji zakłada, że zmienna odpowiedzi (objaśniana) dla każdego poziomu czynnika (zmienna objaśniająca) ma rozkład normalny z tą samą wariancją i jest niezależna od odpowiedzi dla pozostałych poziomów. Rozkłady zmiennej odpowiedzi różnią się między sobą wartością średnią, która, wraz z wariancją błędu, jest parametrem modelu. Model ten opisywany jest równaniem 3.3.

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \tag{3.3}$$

gdzie: Y_ij jest wartością zmiennej odpowiedzi dla j-tej obserwacji i-tego poziomu czynnika (i-tej grupy), μ_i jest nieznaną stałą modelu - średnią w grupie, ϵ_{ij} to błąd losowy dla j-tej obserwacji w i-tej grupie.

Parametry modelu μ_i oraz σ^2 mogą zostać oszacowane za pomocą metody najmniejszych kwadratów lub za pomocą metody największej wiarygodności [17]. Obie te metody prowadzą do uzyskania rozwiązania o postaci:

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \forall i = 1, 2, ..., r, \tag{3.4}$$

gdzie n_i jest liczbą obserwacji w $i-$tej grupie oraz r jest liczbą poziomów czynnika.

Całkowita zmienność obserwacji Y_{ij} jest mierzona w odniesieniu do całkowitej zmienności każdej obserwacji względem średniej ze wszystkich obserwacji. W analizie wariancji ta całkowita zmienność dzielona jest na dwie składowe: zmienność względem estymowanej średniej dla danego poziomu czynnika oraz zmienność estymowanej średniej dla danego czynnika względem średniej ze wszystkich obserwacji. Powyższe można zapisać w formie równania 3.5.

$$Y_{ij} - \overline{Y}_{..} = Y_{ij} - \overline{Y}_{i.} + \overline{Y}_{i.} - \overline{Y}_{..}, \tag{3.5}$$

gdzie $\overline{Y}_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}$, $\overline{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$.

Po podniesieniu obu stron równania do kwadratu i zsumowaniu po wszystkich obserwacjach otrzymuje się podstawowe równanie ANOVA:

$$\underbrace{\sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{..}\right)^2}_{SST} = \underbrace{\sum_{j=1}^{n_i} n_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2}_{SSR} + \underbrace{\sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i.}\right)^2}_{SSE} \tag{3.6}$$

SST (ang. *total sum of square*) oznacza całkowitą zmienność odpowiedzi, *SSR* (ang. *residual sum of square*) oznacza zmienność międzygrupową, *SSE* (ang. *error sum of square*) oznacza zmienność wewnątrzgrupową. Średnie kwadratów zmienności dane są wzorami: $MS_{SSR} = \frac{SSR}{r-1}$ oraz $MS_{SSE} = \frac{SSE}{n_T - r}$, z kolei n_T oznacza liczbę wszystkich obserwacji. Jeżeli hipoteza zerowa jest prawdziwa, wówczas MS_{SSR} i MS_{SSE} powinny różnić się jedynie w granicach losowych odchyleń. Wartości oczekiwane średnich kwadratów odchyleń wynoszą:

$$E(MS_{SSE}) = \sigma^2, \tag{3.7}$$

$$E(MS_{SSR}) = \sigma^2 + \frac{1}{r-1} \sum_{i=1}^{r} n_i (\mu_i - \mu_.)^2, \tag{3.8}$$

w przypadku prawdziwości hipotezy zerowej $E(MS_{SSR}) = \sigma^2$, co znaczy że $E(MS_{SSR})$ i $E(MS_{SSE})$ są sobie równe. Jeśli stosunek MS_{SSR} do MS_{SSE} jest znacząco większy od 1, to świadczy to przeciwko hipotezie zerowej. Statystykę testową można zapisać jako:

$$F^* = \frac{MS_{SSR}}{MS_{SSE}} \overset{H_0}{\sim} F_{r-1,n_T-r}, \tag{3.9}$$

zapis F_{r-1,n_T-r} oznacza rozkład F-Snedecora o $r-1$, n_T-r stopniach swobody. W przypadku, gdy $F^* \leq F_{1-\alpha;r-1,n_T-r}$ nie ma podstaw do odrzucenia hipotezy zerowej. W przeciwnym przypadku przyjmuje się hipotezę alternatywną. $F_{1-\alpha;r-1,n_T-r}$ oznacza kwantyl rzędu $1-\alpha$ rozkładu F-Snedecora o $r-1$, n_T-r stopniach swobody, gdzie α jest zadanym poziomem istotności.

3.2. Testy post-hoc

W przypadku odrzucenia hipotezy zerowej w teście ANOVA przyjmuje się, że co najmniej dwie średnie grupowe są od siebie różne. W celu zweryfikowania, które średnie istotnie różnią się między sobą przeprowadzane są porównania pomiędzy wybranymi parami średnich. Niech n_i oznacza liczbę obserwacji w $i-$tej grupie. Zakłada się, że badane średnie są niezależne

o nieznanych, równych wariancjach. Dla każdej pary średnich badana jest hipoteza zerowa postaci:

$$H_0 : \mu_i = \mu_j, \tag{3.10}$$

przeciwko hipotezie alternatywnej:

$$H_A : \mu_i \neq \mu_j, \tag{3.11}$$

gdzie $i \neq j$. Estymator różnicy średnich jest liczony jako

$$\hat{D} = \hat{\mu_i} - \hat{\mu_j} = \overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}. \tag{3.12}$$

Estymator ten ma rozkład normalny z parametrami:

$$\begin{cases} E(\hat{D}) = E(\overline{Y}_{i\cdot}) - E(\overline{Y}_{i\cdot}) = \mu_i - \mu_j, \\ Var(\hat{D}) = Var(\overline{Y}_{i\cdot}) + Var(\overline{Y}_{j\cdot}) = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_j}. \end{cases} \tag{3.13}$$

Estymator wariancji $\hat{D}$ wynosi $s^2(\hat{D}) = \frac{MS_{SSE}}{n_i} + \frac{MS_{SSE}}{n_j}$. Zatem statystyka testowa jest postaci

$$t^* = \frac{\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}}{\sqrt{\frac{MS_{SSE}}{n_i} + \frac{MS_{SSE}}{n_j}}} \overset{H_0}{\sim} t_{n_T - r}. \tag{3.14}$$

Z powyższego można wywnioskować, że przedział ufności dla statystyki testowej wynosi

$$\left(-t_{1-\frac{\alpha}{2}, n_T - r}, t_{1-\frac{\alpha}{2}, n_T - r}\right),$$

gdzie $t_{1-\frac{\alpha}{2}, n_T - r}$ jest odpowiednim kwantylem rozkładu t-Studenta oraz n_T oznacza liczbę obserwacji. Hipoteza zerowa jest odrzucana, gdy wartość statystyki testowej t^* nie należy do tego przedziału.

W przypadku porównywania r par średnich liczba wykonywanych testów wynosi $\binom{r}{2}$, co oznacza, że należy zmodyfikować poziom istotności pojedynczego testu α_i, tak aby sumaryczny poziom istotności wszystkich porównań był równy α.

Istnieje wiele sposobów pozwalających na dobór poziomu istotności α_i dla pojedynczego testu. Najczęściej przedstawianymi są:

- procedura Bonferroniego [15] - poziom istotności dla pojedynczego testu wynosi $\alpha_i = \alpha/r$, gdzie r jest liczbą porównywanych par. Dla dużej liczby porównań r wielkość α_i jest tak mała, że w większości przypadków procedura nie odrzuca hipotezy zerowej.
- procedura Sheffego [16] - jest wykorzystywana do badania kontrastów - porównań między średnimi, dla których H_0 jest postaci: $\sum_{i=1}^{r} c_i \mu_i = c$, gdzie c_i oraz c to pewne stałe.
- procedura Tukey'a - stosowana do przeprowadzania porównań wielokrotnych, gdzie $H_0 : \mu_i - \mu_j = 0$. Metoda ta zostanie szerzej opisana ze względu na wykorzystanie jej w dalszej części pracy.

Metoda Tukey'a [17] (nazywana również testem Tukey'a) jest modyfikacją testu opisanego statystyką 3.14. Dla wszystkich poziomów czynnika wykonywane są porównania między średnimi - testowane jest $\binom{r}{2}$ hipotez zerowych postaci:

$$H_{0_k} : \mu_i = \mu_j, \tag{3.15}$$

przeciwko:

$$H_{A_k} : \mu_i \neq \mu_j. \tag{3.16}$$

W przypadku równej liczby obserwacji dla każdego poziomu czynnika całkowity poziom istotności dla wszystkich przeprowadzonych testów wynosi dokładnie α. W przeciwnym przypadku poziom istotności jest mniejszy od zadanego α. Oznacza to, że metoda Tukey'a jest konserwatywna - rzadziej odrzucana jest hipoteza zerowa.

Statystyka testowa dla każdego z porównań opisana jest przez:

$$q = \frac{\sqrt{2}\left(\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}\right)}{\sqrt{\frac{MS_{SSE}}{n_i} + \frac{MS_{SSE}}{n_j}}} \overset{H_0}{\sim} q_{r,n_T-r}. \tag{3.17}$$

jeśli $|q| \leq q(1-\alpha, r, n_T - r)$, to nie ma podstaw do odrzucenia H_0, w przeciwnym przypadku przyjmowana jest hipoteza alternatywna. Przy założeniu hipotezy zerowej statystyka q ma podany rozkład, jeżeli n_i jest równe n_j.

Procedura ta wykorzystuje rozkład studentyzowanego rozstępu. Dla r niezależnych obserwacji $Z_1, ..., Z_r$ z rozkładu normalnego o średniej μ i wariancji σ^2, niech w będzie rozstępem dla tego zbioru obserwacji:

$$w = max(Z_i) - min(Z_i). \tag{3.18}$$

Zakładając, że estymatorem wariancji σ^2 jest s^2 o ν stopniach swobody oraz s^2 jest niezależne od Z_i. Wówczas, stosunek w/s nazywany jest studentyzowanym rozstępem i oznaczany jest jako:

$$q(r, \nu) = \frac{w}{s}. \tag{3.19}$$

W przypadku nierównej liczby obserwacji w grupach procedura Tukey'a nazywana jest czasem procedurą Tukey'a - Kramera i dla niedużych różnic między liczebnościami grup może być stosowana.

3.3. Klasteryzacja hierarchiczna

Klasteryzacja (analiza skupień, grupowanie) jest techniką zajmującą się podziałem elementów zbioru na klastry (grupy) charakteryzujące się pewnym podobieństwem elementów wewnątrz danego klastra i wystarczającym brakiem podobieństwa elementów jednego klastra względem elementów drugiego klastra [18][19]. Obiekty przydzielane są do klastrów ze względu na posiadane przez nie cechy (inaczej wartości zmiennych opisujących dany element). Elementy przenoszone są do przestrzeni cech o wymiarze odpowiadającym liczbie cech opisujących dany zbiór elementów. Przy założeniu, że zależności między obiektami są reprezentowane przez opisujące je cechy, znalezione w przestrzeni cech klastry mogą być interpretowane jako grupy elementów na podstawie dodatkowych informacji o elementach.

Metoda klasteryzacji hierarchicznej polega na dokonaniu szeregu podziałów elementów na klastry rozpoczynając od klastra zawierającego wszystkie elementy sukcesywnie dzieląc go na klastry o mniejszej liczbie elementów (metoda deglomeracyjna) lub rozpoczynając od klastrów jednoelementowych i łącząc je do uzyskania jednego klastra zawierającego wszystkie elementy zbioru (metoda aglomeracyjna). Elementy łączone (dzielone) są na podstawie wartości miary niepodobieństwa, która określona jest dla zmiennych ciągłych. Miara jest to funkcja d określona na zbiorze A, której argumenty należą do przedziału $[0, \infty)$ i która dla elementów

a, $b \in A$ spełnia warunki:

$$\begin{aligned} d(a,b) &= d(b,a), \\ d(a,b) > 0 &\Leftrightarrow a \neq b, \\ d(a,b) = 0 &\Leftrightarrow a = b. \end{aligned}$$

Miara jest miarą odległości (metryką), jeżeli spełnia nierówność trójkąta, tzn. dla $a, b, c \,\epsilon\, A$ $d(a,c) \leq d(a,b) + d(b,c)$. Informacje o odległościach między wszystkimi parami elementów zbioru A opisane są za pomocą macierzy niepodobieństwa $\mathbf{D}$, która jest symetryczna.

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & d_{13} & \dots & d_{1N} \\ d_{21} & 0 & d_{23} & \dots & d_{2N} \\ d_{31} & d_{32} & 0 & \dots & d_{3N} \\ \vdots & \vdots & \vdots & & \vdots \\ d_{N1} & d_{N2} & d_{N3} & \dots & 0 \end{bmatrix}$$

Miary niepodobieństwa dzieli się na metryki oraz miary odnoszące się do współczynnika korelacji.

