Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Joanna Zabielska
Nr albumu: 292579

Identyfikacja wariantów splicingowych na podstawie mikromacierzy tilingowych

Praca licencjacka
na kierunku BIOINFORMATYKA I BIOLOGIA SYSTEMÓW

Praca wykonana pod kierunkiem
dr. inż. Przemysław Biecek
Instytut Matematyki Stosowanej i Mechaniki

Wrzesień 2012

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

Praca przedstawia metody analizy wstępnej danych mikromacierzowych oraz sposoby identyfikacji wariantów splicingowych u gatunku *Arabidopsis thaliana*. Ukazuje także różne możliwości oraz darmowe pakiety dostępne w programie R, umożliwiające przeprowadzenie wyżej wymienionych analiz oraz wykonanie statystycznej obróbki danych wielkoskalowych.

Słowa kluczowe

Mikromacierze, splicing index, język R, *Arabidopsis thaliana*

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11300 (Informatyka, nauki komputerowe)

Klasyfikacja tematyczna

D. Software

Tytuł pracy w języku angielskim

Splice variants identyfication with the use of tiling arrays

Spis treści

Wprowadzenie

Celem mojej pracy jest przeprowadzenie analizy danych wielkoskalowych uzyskanych za pomocą mikromacierzy tilingowych tak, by pokazać jakie są różnice w wariantach spicingowych genów u organizmów gatunku *Arabidopsis thaliana* i jak wpływają one na ich fenotyp. Następny krok to napisanie kodu w języku R, który będzie umożliwiał wykonanie takiej analizy. Chcę wykazać, że u organizmów z fenotypem ukazującym się w sytuacjach stresowych, wariant splicingowy danego genu wygląda inaczej, niż w przypadku genu osobnika, u którego w takich warunkach nie obserwujemy żadnego fenotypu.

Dzięki co raz częściej używanym mikromacierzom tilingowym, służącym między innymi do identyfikowania miejsc splicingowch oraz wielu metodom statystycznym i gotowym, bezpłatnym modułom dostępnym w pakiecie R, taka analiza jest możliwa, stosunkowo prosta i nie pochłania zbyt wiele czasu. Pozwala także na otrzymanie wielu cennych dla biologów informacji, które w późniejszym czasie mogą być pomocne w zrozumieniu wielu aspektów biologicznych i w znacznym stopniu ułatwiają im późniejszą pracę w laboratorium.

Praca skład się z pięciu rozdziałów. W pierwszym z nich zostaje krótko opisany badany przez nas organizm modelowy - *Arabidopsis thaliana*. W kolejnych, przedstawione są materiały (rozdział 2) oraz metody (rozdział 3) wykorzystywane do przeprowadzenia eksperymentu mikromacierzowego oraz późniejszej analizy otrzymanych danych. Wytłumaczone zostało jak zbudowane są mikromacierze, jakie są ich rodzaje, a także jakie metody służą do wykonania analiz wyższego oraz niższego rzędu. Kolejny rozdział zawiera opis procesu splicingu i działanie metody używanej do obliczania wartości splicing index, po to by umożliwić identyfikacje wariantów splicingowych genu. Na końcu (rozdział 5), przedstawiono wyniki przeprowadzanej analizy danych oraz wnioski, które dzięki niej, dało się zaobserwować. W skład pracy wchodzi także dodatek, w którym zawarte są kody najistotniejszych funkcji, napisane w języku R.

Rozdział 1

Arabidopsis thaliana

Arabidopsis thaliana (rzodkiewnik pospolity) [1] jest to gatunek rośliny jednorocznej z rodziny kapustowatych (*Brassicaceae / Cruciferae*). Występuje w Azji i Europie oraz na niektórych obszrach Afryki Północnej. Ponieważ zasięg ekologiczny i geograficzny tego gatunku jest stosunkowo duży, obserwuje się dużą zmienność genetyczna między osobnikami różnych populacji. Dotyczy ona głownie takich cech jak wrażliwość na temperaturę, długość dnia, czy pora kwitnienia. Powszechnie roślina ta uważana jest za chwast, jednak dzięki krótkiemu cyklowi rozwojowemu (trwającemu około 6-8 tygodni), dużej ilość wytwarzanych nasion (nawet do 5000 nasion) oraz małej ilość genów (125Mbp DNA) i jedynie 5 parom chromosomów *Arabidopsis thaliana* od dawna jest wykorzystywana jako organizm modelowy w badaniach genetycznych. Nie bez znaczenia jest też fakt, że jest to mało wymagająca i odporna roślina. Nie ma specjalnych wymagań życiowych i na ogół bardzo dobrze rośnie w każdych warunkach. Tak więc jej hodowla jest tania, natomiast samopylność tej rośliny umożliwia łatwe krzyżowanie i utrzymanie czystych linii. Taka charakterystyka sprawia, że jest ona atrakcyjnym obiektem badań i analiz dla biologów.