Do najpopularniejszych miar niepodobieństwa zaliczają się:

- odległość euklidesowa: $d_{ij} = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right]^{1/2}$, (3.21)

- odległość miejska (manhattan): $d_{ij} = \sum_{k=1}^{p} w_k |x_{ik} - x_{jk}|$, (3.22)

- odległość Minkowskiego: $d_{ij} = \left(\sum_{k=1}^{p} {w_k}^r |x_{ik} - x_{jk}|\right)^{1/r} \quad (r \geq 1)$, (3.23)

- odległość Canberra: $d_{ij} = \begin{cases} 0 \text{ dla } x_{ik} = x_{jk} = 0 \\ \sum_{k=1}^{p} w_k |x_{ik} - x_{jk}| / \left(|x_{ik}| + |x_{jk}|\right) \text{ dla } x_{ik} \neq 0 \text{ lub } x_{jk} \neq 0, \end{cases}$ (3.24)

- korelacja Pearsona: $\delta_{ij} = (1 - \phi_{ij}/2)$, gdzie (3.25)

$$\phi_{ij} = \sum_{k=1}^{p} w_k (x_{ik} - \overline{x}_{i\cdot})(x_{jk} - \overline{x}_{j\cdot}) \Bigg/ \left[\sum_{k=1}^{p} w_k (x_{ik} - \overline{x}_{i\cdot})^2 \sum_{k=1}^{p} w_k (x_{jk} - \overline{x}_{j\cdot})^2\right]^{1/2},$$

gdzie x_{ik}, x_{jk} są wartościami k-tej zmiennej z p zmiennych dla elementów i oraz j.

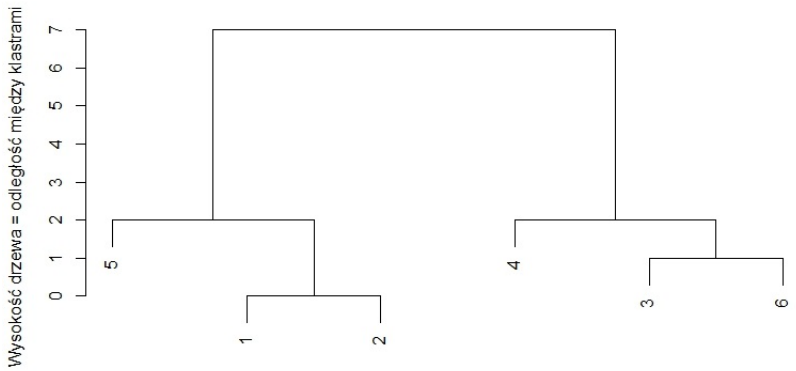
Istnieją różne sposoby złączania klastrów w klasteryzacji hierarchicznej. Należą do nich m.in. metody: pojedynczego wiązania, pełnego wiązania, średniego wiązania, centroidów, Ward'a [18]. W metodzie pojedynczego wiązania odległość między klastrami zdefiniowana jest jako najmniejsza odległość między parami obiektów (jeden obiekt z jednego klastra, drugi obiekt z drugiego klastra). Metoda pełnego wiązania określa tę odległość jako największą odległość między elementami dwóch klastrów. W metodzie centroidów brany jest pod uwagę kwadrat odległości Euklidesa między centroidami (wektorami wartości średnich elementów klastra).

Metoda średniego wiązania bierze pod uwagę średnią odległość między parami elementów klastra, co ilustruje równanie:

$$d(X,Y) = \frac{1}{|X||Y|} \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} d(x_i, y_j), \tag{3.26}$$

gdzie X, Y oznaczają klastry, $|X|$, $|Y|$ oznaczają liczbę elementów klastrów X i Y, a x_i, y_j ich elementy.

Graficznym przedstawieniem klasteryzacji hierarchicznej jest dendrogram. Przykładowy dendrogram przedstawiono na rysunku 3.3. Jest to drzewo, którego węzły i wysokości gałęzi reprezentują powiązania między elementami. Wraz ze wzrostem wysokości drzewa rośnie wartość miary niepodobieństwa między obiektami.



Rysunek 3.1: Przykładowy dendrogram dla 6-elementowego zbioru, jako miarę niepodobieństwa przyjęto odległość euklidesową, do złączania klastrów zastosowano metodę całkowitego wiązania.

Reasumując, zastosowanie aglomeracyjnych metod klasteryzacji hierarchicznej wymaga w pierwszym kroku stworzenia jednoelementowych klastrów o liczebności równej mocy zbioru. Następnie obliczana jest macierz niepodobieństwa za pomocą wybranej miary i stosowana jest wybrana metoda złączania klastrów. Po złączeniu dwóch klastrów macierz niepodobieństwa jest aktualizowana z uwzględnieniem nowo powstałego klastra i powtórnie klastry o najmniejszej wartości miary niepodobieństwa są złączane. Ta procedura powtarzana jest do momentu, gdy pozostaje jeden klaster zawierający wszystkie elementy zbioru.

3.4. Test Fishera

Test Fishera [20] ma zastosowanie do badania niezależności cech. Najczęściej jest on stosowany do badania niezależności dwóch zmiennych jakościowych. W przypadku dwóch zmiennych, z których każda może przyjmować po dwie wartości próbę można zaprezentować w tablicy kontyngencji jak przedstawiono w tabeli 3.1.

Tabela 3.1: Przykład tablicy kontyngencji dla dwóch zmiennych A oraz B, które mogą przyjmować wartości odpowiednio a_1 lub a_2 oraz b_1 lub b_2. Przez n_{ij} oznaczono liczbę obserwacji dla odpowiednich kombinacji wartości cech A i B.

	Cecha A		
Cecha B	a_1	a_2	$\sum$
b_1	n_{11}	n_{12}	n_{1+}
b_2	n_{21}	n_{22}	n_{2+}
$\sum$	n_{+1}	n_{+2}	n_{++}

Iloraz szans jest stosunkiem prawdopodobieństwa sukcesu do prawdopodobieństwa porażki dla dwóch grup. Estymator ilorazu szans θ dla przedstawionej tablicy ma postać:

$$\hat{\theta} = \frac{\frac{n_{11}}{n_{1+}} \Big/ \left(1 - \frac{n_{11}}{n_{1+}}\right)}{\frac{n_{21}}{n_{2+}} \Big/ \left(1 - \frac{n_{21}}{n_{2+}}\right)} = (n_{11}n_{22})/(n_{12}n_{21}). \tag{3.27}$$

W jednostronnym teście Fishera badana jest hipoteza zerowa (brak zależności między cechami):

$$H_0 : \theta = 1, \tag{3.28}$$

przeciwko hipotezie alternatywnej:

$$H_A : \theta < 1 \tag{3.29a}$$

lub

$$H_A : \theta > 1 \tag{3.29b}$$

W teście Fishera wyznaczane jest prawdopodobieństwo zaobserwowania w tabeli kontyngencji wartości bardziej skrajnych niż obserwowane przy założeniu stałych sum elementów wierszy oraz kolumn (n_{1+}, n_{2+}, n_{+1}, n_{+2}). Rozkład warunkowy zmiennej losowej n_{11} pod warunkiem sumy elementów pierwszego wiersza i pierwszej kolumny jest hipergeometryczny. Jeżeli hipoteza zerowa jest prawdziwa, to prawdopodobieństwo zaobserwowania konkretnej wartości n_{11} dane jest wzorem:

$$P(n_{11} = k) = \frac{\binom{n_{1+}}{k}\binom{n_{2+}}{n_{+1}-k}}{\binom{n}{n_{+1}}} \tag{3.30}$$

P-wartość testu jest odczytywana z dystrybuanty rozkładu hipergeometrycznego z odpowiednimi parametrami.

3.5. Problem wielokrotnego testowania

Wielokrotne testowanie ma miejsce w przypadku, gdy dla danej próby testowane jest jednocześnie wiele hipotez. Jeśli wykonywany jest pojedynczy test, wtedy prawdopodobieństwo popełnienia błędu pierwszego rodzaju jest równe α. W sytuacji badania rodziny hipotez (przeprowadzania n testów) prawdopodobieństwo popełnienia co najmniej jednego błędu pierwszego rodzaju (FWER -*Family-Wise Error Rate*) wynosi: $1 - (1 - \alpha)^n$ [21].

Do metod kontroli miary FWER należą: procedura Bonferroniego, procedura Tukey'a stosowana tylko w przypadku porównań par oraz procedura Sheffego przedstawione w podrozdziale

3.2, procedura Sidaka [22], procedura Holma [23], procedura Hochberga [24], test Dunnetta [25].

Inną miarą wykorzystywaną do korekcji poziomu istotności w przypadku wielokrotnego testowania jest miara FDR (False Discovery Rate). FDR kontroluje wartość oczekiwaną proporcji fałszywie odrzuconych hipotez zerowych, co można zapisać jako:

$$FDR = E\left[\frac{V}{R}\right], \tag{3.31}$$

gdzie V jest liczbą odrzuconych prawdziwych hipotez zerowych, zaś R jest liczbą odrzuconych hipotez zerowych. Jeśli $R = 0$, wówczas $FDR = 0$.

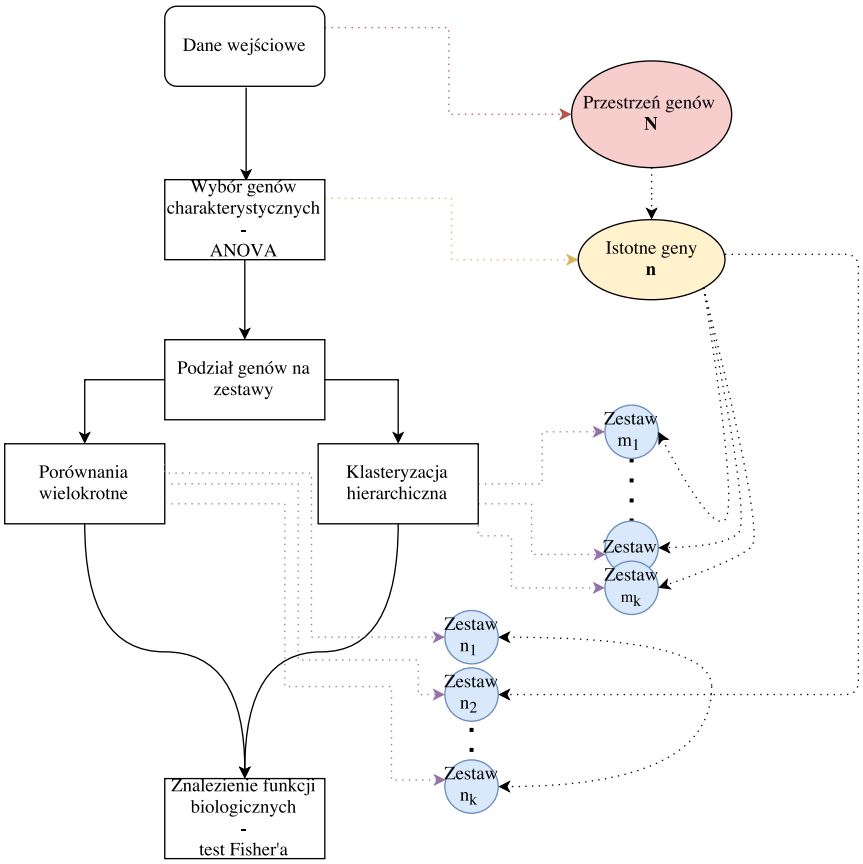
Metodami kontroli FDR są:

- procedura Benjaminiego - Hochberga (BH) [26] stosowana przy braku zależności testów. Otrzymane dla rodziny m testów p-wartości są sortowane rosnąco, następnie dla danego α znajdowana jest największe k takie, że $P(k) \leq \frac{k}{m}\alpha$. Odrzucana jest hipoteza zerowa dla każdego H_i, gdzie $i = 1, ..., k$.
- procedura Benjaminiego - Hochberga - Yakutieliego [27], znajdowane jest największe k takie, że $P(k) \leq \frac{k}{m \cdot c(m)}\alpha$. Dla testów niezależnych lub dodatnio skorelowanych $c(m) = 1$ oraz $c(m) = \sum_{i=1}^{m} \frac{1}{i}$ dla testów zależnych.

Rozdział 4

Algorytm wyznaczania charakterystycznych funkcji biologicznych

Zadaniem algorytmu jest wyznaczenie zestawów interesujących genów oraz znalezienie dla nich charakterystycznych funkcji biologicznych. Zbiór wszystkich interesujących genów wyznaczany jest na podstawie analizy wariancji - znajdowane są geny, których ekspresje istotnie różnią się między wskazanymi grupami. Następnie wybrane geny dzielone są na zestawy z użyciem procedury Tukey'a lub klasteryzacji hierarchicznej. Dla każdego z określonych zestawów genów wyznaczane są funkcje charakterystyczne. Całkowity proces przedstawiono na rysunku 4.1 i dokładnie opisano w punktach poniżej.



Rysunek 4.1: Schemat algorytmu z zaznaczonym na kolorowo przepływem danych: kolorem czerwonym oznaczono dane wejściowe, żółtym - geny wybrane jako charakterystyczne za pomocą analizy wariancji, niebieskim - zestawy genów wyznaczone przy użyciu porównań wielokrotnych lub klasteryzacji hierarchicznej (opracowanie własne).

4.1. Struktura danych wejściowych

Dane muszą odnosić się do genomu ludzkiego, nazwami atrybutów winny być skrótowe nazwy genów (tzw. alias). Wartości atrybutów powinny należeć do zbioru liczb rzeczywistych. Jeśli wśród tych wartości występują informacje niekompletne, wówczas należy odnotować je jako *NA*. Atrybuty powinny znajdować się w wierszach, a obserwacje w kolumnach. Wymagane są co najmniej dwa zestawy informacji identycznych ze względu na atrybuty oraz kolejność ich występowania w strukturze danych a różniących się na podstawie określonej cechy. Przykładowo, mogą to być cztery grupy pacjentów, gdzie czynnikiem je odróżniającym jest stan zaawansowania choroby lub jej typ. Dane powinny mieć rozkład normalny o jednakowej wariancji lub posiadać ten sam rozkład o jednakowej wariancji. Liczba kompletnych obserwacji determinuje moc testów statystycznych i poprawność uzyskanych wyników jednak wskazanie dokładnej jej wartości *a priori* nie jest możliwe ze względu na jej ścisłą zależność z liczbą grup. w tabeli 4.1 zaprezentowano przykład poprawnej struktury danych wejściowych. Dla przedstawionych informacji można wyróżnić 2 cechy (typ), a dla każdej z nich po 7 identycznych atrybutów wymienionych w tej samej kolejności oraz po 3 niezależne próbki. Wartości pokazane w tabeli to znormalizowane ekspresje genów.

Typ: BRCA							
	CHST7	AGFG2	AMD1	ABL1	PLAA	ATP1A2	MTUS2
TCGA.B6.A0WY.01	0.23	-0.89	-0.74	0.32	0.32	0.11	na
TCGA.B6.A0WZ.01	-0.53	0.81	-1.51	-0.04	0.53	-1.89	2.78
TCGA.B6.A0X0.01	-1.59	-1.21	-0.36	0.29	0.26	-1.83	-1.19
Typ: OVARIAN							
	CHST7	AGFG2	AMD1	ABL1	PLAA	ATP1A2	MTUS2
TCGA.24.1560.01	0.04	0.92	0.70	-0.34	0.17	-1.63	1.66
TCGA.24.1562.01	-1.55	2.31	-0.43	-0.07	0.10	0.59	-0.71
TCGA.24.1563.01	-1.19	0.12	0.87	-0.05	-0.02	na	-0.45

Tabela 4.1: Przykład poprawnej struktury danych na podstawie wybranych informacji o znormalizowanej ekspresji genów z TCGA. Nazwami kolumn są nazwy genów, nazwami wierszy są nazwy próbek. Typ BRCA oznacza raka piersi, typ OVARIAN oznacza raka jajnika. Nazwa próbki opisywana jest przez kod TCGA. Sposób odczytywania takiego kodu dla próbki TCGA.B6.A0WY.01 jest następujący: TCGA oznacza nazwę projektu, B6 rodzaj próbki i ośrodek, w którym została pobrana (inwazyjny rak piersi, Uniwersytet Duke'a), A0WY to indywidualny kod pacjenta, 01 oznacza typ próbki (pierwotny guz lity)[29, 30].

4.2. Wybór genów charakterystycznych

Wyszczególnienie genów, które są charakterystyczne polega na przeprowadzeniu analizy wariancji, opisanej w rozdziale 3.1, dla zdeterminowanych przez daną cechę grup obserwacji. Test ten przeprowadzany jest dla każdego genu oddzielnie. Ze względu na problem wielokrotnego testowania wprowadzona jest kontrola miary FDR procedurą Benjaminiego - Hochberga. Jeżeli hipoteza zerowa o równości średnich między grupami zostanie odrzucona, wówczas gen jest klasyfikowany jako charakterystyczny. Poziom istotności testu (z uwzględnieniem korekcji na liczbę testów) oraz maksymalna liczba znalezionych genów są parametrami algorytmu. Geny są sortowane rosnąco ze względu na p-wartość testu. Lista genów charakterystycznych może zawierać więcej elementów niż wskazana maksymalna liczba znalezionych genów. Ma to miejsce w przypadku, gdy co najmniej jeden gen ma przypisaną taką samą p-wartość testu jak gen o numerze porządkowym równym wskazanej liczbie. Przykładowy wynik zastosowania testu ANOVA przedstawiono w tabeli 4.2. Poniżej podwójnej poziomej linii wypisane są geny nieistotne w przypadku przyjęcia za poziom istotności wartości 0,1. Dla parametru maksymalnej liczby znalezionych genów równego 5 gen *SEC61B* zostanie wskazany jako istotny, ponieważ p - wartość testu dla tego genu jest równa p - wartości genu wymienionego jako 5 z kolei.

4.3. Podział genów na podzbiory

Oznaczone jako interesujące geny dzielone są na zestawy. Możliwe są dwa sposoby podziału: za pomocą wielokrotnych porównań metodą Tukey'a lub przy użyciu klasteryzacji hierarchicznej.

4.3.1. Porównania wielokrotne metodą Tukey'a

Za pomocą porównań wielokrotnych, opisanych w rozdziale 3.2, każdy gen przyporządkowywany jest do dokładnie jednego zbioru. Maksymalna liczba zestawów genów wyznaczonych

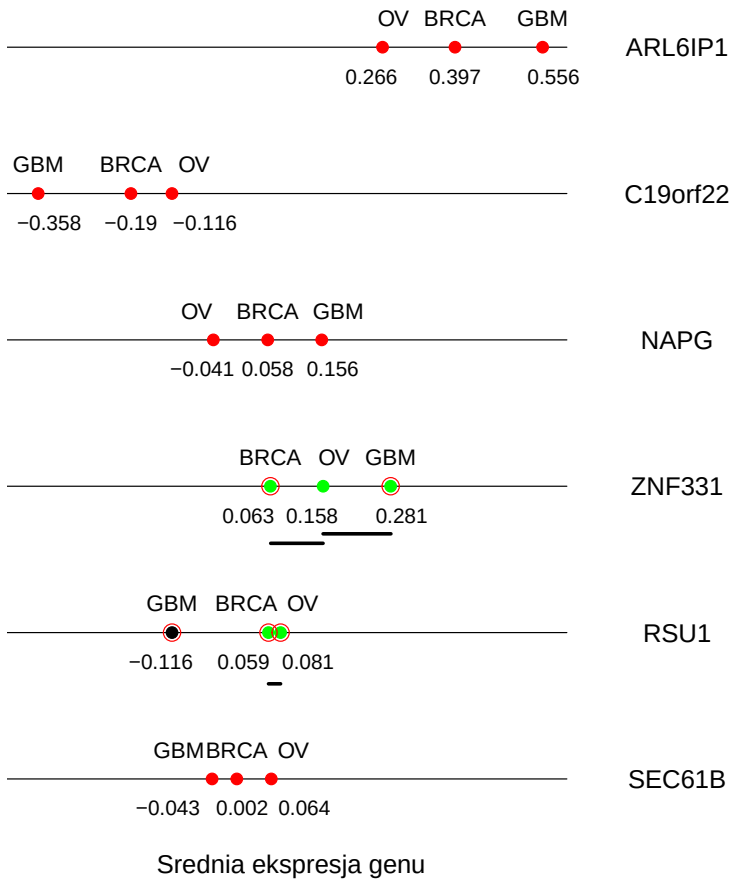
	alias genu	p-wartość	p-wartość (FDR)
1	ARL6IP1	0.00000	0.00001
2	C19orf22	0.00001	0.00004
3	NAPG	0.00017	0.00056
4	ZNF331	0.00043	0.00107
5	RSU1	0.00075	0.00151
6	SEC61B	0.00075	0.00151
7	NCRNA00120	0.17420	0.24885
8	PFDN6	0.33213	0.41517
9	TMEM43	0.47488	0.52765
10	CCNI2	0.74718	0.74718

Tabela 4.2: Przykładowe wyniki analizy wariancji 10 genów wśród 3 różnych grup dla danych z TCGA. Dla poziomu istotności 0,1 geny od 1 do 6 uważane są za istotnie różne między sobą ze względu na wartość ekspresji, co znaczy że wybierane są jako charakterystyczne. W środkowej kolumnie przedstawione są p-wartości testów wyznaczone bez uwzględniania liczby przeprowadzanych testów. W prawej kolumnie p-wartości są skorygowane ze względu na liczbę testów.

dla metody Tukey'a równa jest $n!!$, gdzie zapis n oznacza liczbę grup. Silnia podwójna !! definiowana jest jako:

$$n!! = \begin{cases} 1 & \text{dla } n = 0 \text{ lub } n = 1 \\ n \cdot (n-2)!! & \text{dla } n \geq 2 \end{cases} \tag{4.1}$$

Postępowanie jest analogiczne do wykorzystywanego przy wyborze genów charakterystycznych, z tą różnicą, że porównywane są średnie między każdymi dwoma grupami. Liczba tych zestawień uwzględniana jest przez zastosowanie procedury Tukey'a przy wyznaczaniu przedziałów ufności dla statystyki testowej. Geny o odpowiadających zależnościach między wartościami średnimi należą do jednego zestawu. Przykładowo, geny, których średnie ekspresje w grupie i oraz II są sobie równe i są mniejsze od średniej ekspresji w grupie III zaklasyfikowane zostaną do tego samego zestawu. na rysunku 4.2 schematycznie przedstawiono wynik porównań wielokrotnych. Istotna jest kolejność grup przedstawionych na osi oraz kolor punktu, którym zostały oznaczone. Kolorem czerwonym wyróżniono grupy, dla których hipoteza zerowa została odrzucona. na przykład, dla genu *ARL6IP1* wszystkie trzy porównania średnich między grupami wskazywały na istotne różnice między nimi. Średnia ekspresja genów w grupie *OV* była różna od średniej w grupie *GBM*, ponadto była różna od średniej w grupie *BRCA* oraz *GBM* różniła się względem *BRCA*. na podstawie schematu można stwierdzić, że geny *ARL6IP1* i *NAPG* utworzą jeden zestaw, geny *C19orf22* i *SEC61B* kolejny, a geny *ZNF331* i *RSU1* dwa jednoelementowe zestawy. w przypadku genu *RSU1* średnia w grupie *GBM* jest istotnie różna od średniej w grupie *OV* i *BRCA*, a średnie dla grup *OV* i *BRCA* nie są istotnie różne i z tego powodu zostały oznaczone kolorem zielonym.

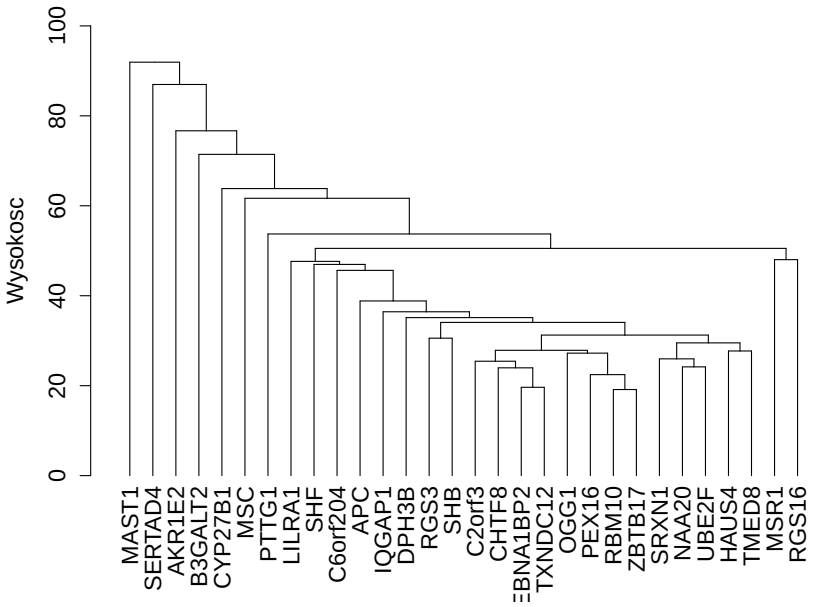


Rysunek 4.2: Wizualizacja idei oraz wyników porównań wielokrotnych. Po prawej stronie rysunku przedstawiono nazwy genów, na osiach - nazwy grup (BRCA, OV, GBM). Kolor czerwony oznacza odrzucenie hipotezy zerowej, zielony - brak podstaw do jej odrzucenia. Dodatkowo, pary grup, przy porównaniu których hipoteza zerowa nie została odrzucona, zostały podkreślone pogrubioną linią (opracowanie własne).

4.3.2. Klasteryzacja hierarchiczna

W przypadku klasteryzacji hierarchicznej możliwy jest dobór dowolnych miar niepodobieństwa oraz metod złączania, co zostało opisane w rozdziale 3.3. Sposób tworzenia klastrów polega na złączaniu zbiorów o najmniejszej wartości niepodobieństwa, rozpoczynając od jednoelementowych zbiorów każdego genu. W wyniku działania algorytmu każdy gen zostaje przyporządkowany do co najmniej jednego zestawu, przy czym nie może należeć do więcej niż jednego zestawu na tym samym poziomie dendrogramu. Maksymalna liczba zestawów wynosi $k + (k - 1)$, gdzie k jest liczbą genów. Na rysunku 4.3 przedstawiono wynik działania klasteryzacji hierarchicznej na 30 genach wybranych jako charakterystyczne. Uwzględnianą

cechą była ekspresja każdego z genów. Jako miarę niepodobieństwa zastosowano odległość euklidesową, a wykorzystaną metodą złączania była metoda średniego wiązania.



Rysunek 4.3: Dendrogram wyznaczony dla 30 genów z TCGA za pomocą grupowania hierarchicznego przy użyciu odległości euklidesowej i metody średniego wiązania[18]. Odległość euklidesową pomiędzy dwoma genami wyznaczono według wzoru: $\sqrt{(A_1 - B_1)^2 + ... + (A_i - B_i)^2}$, gdzie A_i, B_i oznaczają ekspresję genów odpowiednio a i B, i jest numerem porządkowym próbki.