Rysunek 1.1: *Arabidopsis thaliana*. Źródło: http://commons.wikimedia.org/wiki/File:Arabidopsis_thaliana_inflorescencias.jpg

Istnieje wiele baz danych zawierających informacje o *Arabidopsis thaliana* takich jak **The Arabidopsis thaliana Integrated Database (ATIDB)**[1], czy **The Arabidopsis Information Resource (TAIR)**[2]. Jej genom został zsekwencjonowany w całości w 2000r. Zachęca to wielu badaczy i ułatwia im przeprowadzanie eksperymentów na tym organizmie modelowym.

W laboratoriach, gdzie warunki hodowli są inne niż w naturze, kontrolowane i stałe, wykorzystuje się najczęściej linie homozygotyczne, pochodzące z nasion „dzikich". Powszechnie używanymi liniami są Columbia Wild-type i Landsberg erecta, obie wyselekcjonowane z ekotypu Landsberg (Niemcy), skąd *Arabidopsis* pochodzi.

[1] http://atidb.org/
[2] http://www.arabidopsis.org/

Rozdział 2

Mikromacierze DNA

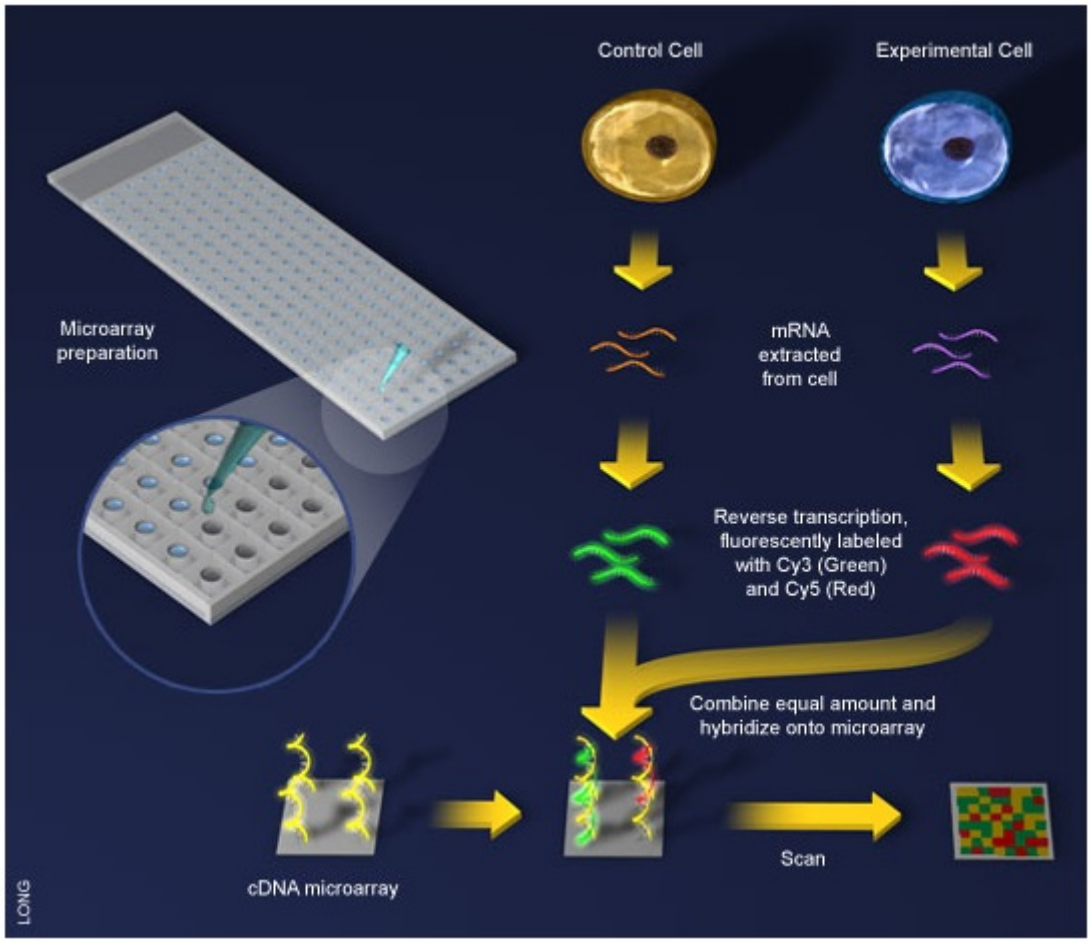
Mikromacierz jest to płytka szklana lub plastikowa, do której dołączone są w ściśle określonym porządku sondy molekularne. Służą do analizy transktyptomicznej lub badania DNA organizmów. Wykorzystywane są głównie przy określaniu zmian w poziomie ekspresji genów, ale także do identyfikacji miejsc wiązania czynników transkrypcyjnych, metylacji DNA i innych elementów związanych z regulacją transkrypcji. Dzięki ich zminiaturyzowanej budowie, możliwe jest jednoczesne badanie ekspresji nawet kilkudziesięciu tysięcy sekwencji jednocześnie. Mikromacierze w dzisiejszych czasach są powszechnie stosowane. Korzystanie z nich staje się co raz tańsze, dlatego też znalazły swoje zastosowanie w wielu dziedzinach związanych z biologią, takich jak medycyna, diagnostyka, farmacja czy badania laboratoryjne.