4.4. Wyznaczanie funkcji charakterystycznych

W celu znalezienia funkcji charakterystycznych dla każdego zestawu znajdowane są w bazie danych *org.Hs.eg.db* [31] wszystkie funkcje danego genu z przestrzeni genów dla wybranej domeny (MF, CC, BP przedstawione w rozdziale 2.1). Przestrzeń genów określona jest jako zbiór wszystkich genów uwzględnionych w analizie, na rysunku 4.1 została oznaczona kolorem czerwonym. Przy wyborze wszystkich funkcji możliwe jest pominięcie tych, które są bardzo rzadkie lub bardzo częste w danym zbiorze analizowanych genów. Służy do tego parametr określający minimalną i maksymalną liczbę genów, którym przypisana jest wskazana funkcja. Następnie dla każdej znalezionej funkcji tworzona jest tablica kontyngencji. Składa się ona z czterech elementów:

- liczby genów, które posiadają daną funkcję i należą do wskazanego zestawu - n_{11},
- liczby genów, które posiadają daną funkcję i nie należą do wskazanego zestawu - n_{12},
- liczby genów, które nie posiadają danej funkcji i należą do wskazanego zestawu- n_{21},
- liczby genów, które nie posiadają danej funkcji i nie należą do wskazanego zestawu- n_{22},

przy czym uwzględniane są jedynie geny z danej przestrzeni genów. Przykład tablicy kontyngencji dla funkcji biologicznej oznaczonej identyfikatorem *GO:0045116* przedstawiono w tabeli 4.3.

ID funkcji: GO:0045116 - neddylacja białek [32]		
	w zestawie	Poza zestawem
Ma funkcję	$n_{11} = 2$	$n_{12} = 15$
Nie ma funkcji	$n_{21} = 3$	$n_{22} = 16095$

Tabela 4.3: Przykładowa tablica kontyngencji dla zestawu genów: SRXN1, NAA20, UBE2F, HAUS4, TMED8. Funkcja oznaczona przez identyfikator GO:0045116 to *kowalencyjne wiązanie białka NEDD8 do innego białka* [32].

Tworzone jest n takich tablic dla wszystkich zestawów genów, gdzie n oznacza liczbę funkcji reprezentowanych przez dany zestaw. Dla każdej z tabel przeprowadzany jest test Fishera, opisany w rozdziale 3.4. Parametrem tego testu jest współczynnik istotności. Dla znalezionych funkcji przypisanych do danego zestawu genów przeprowadzana jest kontrola FDR z uwzględnieniem liczby wykonywanych testów dla tego zestawu, czyli liczby znalezionych funkcji (n). Funkcja wybierana jako charakterystyczna cechuje się najniższą p-wartością testu Fishera, mniejszą niż zadany próg istotności. Opisy kategorii GO znajdowane są w bazie danych *GO.db* [34]. Sposób znajdowania funkcji charakterystycznych jest taki sam dla zestawów uzyskanych za pomocą algorytmu klasteryzacji hierarchicznej oraz testu Tukey'a.

Rozdział 5

Przykład zastosowania algorytmu do danych z TCGA

Działanie algorytmu zostało przetestowane na danych pochodzących z The Cancer Genome Atlas (TCGA), które pobrano ze strony internetowej [4] Narodowego Instytutu Zdrowia Stanów Zjednoczonych za pomocą pakietu *RTCGA.PANCAN12* [28].

5.1. Cel analizy danych z TCGA

Analiza danych dotyczących ekspresji genów z The Cancer Genome Atlas ma na celu znalezienie genów różnicujących pacjentów o różnym typie nowotworu i przedstawienie grup tych genów za pomocą charakterystycznych dla nich funkcji biologicznych. Jest to zamiana informacji o nazwach genów, które różnicują pacjentów na informację o funkcjach biologicznych charakterystycznych dla grup genów wyznaczonych na podstawie pewnego podobieństwa.

5.2. Analiza ekspresji genów z wykorzystaniem algorytmu

Z TCGA wybrano informacje dotyczące ekspresji genów dla trzech grup pacjentów. Grupy te rozróżnialne są ze względu na rodzaj nowotworu pacjentów, od których zostały pobrane próbki do określenia ekspresji genów. W tej analizie wzięto pod uwagę następujące typy raka:

- rak piersi - grupa oznaczona jako **BRCA** (ang. *breast cancer*),
- rak jajnika - grupa oznaczona jako **OV** (ang. *ovarian*),
- glejak wielkopostaciowy - grupa oznaczona jako **GBM** (ang. *glioblastoma*).

Dla każdej z próbek wyznaczono ekspresję tych samych 16115 genów. W tabeli 5.2 przedstawiono liczebności poszczególnych grup pacjentów.

Ekspresje dla każdej z grup posiadają rozkład normalny. Pierwszym krokiem analizy jest znalezienie genów charakterystycznych, czyli takich, których średnia wartość ekspresji jest istotnie różna dla co najmniej dwóch z trzech przedstawionych grup. Model analizy wariancji

Grupa	**BRCA**	**GBM**	**OV**
Liczba próbek	841	166	265
L. oznaczonych genów	16115	16115	16115
Zakres ekspresji	-10,24 - 11,15	-13,62 - 16,35	-11,48 - 10,03

Tabela 5.1: Liczebności grup, liczba oznaczonych genów oraz zakres ekspresji dla każdej z trzech grup pacjentów: BRCA, GBM oraz OV analizowanych w przykładzie zastosowania algorytmu.

dla rozważanego przypadku można zapisać następująco:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{5.1}$$

gdzie Y_{ij} jest obserwowaną wartością ekspresji genu dla j-tej obserwacji i-tej grupy, μ_i jest nieznaną średnią dla i-tej grupy, ε_{ij} to błąd losowy dla j-tej obserwacji w i-tej grupie. Indeks i przyjmuje wartości 1, 2, 3 i oznacza grupę pacjentów odpowiednio: BRCA, GBM, OV. Dla każdego z genów przeprowadzany jest test ANOVA. Liczba przeprowadzonych testów brana jest pod uwagę przy wnioskowaniu o istotności różnicy średnich ekspresji między grupami. Jako parametry funkcji znajdującej charakterystyczne geny przyjęto:

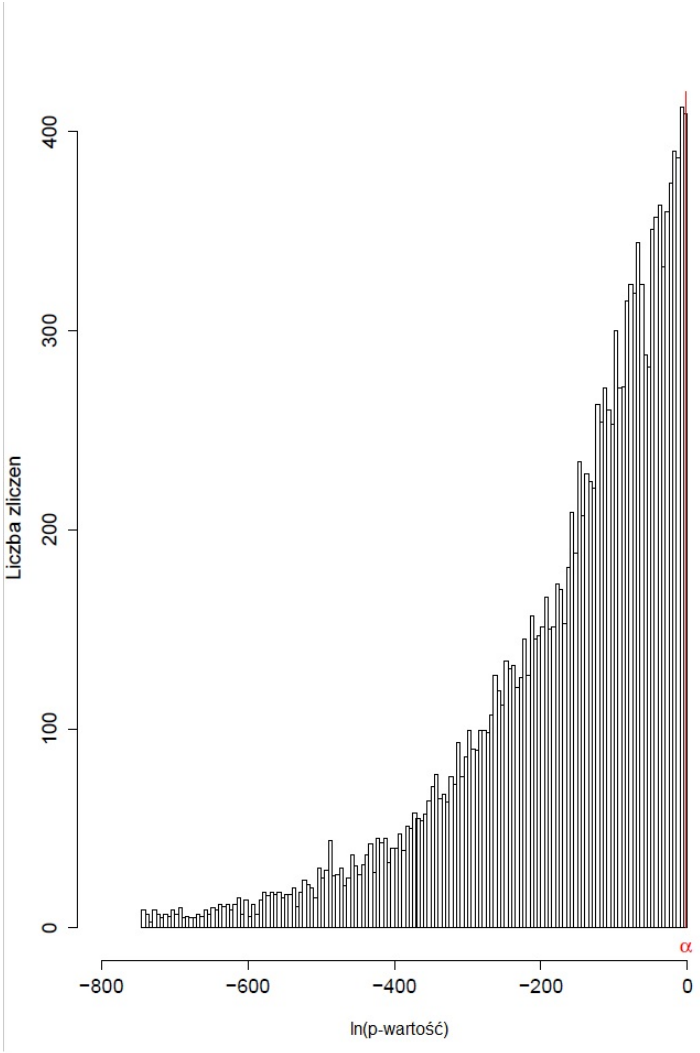
- poziom istotności - 0,1,
- liczba genów charakterystycznych - 200,
- minimalna liczba obserwacji dla każdej grupy genów - 20.

Dodatkowym parametrem funkcji jest możliwość wykonywania obliczeń równolegle, co istotnie skraca czas wykonywanych operacji. Dla przedstawionego przykładu czas obliczeń maleje proporcjonalnie do liczby wykorzystanych rdzeni. Podsumowanie wyników przedstawiono w tabeli 5.2. Rysunek 5.1 przedstawia histogram p - wartości dla wykonanych testów. Czer-

Liczba testów	16026
Liczba odrzuconych H_0 dla $\alpha = 0,1$	15812
Liczba nieodrzuconych H_0 dla $\alpha = 0,1$	214
Liczba genów wybranych jako charakterystyczne	241

Tabela 5.2: Podsumowanie wyników testów ANOVA przeprowadzanych na 16026 genach dla trzech grup: BRCA, GBM, OV.

wona linia reprezentuje poziom istotności. Geny dla których p - wartość testu znajduje się po lewej stronie od tej linii uznane są za charakterystyczne (nie uwzględniając wartości parametru liczby genów charakterystycznych).



Rysunek 5.1: Histogram logarytmu naturalnego p - wartości wyznaczonych dla testów ANOVA przeprowadzonych na 16026 genach z zaznaczonym czerwoną linią poziomem istotności.

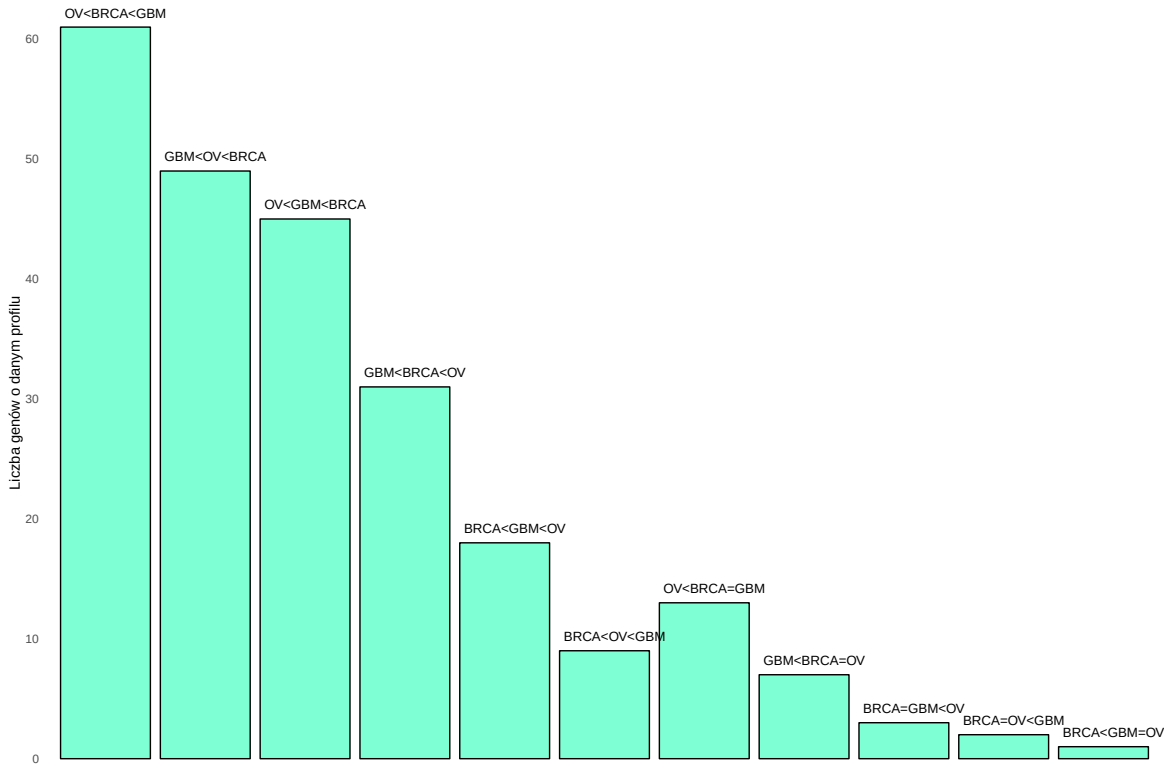
Na podstawie tabeli 5.2 oraz rysunku 5.1 można zauważyć, że liczba genów wybranych jako charakterystyczne przewyższa tę zadaną parametrem. Wynika to z równości p - wartości dla przeprowadzanych testów. Efektem tego etapu analizy jest wybranie 241 charakterystycznych genów.

Kolejnym krokiem analizy jest podział genów na zestawy. Podział ten zostanie przeprowadzony za pomocą obydwu zaimplementowanych podejść: porównań wielokrotnych oraz grupowania hierarchicznego.

Porównania wielokrotne

Znając wyniki testu analizy wariancji należy dokładniej zbadać rodzaj zależności pomiędzy średnimi ekspresjami między grupami, by podzielić geny na zestawy. Zbiór 241 genów dzielony jest za pomocą porównań wielokrotnych metodą Tukey'a na zestawy podobne ze względu na profil ekspresji (np. ekspresja w grupie BRCA jest mniejsza od ekspresji w grupie OV, która jest mniejsza od ekspresji w grupie GBM). Geny o takim samym profilu są klasyfikowane jako

należące do tego samego zestawu. Poziom istotności testów jest parametrem funkcji dzielącej geny na zestawy. W tym przykładzie przyjęto poziom istotności jako 0,1. W tym kroku również istnieje możliwość wykonania obliczeń równolegle, co skraca czas ich wykonywania. Wyniki grupowania przedstawiono na rysunku 5.2.



Rysunek 5.2: Wykres słupkowy liczby zliczeń genów o poszczególnych profilach. Zapis „BRCA≤GBM=OV" oznacza, że średnia ekspresja w grupie BRCA była mniejsza niż w grupie GBM oraz mniejsza niż w grupie OV, a średnie ekspresje w grupach GBM i OV były sobie równe.

Na podstawie porównań wielokrotnych wyróżniono 11 odrębnych zestawów genów:

- ZT_1 - BRCA≤GBM=OV, liczba elementów: 1,
- ZT_2 - BRCA=OV≤GBM, liczba elementów: 2,
- ZT_3 - BRCA=GBM≤OV, liczba elementów: 3,
- ZT_4 - GBM≤BRCA=OV, liczba elementów: 7,
- ZT_5 - BRCA≤OV≤GBM, liczba elementów: 9,
- ZT_6 - OV≤BRCA=GBM, liczba elementów: 13,
- ZT_7 - BRCA≤GBM≤OV, liczba elementów: 18,
- ZT_8 - GBM≤BRCA≤OV, liczba elementów: 31,
- ZT_9 - OV≤GBM≤BRCA, liczba elementów: 45,

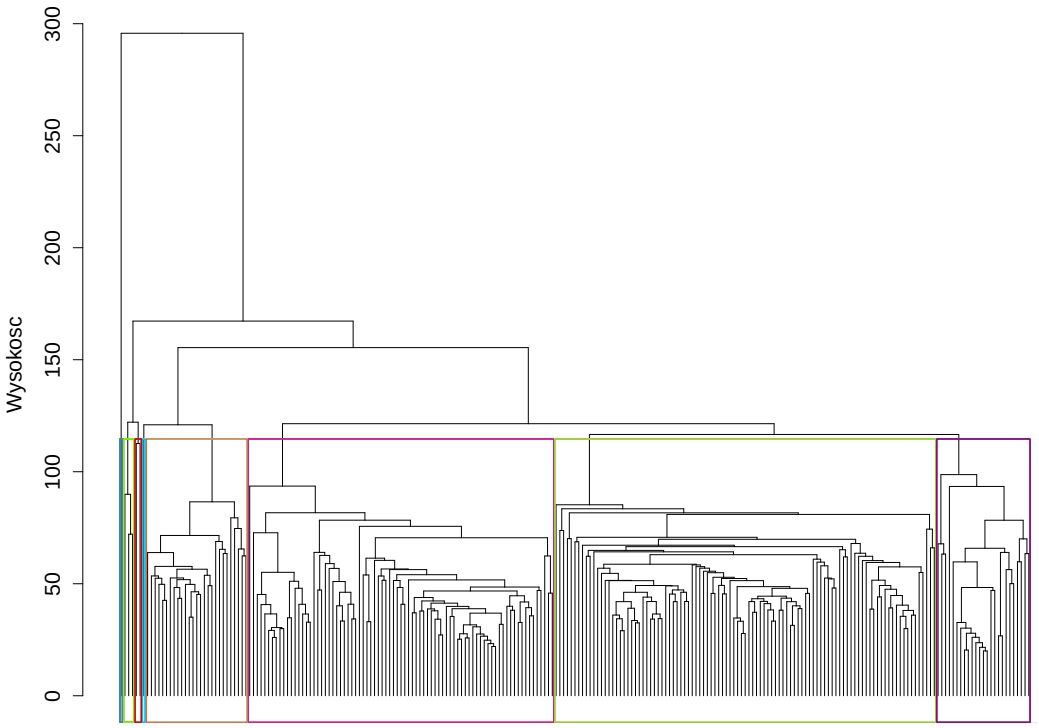
- ZT_{10} - GBM≤OV≤BRCA, liczba elementów: 49,

- ZT_{11} - OV≤BRCA≤GBM, liczba elementów: 61.

Zapis „OV≤BRCA=GBM" oznacza, że średnia ekspresja w grupie OV była mniejsza niż w grupie BRCA oraz mniejsza niż w grupie GBM, a średnie ekspresje w grupach BRCA i GBM były sobie równe. Suma elementów zestawów wynosi 239 jest różna od liczby genów charakterystycznych znalezionych przy użyciu testu ANOVA (241 genów). Wynika to z wykonywania w tym kroku porównań parami. W analizie wariancji wymagano, aby co najmniej 2 z 3 grup posiadały minimum 20 obserwacji. Oznacza to, że jedna z grup mogła posiadać liczbę obserwacji niewystarczającą do przeprowadzenia porównań wielokrotnych.

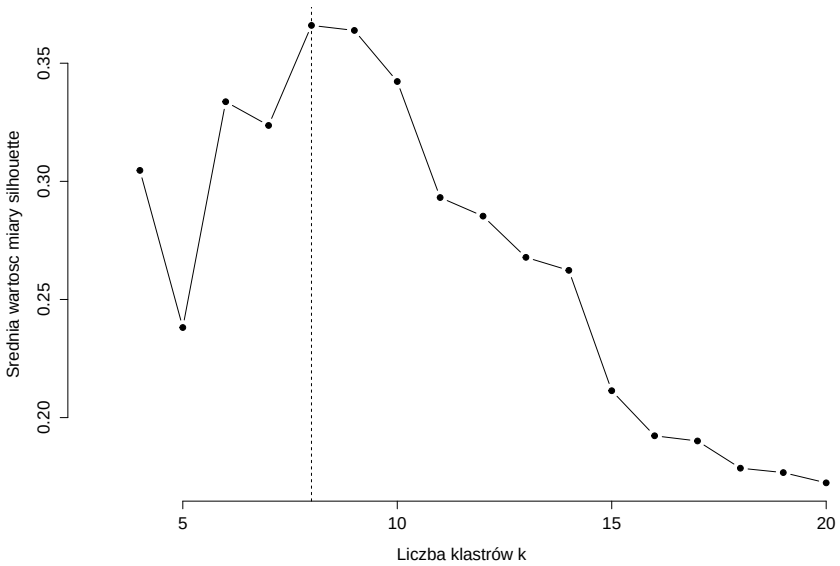
Klasteryzacja hierarchiczna

Geny mogą zostać podzielone na zestawy za pomocą jednej z nienadzorowanych technik uczenia maszynowego - klasteryzacji hierarchicznej. Do podziału 241 genów zastosowano jako miarę niepodobieństwa odległość euklidesową, której wzór dla przedstawionego przypadku przyjmuje postać
$\sqrt{(A_1 - B_1)^2 + ... + (A_{1241} - B_{1241})^2}$, gdzie A_i oraz B_i są ekspresjami genu a i B zmierzonymi dla pacjenta i. Liczba obserwacji 1241 jest równa sumie próbek w trzech grupach (5.2). Miara niepodobieństwa między klastrami jest wyznaczana za pomocą algorytmu średniego wiązania. Odległość między klastrami jest średnią wszystkich odległości między parami elementów dwóch klastrów, co dla rozważanego problemu można zapisać jako
$\frac{1}{|K1|\cdot|K2|} \sum_{a \epsilon K1} \sum_{b \epsilon K2} d(a, b)$, gdzie zapis $|K1|$, $|K2|$ oznacza liczbę genów w klastrach $K1$ oraz $K2$, $d(a, b)$ to, dla rozważanego przypadku, odległość euklidesowa między parą genów. Dendrogram, będący graficzną reprezentacją wyniku działania algorytmu klasteryzacji hierarchicznej, przedstawiono na rysunku 5.3.



Rysunek 5.3: Dendrogram 241 genów zgrupowanych na podstawie odległości euklidesowej i algorytmu średniego wiązania. Kolorami oznaczono klastry, których liczbę wyznaczono za pomocą miary silhouette [36].

Na dendrogramie zaznaczono kolorami 8 klastrów. Optymalna liczba klastrów została wyznaczona jako średnia wartość miary silhouette. Miara silhouette zdefiniowana jest jako $s_i = \frac{b_i - a_i}{max(a_i, b_i)}$, gdzie a_i jest średnią odległością między i-tym elementem a pozostałymi elementami, które należą do tego samego klastra, b_i to minimum średniej odległości między i-tym elementem klastra a elementami z pozostałych klastrów. Miara ta przyjmuje wartości od -1 do 1, im większa jej wartość tym większe prawdopodobieństwo poprawnego przypisania i-tego elementu do klastra. Wykres zmienności miary silhouette dla liczby klastrów od 4 do 20 dla analizowanego przypadku klasteryzacji 241 genów przedstawiono na rysunku 5.4. Największą średnią wartość tej miary uzyskano dla 8 klastrów, co oznacza że podział analizowanych genów na 8 zestawów jest optymalny.



Rysunek 5.4: Wykres silhouette - wykres średniej wartości miary silhouette [36] w zależności od liczby klastrów. Przerywaną linią oznaczono liczbę klastrów (8), dla której uzyskano największą wartość średniej miary silhouette.

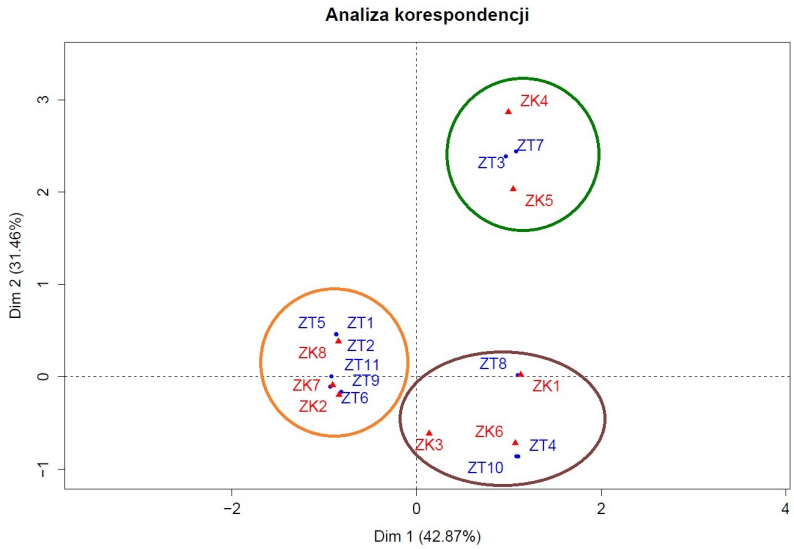
Za pomocą klasteryzacji hierarchicznej i miary silhouette wyróżniono 8 zestawów genów:

- ZK_1 - liczba elementów: 1,
- ZK_2 - liczba elementów: 3,
- ZK_3 - liczba elementów: 2,
- ZK_4 - liczba elementów: 1,
- ZK_5 - liczba elementów: 27,
- ZK_6 - liczba elementów: 81,
- ZK_7 - liczba elementów: 101,
- ZK_8 - liczba elementów: 25.

W tabeli 5.2 zaprezentowana jest liczba elementów wspólnych poszczególnych zestawów uzyskanych za pomocą metody porównań wielokrotnych i klasteryzacji hierarchicznej.

Tabela 5.3: Macierz zgodności wyników podziału 241 genów na zestawy na podstawie ekspresji genów metodą porównań wielokrotnych (ZT_i) i klasteryzacji hierarchicznej (ZK_i).

	ZK_1	ZK_2	ZK_3	ZK_4	ZK_5	ZK_6	ZK_7	ZK_8	$\sum$
ZT_1	0	0	0	0	0	0	0	1	1
ZT_2	0	0	0	0	0	0	0	2	2
ZT_3	0	0	0	0	3	0	0	0	3
ZT_4	0	0	0	0	0	7	0	0	7
ZT_5	0	0	0	0	0	0	0	9	9
ZT_6	0	0	0	0	0	0	13	0	13
ZT_7	0	0	0	1	16	0	0	1	18
ZT_8	1	0	0	0	8	22	0	0	31
ZT_9	0	3	1	0	0	2	39	0	45
ZT_{10}	0	0	1	0	0	48	0	0	49
ZT_{11}	0	0	0	0	0	0	49	12	61
$\sum$	1	3	2	1	27	81	101	25	241



Rysunek 5.5: Wykres analizy korespondencji [35] dla wartości przedstawionych w Tabeli 5.2.

Dla każdego z zestawów znajdowana jest jego charakterystyczna funkcja biologiczna. Wyszukanie takiej funkcji polega na zastosowaniu jednostronnego testu Fishera. Oddzielnie przeprowadzono wyszukanie funkcji dla zestawów genów uzyskanych za pomocą dwóch przedstawionych metod. Parametrami determinowanymi dla tego etapu analizy są: minimalna liczba genów posiadających daną funkcję (przyjęto wartość równą 5), maksymalna liczba genów posiadających daną funkcję (przyjęto 500), poziom istotności dla testu Fishera (z uwzględnieniem korekcji na liczbę testów w zestawie, przyjęto 0,1), maksymalna liczba zwracanych funkcji charakterystycznych dla zestawu (przyjęto 1) oraz brane pod uwagę subontologie (uwzględniono wszystkie subontologie - MF, BP, CC). Zestawom genów wyznaczonych przy użyciu metody porównań wielokrotnych przypisano jako charakterystyczne funkcje biologiczne przedstawione w tabeli 5.4. Występowanie kilku funkcji dla jednego zestawu genów wynika z równości p-wartości przeprowadzonych testów dla maksymalnej liczby znalezionych funkcji. W przypadku klasteryzacji hierarchicznej wyznaczane są funkcje charakterystyczne

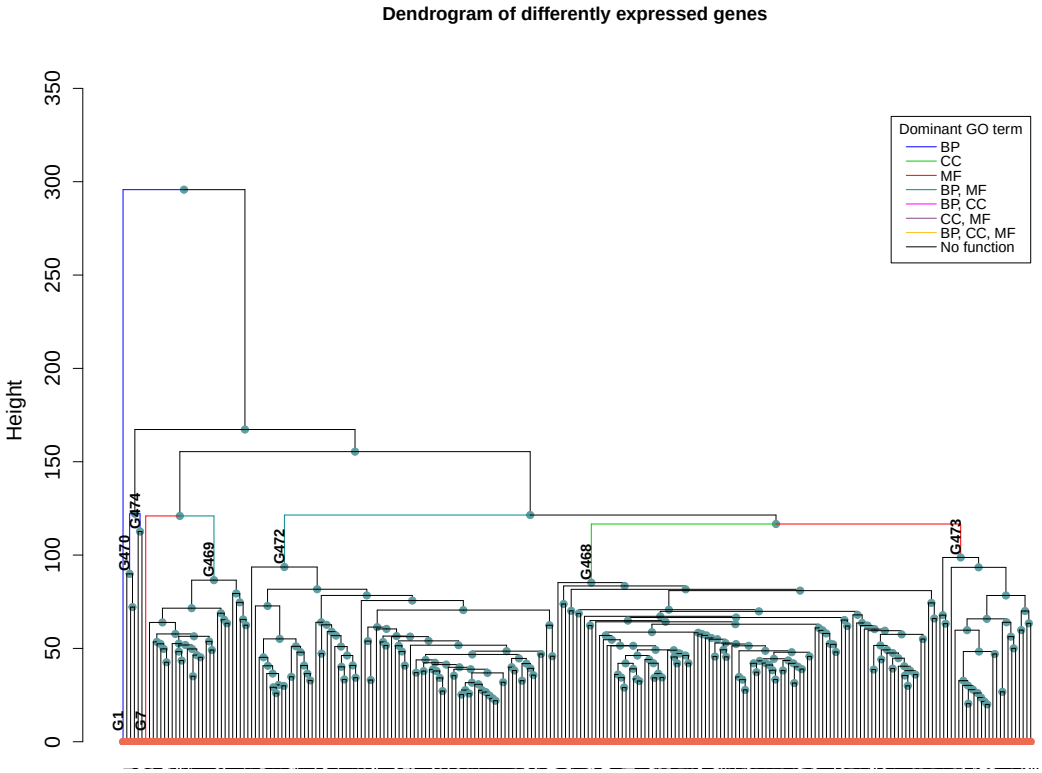
dla wszystkich znalezionych zestawów genów. Funkcje 8 wyznaczonych uprzednio zestawów genów zebrano w tabeli 5.5 wraz z informacją o nazwach genów znajdujących się w danym zestawie i opisano numerem grupy. Numer ten przedstawiony jest również na wynikowym dendrogramie 5.6, co umożliwia interpretację dendrogramu w oparciu o charakterystyczne funkcje biologiczne zestawów genów. Dodatkową informacją zaprezentowaną na wykresie jest dominująca (występująca najczęściej wśród funkcji oznaczonych jako charakterystyczne) subontologia dla danej grupy genów. Na dendrogramie przedstawiającym wynik przeprowadzonej analizy dla czytelności wykresu opisano numerem grupy 8 wybranych wcześniej zestawów.

Tabela 5.4: Przypisane charakterystyczne funkcje biologiczne do każdego z zestawów genów wyznaczonych za pomocą porównań wielokrotnych.

zestaw	identyfikator	termin
ZT_1	GO:0050804	modulation of synaptic transmission
ZT_2	GO:0005865	striated muscle thin filament
ZT_2	GO:0030239	myofibril assembly
ZT_2	GO:0051694	pointed-end actin filament capping
ZT_2	GO:0070307	lens fiber cell development
ZT_3	GO:0072372	primary cilium
ZT_4	GO:0050808	synapse organization
ZT_5	GO:0003680	AT DNA binding
ZT_6	GO:0004619	phosphoglycerate mutase activity
ZT_7	GO:0001942	hair follicle development
ZT_7	GO:0042475	odontogenesis of dentin-containing tooth
ZT_7	GO:0046854	phosphatidylinositol phosphorylation
ZT_8	GO:0051082	unfolded protein binding
ZT_8	GO:0000042	protein targeting to Golgi
ZT_8	GO:0002181	cytoplasmic translation
ZT_8	GO:0003735	structural constituent of ribosome
ZT_8	GO:0006412	translation
ZT_8	GO:0006415	translational termination
ZT_8	GO:0022625	cytosolic large ribosomal subunit
ZT_9	GO:0005923	bicellular tight junction
ZT_9	GO:0030057	desmosome
ZT_{10}	GO:0045095	keratin filament
ZT_{10}	GO:0097284	hepatocyte apoptotic process
ZT_{11}	GO:0003382	epithelial cell morphogenesis

Tabela 5.5: Przypisane charakterystyczne funkcje biologiczne do 8 wybranych zestawów genów wyznaczonych za pomocą klasteryzacji hierarchicznej.

zestaw	nr grupy	identyfikator	termin
ZK_1	G1	GO:0014002	astrocyte development
ZK_1	G1	GO:0031102	neuron projection regeneration
ZK_1	G1	GO:0045109	intermediate filament organization
ZK_1	G1	GO:0060020	Bergmann glial cell differentiation
ZK_2	G470	GO:0007171	activation of transmembrane receptor protein tyrosine kinase activity
ZK_2	G470	GO:0030856	regulation of epithelial cell differentiation
ZK_2	G470	GO:0031904	endosome lumen
ZK_2	G470	GO:0060644	mammary gland epithelial cell differentiation
ZK_2	G470	GO:0060736	prostate gland growth
ZK_2	G470	GO:0060576	intestinal epithelial cell development
ZK_3	G474	GO:0048646	anatomical structure formation involved in morphogenesis
ZK_4	G7	GO:0008757	S-adenosylmethionine-dependent methyltransferase activity
ZK_5	G469	GO:0005923	bicellular tight junction
ZK_5	G469	GO:0050804	modulation of synaptic transmission
ZK_5	G469	GO:0042043	neurexin family protein binding
ZK_6	G472	GO:0007017	microtubule-based process
ZK_6	G472	GO:0004619	phosphoglycerate mutase activity
ZK_7	G468	GO:0005923	bicellular tight junction
ZK_7	G468	GO:0030057	desmosome
ZK_8	G473	GO:0003705	transcription factor activity, RNA polymerase II distal enhancer sequence-specific binding
ZK_8	G473	GO:0045216	cell-cell junction organization
ZK_8	G473	GO:0003680	AT DNA binding
ZK_8	G473	GO:0005667	transcription factor complex



Rysunek 5.6: Dendrogram dla 241 genów z oznaczonymi etykietami zestawami genów wybranych na podstawie miary silhouette [36]. Kolorami oznaczone są domeny funkcji biologicznych,

5.3. Podsumowanie przeprowadzonej analizy

W wyniku analizy danych z TCGA dla trzech grup pacjentów BRCA, GBM i OV z 16115 genów wybrano 241 genów istotnie różnicujących te grupy. Za pomocą porównań wielokrotnych podzielono charakterystyczne geny na 11 zestawów o różnych profilach ekspresji o liczności od 1 do 61 elementów. Do każdego z zestawów przypisano najbardziej specyficzne dla niego funkcje biologiczne. 6 zestawom przypisano po 1 funkcji, dla 2 zestawów znaleziono po 2 charakterystyczne funkcje biologiczne oraz dla 3 zestawów znaleziono odpowiednio 3, 4 i 7 funkcji. Na tej podstawie można zauważyć brak zależności między licznością zestawu a liczbą funkcji dla niego charakterystycznych. Biologiczna interpretacja wskazanych funkcji pozostaje poza zakresem tej pracy.

Drugą zastosowaną do podziału genów na zestawy metodą była klasteryzacja hierarchiczna. Przy jej użyciu wyróżniono 8 zestawów genów o liczbie elementów od 1 do 101, połowa wyodrębnionych zestawów zawierała mniej niż 4 geny (grupy od ZK_1 do ZK_4). 2 zestawom przypisano po 1 charakterystycznej funkcji biologicznej, również 2 zestawom przypisano po 2 funkcje, 1 zestawowi przypisano 3 funkcje. 4 charakterystyczne funkcje przyporządkowano 3 zestawom genów i 6 funkcji przypisano do 1 zestawu. Także w tym przypadku nie obserwuje się związku między wielkością zestawu a liczbą przypisanych do niego funkcji. Obserwacje poczynione dla liczby funkcji znalezionych dla zestawów wyznaczonych za pomocą obu metod podziału są pożądane, ponieważ pokazują, że liczba znalezionych funkcji nie jest zależna od liczby genów w zestawie.

Porównując zgodność wyników podziału genów na zestawy dla obu metod, przedstawione w tabeli 5.2, można zauważyć, że elementy zestawów o niewielkiej liczności wyznaczone za pomocą jednej metody w przypadku drugiej metody są przypisywane do zestawów o znacznie większej liczbie elementów. Świadczy to o tym, że cechy, które posiadają geny, pozwalają na określenie ich jako niepodobnych do pozostałych dla jednej metody, a przy zastosowaniu drugiej metody geny te na podstawie tej samej cechy są interpretowane jako podobne do większości pozostałych. Ponadto w macierzy zgodności nie obserwuje się dużego rozproszenia wyników (tzn. 76% elementów macierzy stanowią 0), co przemawia za pewnym podobieństwem między grupami znalezionych genów.

5.4. Wnioski

Za pomocą przedstawionej analizy można wyznaczać sygnatury genetyczne opisane przez funkcje biologiczne. Zaprezentowane podejście umożliwia zrozumienie, jakie procesy biologiczne dyskryminują grupy genów o istotnie zmienionych ekspresjach.

W porównaniu do metod data mining zastosowanie porównań wielokrotnych dostarcza informacji o strukturze grupy. W przypadku, przykładowo, klasteryzacji hierarchicznej wiadome jest jedynie, że geny należące do jednego klastra są do siebie podobne w większym stopniu niż do genów należących do innego klastra. Znane jest, jakie geny zostały przypisane do danej grupy, lecz porównanie grup opiera się jedynie na informacji o nazwach genów.

Zastosowanie metody porównań wielokrotnych pozwala na podział genów na grupy scharakteryzowane przez prostą oraz informatywną różnicę w ekspresjach genów. Przekłada się to na informację o aktywnościach poszczególnych genów, np. w toku analizy wyznaczono

9-elementową grupę genów o profilu „GBM≤BRCA=OV" (co oznacza, że aktywność wskazanych genów u pacjentów z glejakiem wielopostaciowym była zmniejszona względem aktywności tych genów u pacjentów z rakiem piersi oraz rakiem jajnika, których aktywność była statystycznie równa), do której przypisano, jako charakterystyczną, funkcję biologiczną odpowiadającą za organizację synapsy: *synapse organization - a process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of a synapse, the junction between a neuron and a target (neuron, muscle, or secretory cell).*

Opisywanie każdego elementu dendrogramu przy użyciu funkcji biologicznych umożliwia wnioskowanie, w jaki sposób zmienia się charakterystyczna funkcja zbioru genów, gdy dołączane są do niego inne geny. Innymi słowy, pozwala na śledzenie zmian zachowania grupy genów w zależności od elementów, które dana grupa zawiera oraz umożliwia porównywanie dowolnych grup genów ze względu na charakterystyczne dla nich funkcje biologiczne.

Rozdział 6

Podsumowanie i wnioski

Analiza danych związanych z genami może dotyczyć różnych zagadnień i być przeprowadzana różnymi metodami. Jednym z podejść jest analiza funkcji biologicznej genów. W tym celu należy określić jakim zestawom genów zostaną przypisane charakterystyczne funkcje biologiczne. Istnieją różne metody wyznaczania takich zestawów, w tym najpowszechniej używane do tego zagadnienia, metody data mining. Funkcje biologiczne przypisane do produktów genów przechowywane są w ogólnodostępnej bazie danych Gene Ontology. Często wykorzystywanym w analizie genów narzędziem jest środowisko R. Powodem popularności R jest możliwość udostępniania w ogólnodostępnym repozytorium *Bioconductor* pakietów dedykowanych poszczególnym zagadnieniom analizy genów. W przeciwieństwie do gotowych programów, funkcje zawarte w różnych pakietach można modyfikować i dowolnie zestawiać tak, aby przeprowadzić analizę zgodnie z własnymi wytycznymi.

Za pomocą przedstawionej w pracy analizy możliwe jest wyznaczanie sygnatur genetycznych przedstawionych przy użyciu funkcji biologicznych. Dzięki omówionemu podejściu do analizy genów możliwe jest zrozumienie, jakie procesy biologiczne rozróżniają zestawy genów o istotnie zmienionych ekspresjach.

W pracy przedstawiono podstawowe sposoby wyznaczania istotnie zmienionych genów, dzielenia ich na zestawy oraz wyznaczania charakterystycznych funkcji biologicznych. Szczegółowo opisano algorytm przejścia od informacji o ekspresjach genów dla danych zbiorów (np. dwóch grup pacjentów: cierpiących na daną chorobę i zdrowych) do uzyskania wiedzy na temat funkcji biologicznych genów różnicujących dane zbiory.

Do wyznaczenia genów charakterystycznych wykorzystano analizę wariancji. Stworzono możliwość podziału genów na zestawy dwiema metodami. Pierwszą z nich były porównania wielokrotne, których wyróżniającą zaletą jest sposobność interpretacji rodzaju utworzonego zestawu i porównanie tych zestawów między sobą (ze względu na utworzone profile). Drugi sposób należący do technik data mining to klasteryzacja hierarchiczna. Umożliwia ona budowanie zestawów począwszy od jednoelementowych, a kończąc na jednym zestawie zwierającym wszystkie wejściowe elementy.

Znalezienie funkcji charakterystycznych polegało na przeprowadzeniu jednostronnego testu Fishera nadprezentacji danej funkcji biologicznej w analizowanym zestawie genów. W przypadku klasteryzacji hierarchicznej prowadzi to do uzyskania wiedzy o zmianie rodzaju biologicznej aktywności danego zestawu wraz z dołączaniem do niego nowych elementów. Ponadto, można porównywać między sobą funkcje charakterystyczne przypisane do różnych węzłów drzewa.

Przedstawiony algorytm zaimplementowano w postaci funkcji w języku R oraz C++, które następnie zebrano w formie pakietu **GOpro**. W porównaniu do istniejącego pakietu *profilesGO*, który umożliwia znajdowanie funkcji charakterystycznych, utworzony w ramach tej pracy pakiet pozwala na jednoczesne wyszukanie funkcji charakterystycznych genów dla wielu zestawów oraz wykonuje tę analizę w istotnie krótszym czasie. Dodatkowo, większość zaimplementowanych funkcji może być uruchamianych na wielu rdzeniach, co znacząco redukuje czas obliczeń w przypadku tak dużych danych, jakimi są dane wykorzystywane w analizie genów. Możliwa jest analiza co najmniej dwóch zbiorów (np. chorych i zdrowych pacjentów) jednak przedstawiony algorytm został stworzony, aby umożliwić przeprowadzanie analiz większej liczby zbiorów.

Algorytm zawarty w opisanym pakiecie został zastosowany do danych z The Cancer Genome Atlas. Do analizy wybrane zostały grupy pacjentów o następujących jednostkach chorobowych: glejaku wielopostaciowym, raku jajnika oraz raku piersi. Zastosowano obydwie metody grupowania genów, których wyniki zostały między sobą porównane. W celu porównania zestawów, spośród tych, które wyznaczono metodą klasteryzacji hierarchicznej wyodrębniono 8 zestawów za pomocą miary silhouette, co nie jest dostępne w pakiecie. Dla porównywanych metod zauważono różnicę w liczbie otrzymanych grup oraz w liczbie elementów w poszczególnych grupach. Ponadto wywnioskowano, że cechy, które posiadają geny pozwalają na określenie ich jako niepodobne do pozostałych przy użyciu jednej metody, a przeciwnie przy zastosowaniu drugiej. W przypadku obu metod nie istnieje zależność między liczbą elementów w zestawach a liczbą przypisanych zestawom charakterystycznych funkcji biologicznych.

W załącznikach do pracy umieszczono winietkę do pakietu **GOpro**, w której zaprezentowano sposób przeprowadzania przedstawionej analizy z wykorzystaniem pakietu **GOpro**. Również w tym przykładzie zastosowano opisywaną w pracy metodologię do danych z The Cancer Genome Atlas jednak w tym przypadku wybrano próbki pobrane od chorych z rakiem odbytu, rakiem jelita oraz glejakiem wielopostaciowym.

Zaprezentowana w pracy metodologia może być rozwijana w różnych kierunkach. Przykładem takiej możliwości jest stworzenie alternatywnego sposobu wyboru genów charakterystycznych, dla którego nie wymagane byłoby założenie o normalności rozkładu ekspresji genów lub o równości wariancji. Kolejnym krokiem może być rozwinięcie pakietu o dodatkowe funkcje umożliwiające grupowanie genów lub zastosowanie bardziej adekwatnego sposobu korekcji na liczbę testów w przypadku wykonywania testu Fishera dla funkcji biologicznych.

Załączniki

A. Funkcje pakietu GOpro

W tym załączniku przedstawiona jest tabela z nazwami, argumentami oraz krótkim opisem funkcji, które nie zostały zawarte w winietce pakietu GOpro (dostępnego w repozytorium pod adresem *https://github.com/lidiaad/GOpro* - stan na 7.09.2016).

Nazwa funkcji	**Argumenty**	**Opis**
ordering	x - iterator merged - wartość pola *merge* obiektu klasy *hclust*	Funkcja rekurencyjna znajduje kolejność etykiet węzłów dendrogramu.
unbundleCluster	hc - obiekt klasy *hclust*	Funkcja zwraca nazwy obiektów przypisanych do odpowiednich węzłów dendrogramu.
prepareData	RTCGA - wskazuje czy wczytać dane RTCGA cohorts - nazwy rodzajów nowotworów data - lista ramek danych z ekspresjami genów	Funkcja zwraca obiekt o uporządkowanej strukturze do wykorzystania w kolejnych krokach algorytmu.
makeNames	part.tukey.res - element listy zwracanej przez *TukeyHSDTest* sig.level - poziom istotności n - liczba porównywanych grup.	Funkcja zwraca pojedynczy string zawierający nazwy zestawów genów i relacje między nimi.
IsConflict	part.tukey.res - element listy zwracanej przez *TukeyHSDTest* sig.level - poziom istotności n - liczba porównywanych grup.	Funkcja sprawdza czy w podanych danych występuje konflikt w relacji średnich porównywanych grup.
findGO	wiele argumentów	Funkcja umożliwiająca wygodne przeprowadzenie kompleksowej analizy.
nameByProfiles	groups - lista ramek danych genów o zróżnicowanej ekspresji tukey.results - rezultaty testu Tukey'a sig.level - poziom istotności testu	Funkcja przypisuje genom nazwy zestawów, do których przynależą (na podstawie wyników testu Tukey'a).
groupByProfiles	groups - lista ramek danych genów o zróżnicowanej ekspresji tukey.results - rezultaty testu Tukey'a sig.level - poziom istotności testu	Funkcja grupuje geny w zestawy w oparciu o nazwy (na podstawie wyników testu Tukey'a).
fisher_test	a, b, c, d - elementy tablicy kontyngencji 2×2	Funkcja C++, która wykonuje test Fishera, korzysta z biblioteki *boost*. Wykorzystuje rozkład hipergeometryczny [37].
TukeyCore	z - iterator groups - lista ramek danych genów o zróżnicowanej ekspresji n - liczba grup group.names - nazwy grup	Funkcja wykonuje test Tukey'a.
tukeyHSDTest	groups - lista ramek danych genów o zróżnicowanej ekspresji parallel - wartość logiczna group.names - nazwy grup	Funkcja domyślnie wykonuje test Tukey'a równolegle (wielowątkowo).

Tabela 1: Tabela opisująca funkcje pakietu GOpro.

B. Winietka do pakietu GOpro

GOpro: Determine groups of genes and find their most characteristic GO term

Lidia Chrabaszcz

Overview

This document presents an overview of the GOpro package. This package is for determining groups of genes and finding characteristic functions for these groups. It allows forinterpreting groups of genes by their most characteristic biological function. It provides a tool for determining significantly different genes between at least two distinct groups (i.e. patients with different medical condition) which is ANOVA test with correction for multiple testing. It also gives two methods for grouping these genes. One of them is statistical that is pairwise comparisons between groups of patients' with different medical condition genes using Tukey's method. By this method profiles of genes are determined, i.e. in terms of expressions genes are grouped according to the differences in the expressions between given groups of patients with different medical condition. Another method of grouping is hierarchical clustering. Next method provides finding the most characteristic GO terms (biological functions of genes) for anteriorly obtained groups using one-sided Fisher's test for overrepresentation. If genes are grouped by hierarchical clustering, then the most characteristic function is found for all groups (for each node in the dendrogram).

Data

Data used in this example come from The Cancer Genome Atlas. They are loaded via RTCGA.PANCAN12 package. The data represent expressions of 16115 genes determined for each patient (sample). Four different medical conditions are included: acute myeloid leukemia, lung adenocarcinoma, lung squamous cell carcinoma, and endometrioid cancer.

Load the RTCGA.PANCAN12 package.

```
library(RTCGA.PANCAN12)
```

```
## Loading required package: RTCGA
```

```
## Welcome to the RTCGA (version: 1.3.3).
```

```
data("expression.cb1")
data("expression.cb2")
expr.all <- rbind(expression.cb1, expression.cb2)
rownames(expr.all) <- NULL
data("clinical.cb")
clinical <- clinical.cb
TCGAcohorts <- c('TCGA Lung Adenocarcinoma', 'TCGA Endometrioid Cancer', 'TCGA Lung Squamous Cell
Carcinoma',
                 'TCGA Acute Myeloid Leukemia')

expr <- lapply(TCGAcohorts, function (x)
  expr.all[ , c(1, na.omit(match(gsub('-', '.', clinical$sampleID[clinical$X_cohort == x]),
names(expr.all))))])

# find and remove rows with a missing observation
n <- rowSums(sapply(expr, function (x) rowSums(apply(x, 2, (is.na)))))
expr <- lapply(expr, function(x) x[n == 0, ])
expr <- lapply(expr, function(x) {row.names(x) <- as.character(x[ , 1]); x})
expr <- lapply(expr, function(x) x[, -1])
names(expr) <- c('lung adenocarcinoma', 'endometrioid', 'lung squamous', 'leukemia')
```

An example of the data:

```
##         TCGA.05.4244.01 TCGA.05.4249.01 TCGA.05.4250.01 TCGA.05.4382.01
## ?|10357           -1.13           -0.26           -1.13           -0.39
## ?|10431           -0.09            0.03            0.34           -0.16
## ?|57714            0.86            0.11            0.74           -0.13
## ?|8225            -0.78           -0.44           -0.43           -0.12
## A2LD1              0.56           -0.04            0.72            0.29
##         TCGA.05.4384.01
## ?|10357           -0.58
## ?|10431           -0.69
## ?|57714            0.40
## ?|8225             0.24
## A2LD1             -0.07
```

Example

Genes aliases must be used as genes names and they must be arranged in the same order for each group. Genes must be stored in rows and probes (samples) in columns. Data frames of genes and patients observations must be arranged into a list. The length of the list should be consistent with the number of data frames (different groups of probes).

First, load the GOpro package.

```
library(GOpro)
```

```
##
```

```
##
```

Having in mind that data should be normally distributed, the analysis of variance test is conducted by function *aovTopTest*. The *top* parameter denotes the number of significantly different genes to be given by this function. In the case of ties, all genes for which the p-value of the test is the same as for the last *top* are included in the result.

Tukey all pairwise comparisons

Significantly different genes are then grouped by pairwise comparisons.

```
tukey.results <- groupByTukey(aov.results)
```

The output is stored as a list where notation ``bladder＜colon＜gbm＜leukemia" denotes that the mean expression of bladder cancer probes is smaller than the mean expression of colon cancer probes, and they both are smaller than the mean expression of glioblastoma probes and so on for the leukemia group. The third element of the list is presented below.

```
tukey.results[[3]]
```

```
##  [1] "ARHGEF6"  "BIN2"     "EIF2C4"   "FLI1"     "FUT4"     "GIT2"
##  [7] "KCNAB2"   "LYST"     "MALAT1"   "NR3C1"    "SPN"      "TAOK3"
## [13] "TMEM106A" "ZEB2"
```

The next step is to find all GO terms annotated with each of the gene groups. The *min* and *max* parameter denote the range of counts of genes annotated with each GO term. All GO terms with counts above or below this range are omitted from the analysis. For groups obtained from all pairwise comparisons.

```
all.gos <- findAllGOs(tukey.results, geneUni = rownames(expr[[1]]), onto = c('MF', 'BP', 'CC'), min =
7, max = 600)
```

A part of the result (first five lines of the second element of a returned list) is presented below. The first value in a row is a p-value of a Fisher's test not corrected for the number of comparisons within a group.

```
all.gos[[2]][1:5]
```

```
## [[1]]
## [1] "0.0930890714696477" "GO:0003700"         "C14orf106"
## [4] "MLLT10"
##
## [[2]]
## [1] "0.118009263274349" "GO:0045944"        "C14orf106"
## [4] "MLLT10"
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
```

Then, find 2 the most characteristic biological functions (GO terms) from all annotated terms for each group. The signficance level for BH correction is set to 0.1.

```
tukey.top <- findTopGOs(all.gos, sig.level = 0.1, top = 2)
```

A part of the result (first three elements of a returned list) is presented here. Each element of a list consists of p-values annotated with selected most characteristic GO terms for the particular group. These p-values are named with the GO ID and they are corrected for the number of comparisons within a group.

```
tukey.top[1:3]
```

```
## $`endometrioid<lung adenocarcinoma<lung squamous<leukemia`
##   GO:0016925   GO:0007179
## 6.421444e-05 1.657692e-03
##
## $`endometrioid<lung adenocarcinoma=lung squamous<leukemia`
## [1] NA
##
## $`endometrioid<lung squamous<lung adenocarcinoma<leukemia`
##  GO:0036065  GO:0042742  GO:0043507
## 0.005201751 0.018835198 0.018835198
```

The results of the analysis can be presented in a more descriptive way or in a concise one. The concise way is obtained by using *GO* function (additionally information about p-values returned by *findTopGOs* are included). To get more descriptive results use *extendGO* function. First elements of the output are presented here.

```
print(GO(all.gos, tukey.top))[1:2, ]
```

```
##                                                    profile
## endometrioid<lung adenocarcinoma<lung squamous<leukemia "endometrioid<lung adenocarcinoma<lung
squamous<leukemia"
## endometrioid<lung adenocarcinoma=lung squamous<leukemia "endometrioid<lung adenocarcinoma=lung
squamous<leukemia"
##                                                    GOs
## endometrioid<lung adenocarcinoma<lung squamous<leukemia Character,2
## endometrioid<lung adenocarcinoma=lung squamous<leukemia NULL
##                                                    p.values
```

```
## endometrioid<lung adenocarcinoma<lung squamous<leukemia Numeric,2
## endometrioid<lung adenocarcinoma=lung squamous<leukemia NA
##                                                      GENES
## endometrioid<lung adenocarcinoma<lung squamous<leukemia "C14orf118 CDK6 CRLF3 DENND4A PIAS1 RNF24
USP15"
## endometrioid<lung adenocarcinoma=lung squamous<leukemia "C14orf106 MLLT10"
```

```
extendGO(tukey.top)[1:2, ]
```

```
##   GROUP      GO ID
## 1     1 GO:0016925
## 2     1 GO:0007179
##                                                      TERM
## 1                                    protein sumoylation
## 2 transforming growth factor beta receptor signaling pathway
##
DEFINITION
## 1      The process in which a SUMO protein (small ubiquitin-related modifier) is conjugated to a
target protein via an isopeptide bond between the carboxyl terminus of SUMO with an epsilon-amino group
of a lysine residue of the target protein.
## 2 A series of molecular signals initiated by the binding of an extracellular ligand to a
transforming growth factor beta receptor on the surface of a target cell, and ending with regulation of
a downstream cellular process, e.g. transcription.
##   ONTOLOGY
## 1       BP
## 2       BP
```

Hierarchical clustering

Alternatively, significantly different genes may be grouped by hierarchical clustering.

```
cluster.results <- clustering(aov.results, clust.metric = "euclidean", clust.method = "centroid")
```

Then, find all GO terms annotated with each of the gene groups.

```
all.gos <- findAllGOs(cluster.results, geneUni = rownames(expr[[1]]), onto = c('MF', 'BP', 'CC'), min =
7, max = 600)
```

A part of the result (first five lines of the second element of a returned list) is presented below. The first value in a row is a p-value of a Fisher's test not corrected for the number of comparisons within a group.

```
all.gos[[2]][1:5]
```

```
## [[1]]
## [1] "0.00146382266833961" "GO:0001942"          "EDARADD"
##
## [[2]]
## [1] "0.0155792555416144" "GO:0030154"         "EDARADD"
##
## [[3]]
## [1] "0.0108741112505228" "GO:0033209"         "EDARADD"
##
## [[4]]
## [1] "0.0019866164784609" "GO:0042475"         "EDARADD"
##
## [[5]]
## NULL
```

Then, find 2 the most characteristic biological functions (GO terms) from all annotated terms for each group. The signficance level for BH correction is set to 0.1.

```
cluster.top <- findTopGOs(all.gos, sig.level = 0.1, top = 2)
```

A part of the result (first three elements of a returned list) is presented here. Each element of a list consists of p-values annotated with selected most characteristic GO terms for the particular group. These p-values are named with the GO ID and they are corrected for the number of comparisons within a group.

```
cluster.top[1:3]
```

```
## $G1
## GO:0001932 GO:0002931 GO:0005518 GO:0006953 GO:0009611 GO:0010952
## 0.00476788 0.00476788 0.00476788 0.00476788 0.00476788 0.00476788
## GO:0031093 GO:0035924 GO:0035987 GO:0036120 GO:0045773 GO:0050921
## 0.00476788 0.00476788 0.00476788 0.00476788 0.00476788 0.00476788
## GO:0070372 GO:0071380 GO:0071560
## 0.00476788 0.00476788 0.00476788
##
## $G2
##  GO:0001942  GO:0042475
## 0.003973233 0.003973233
##
## $G3
## GO:0001523 GO:0005796 GO:0006027 GO:0006775 GO:0017134 GO:0031225
## 0.00448109 0.00448109 0.00448109 0.00448109 0.00448109 0.00448109
## GO:0043236
## 0.00448109
```

The results of the analysis can be presented in a more descriptive way or in a concise one. The concise way is obtained by using *GO* function (additionally information about p-values returned by *findTopGOs* are included). To get more descriptive results use *extendGO* function. The first five rows of a result are presented below for each function.

```
print(GO(all.gos, cluster.top))[1:2, ]
```

```
##    profile GOs          p.values   GENES
## G1 "G1"    Character,15 Numeric,15 "FN1"
## G2 "G2"    Character,2  Numeric,2  "EDARADD"
```

```
extendGO(cluster.top)[1:2, ]
```

```
##   GROUP      GO ID                                  TERM
## 1     1 GO:0001932 regulation of protein phosphorylation
## 2     1 GO:0002931                 response to ischemia
##
DEFINITION
## 1                                                                  Any process that modulates
the frequency, rate or extent of addition of phosphate groups into an amino acid in a protein.
## 2 Any process that results in a change in state or activity of an organism (in terms of movement,
secretion, enzyme production, gene expression, etc.) as a result of a inadequate blood supply.
##   ONTOLOGY
## 1       BP
## 2       BP
```

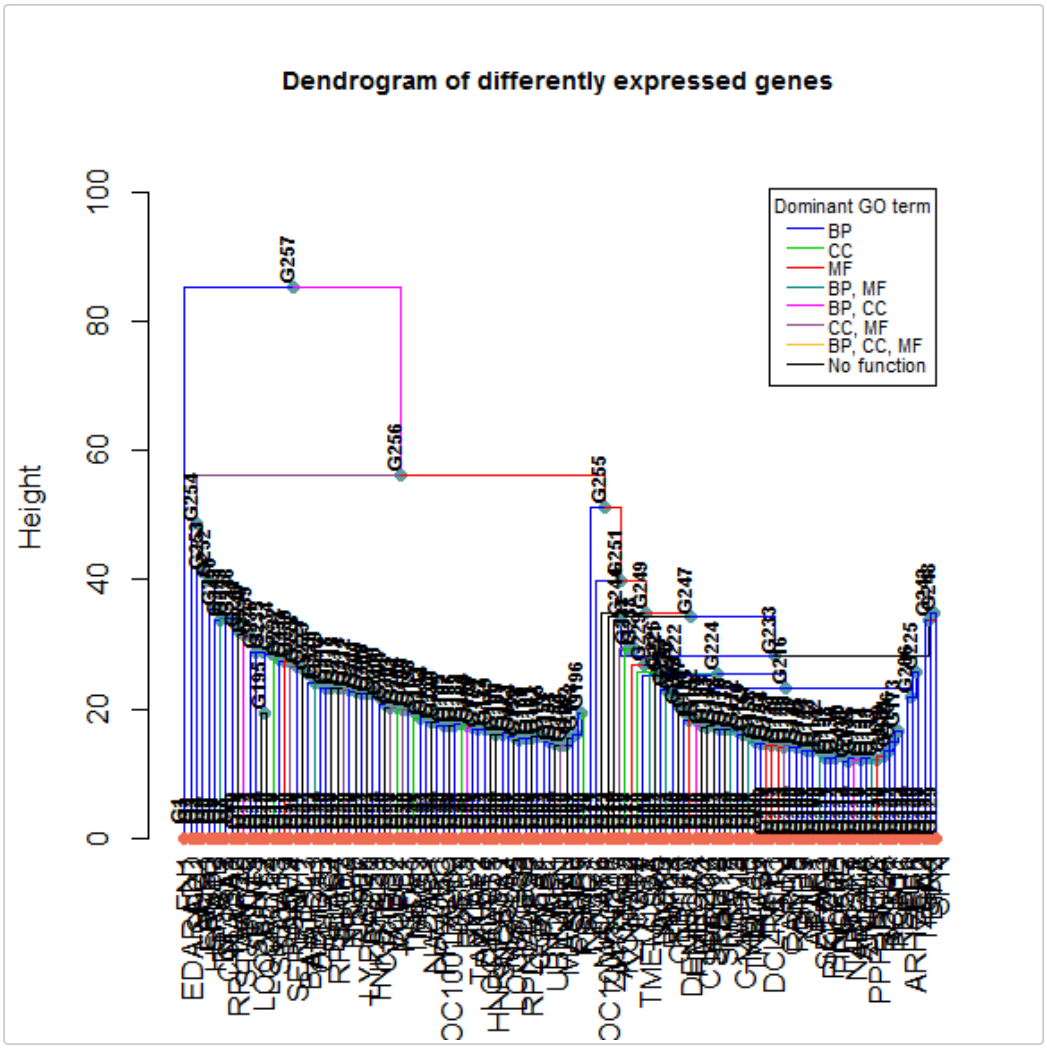
In order to present the results graphically the following functions must be used.

```
unbundled <- unbundleCluster(cluster.results)
printout <- GO(all.gos, cluster.top)
```

```
dendro <- geneclusterplot(printout, unbundled, cluster.top)
```

The result of the clustering can be presented in the dendrogram. The parameter *over.represented* indicates if the dominant GO terms should be calculated and presented in the plot. The dominant GO term is the most frequently represented ontology by all GO terms characteristic for each node of the dendrogram.

```
plotg(dendro, over.represented = TRUE)
```



Bibliografia

[1] Y. Rahmatallah, F. Emmert-Streib, G. Glazko, *Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline,* Briefings in Bioinformatics, 2015, s. 1-15.

[2] O. Huber et al., *rchestrating high-throughput genomic analysis with Bioconductor*, Nature Methods, 2015, 115.

[3] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2013.

[4] The Cancer Genome Atlas: *https://tcga-data.nci.nih.gov/tcga/* (data dostępu: 10.02.2016).

[5] L. Peterson, *Classification Analysis of DNA Microarrays*, Wiley, 2013.

[6] H. Itadani, S. Mizuarai, H. Kotani, *Can systems biology understand pathway activation? Gene expression signatures as surrogate markers for understanding the complexity of pathway activation,* Curr Genomics, s. 349-360.

[7] S. Corsello et al., *Identification of AML1-ETO modulators by chemical genomics,* Blood 2009, s. 6193-6205.

[8] Gene Ontology Consortium: *www.geneontology.org* (data dostępu: 25.02.2016).

[9] Notatki z wykładu: *Wielkoskalowe metody pomiarowe w biologii molekularnej*, prowadzący: T. Rubel.

[10] Struktura ontologii genów: *http://geneontology.org/page/ontology-structure* (data dostępu: 25.02.2016).

[11] Baza danych GO: *http://geneontology.org/page/lead-database-guide* (data dostępu: 25.02.2016).

[12] Informacje dotyczące analizy wzbogacenia genów: *http://lectures.molgen.mpg.de/Algorithmische_Bioinformatik_WS1213/lec14a_Shamir_GO_GSEA.pdf* (data dostępu: 18.02.2016).

[13] A. Subramanian et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,* PNAS, 2009, 102.

[14] M. Kutner, C. Nachtsheim, J. Neter, W. Li, *Applied Linear Statistical Models*, McGraw-Hill/Irwin, 2004.

[15] O. Dunn, *Estimation of the Medians for Dependent Variables*, Annals of Mathematical Statistics 30, s. 192–197.

[16] H. Sheffe, *The Analysis of Variance*, Wiley, 1999.

[17] J. Tukey, *Comparing Individual Means in the Analysis of Variance*, Biometrics, 1949, 5, s. 99-114.

[18] B. Everitt, S. Landau, M. Leese, *Cluster Analysis, 5th Edition*, Wiley, 2011.

[19] A. Jain, M. Murty, P. Flynn, *Data Clustering: A Review*, ACM Computing Surveys, 1999, 31, s. 265-323.

[20] R. Fisher, *Statistical Methods for Reaserch Workers*, Hafner Publishing Company Inc., 1954.

[21] Y. Hochberg, A. C. Tamhane, *Multiple Comparison Procedures*, Wiley, 1987.

[22] Z. Šidák, *Rectangular Confidence Regions for the Means of Multivariate Normal Distributions*, Journal of the American Statistical Association 1967, 62, s. 626–633.

[23] S. Holm, *A simple sequentially rejective multiple test procedure*, Scandinavian Journal of Statistics, 1979, 6, s. 65–70.

[24] Y. Hochberg, *A Sharper Bonferroni Procedure for Multiple Tests of Significance*, Biometrika, 1988, 75, s. 800–802.

[25] C. W. Dunnett, *A multiple comparison procedure for comparing several treatments with a control*, Journal of the American Statistical Association, 1955, 50, s. 1096–1121.

[26] Y. Benjamini, Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society, 1995, 57, s. 289–300.

[27] Y. Benjamini, D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, Annals of Statistics, 2001, 29, s. 1165–1188.

[28] P. Biecek, *RTCGA.PANCAN12: PanCan 12 from Genome Cancer Browser*, R package version 0.1, 2015.

[29] Narodowy Instytut Zdrowia Stanów Zjednoczonych - National Cancer Institute: *https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode* (data dostępu: 10.12.2015).

[30] National Cancer Institute, National Human Genome Research Institute: *https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=Tissue%20Source%20Site* (data dostępu: 10.12.2015).

[31] M. Carlson, *org.Hs.eg.db: Genome wide annotation for Human*, R package version 3.2.3.

[32] Definicja neddylacji: *https://en.wikipedia.org/wiki/Neddylation* (data dostępu: 25.02.2016).

[33] M. Carlson, *org.Hs.eg.db: Genome wide annotation for Human*, R package version 3.2.3.

[34] M. Carlson, *GO.db: A set of annotation maps describing the entire Gene Ontology*, R package version 3.2.2.

[35] H. Hirschfeld, *A connection between correlation and contingency*, Proc. Cambridge Philosophical Society, 1935, 31, s. 520–524.

[36] P. Rousseeuw, *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis*, Computational and Applied Mathematics, 1987, 20, s. 53–65.

[37] Rozkład hipergeometryczny w bibliotece *boost*: *http://www.boost.org/doc/libs/1_60_0/libs/math/doc/html/math_toolkit/dist_ref/dists/hypergeometric_dist.html* (data dostępu: 26.11.2015).