Rysunek 2.1: Przykładowa mikromacierz. Źródło: http://neurophilosophy.wordpress.com/2006/08/21/reseachers-will-use-dna-microarrays-to-probe-for-autism-genes/

Każda mikromacierz składa się z kilkudziesięciu tysięcy sond, umieszczonych w równych odstępach. Sonda jest to krótki łańcuch nukleotydów o sekwencji charakterystycznej dla określonego genu. Długość łańcucha nukleotydowego w sondzie oraz odstępy między nimi są zależne od rodzaju mikromacierzy. Na ogół sondy są pogrupowane w zespoły kilkunastu par komplementarnych do różnych regionów tego samego transkryptu, co zapewnia większą czułość. Każda taka para składa się z sondy perfect match (PM), czyli sondy w pełni

komplementarnej oraz sondy posiadającej jeden, najczęściej centralny, niekomplementarny nukleotyd nazywanej sondą mismatch (MM). Służy to zwiększeniu specyficzności sygnału po hybrydyzacji oraz umożliwia określenie wartości tła.

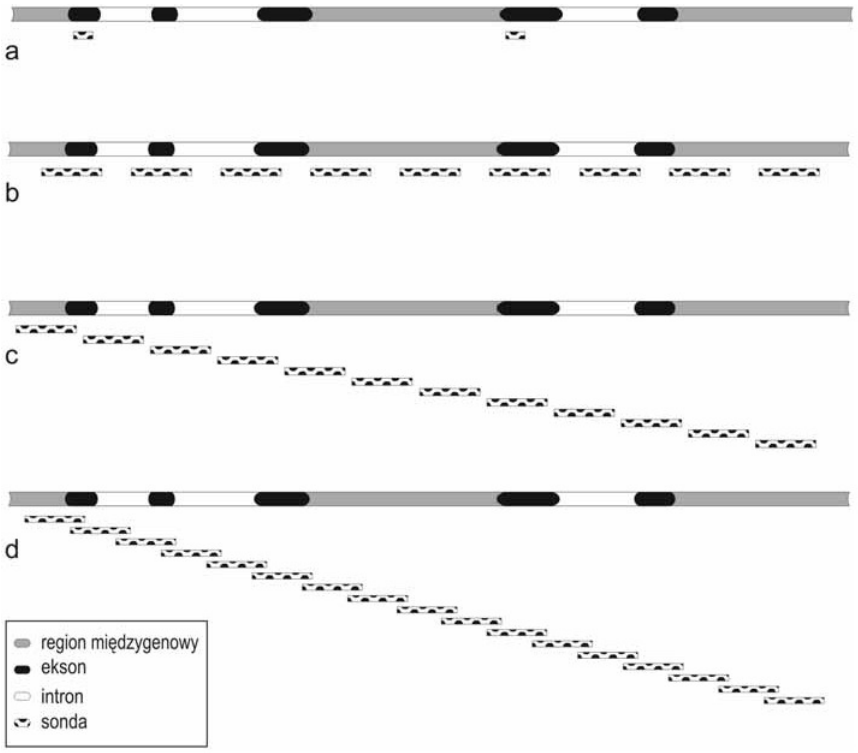
W celu przeprowadzenia eksperymentu z wykorzystaniem mikromacierzy, izoluje się RNA z komórek, które chce się przebadać, następnie z oczyszczonego RNA syntetyzuje się cDNA. Kolejnym etapem jest wyznakowanie cDNA znacznikiem fluorescencyjnym, a następnie poddanie hybrydyzacji z mikromacierzą. Cząsteczki badanego materiału wiążą się do komplementarnych sekwencji, którymi są sondy molekularne związane z płytką. Następnie skanuje się mikromacierz za pomocą odpowiedniego skanera fluorescencyjnego. W rezultacie otrzymuje się siatkę plamek fluorescencyjnych, oznaczających hybrydyzacje sekwencji komplementarnych do mikromacierzy, co identyfikuje geny, które ulegają ekspresji w badanej próbce. Ostatnim etapem jest analiza wyników za pomocą cyfrowej analizy obrazu. Na podstawie intensywności fluorescencji, dostaje się informację o poziomie ekspresji genów.



Rysunek 2.2: Przebieg eksperymentu mikromacierzowego do momentu uzyskania surowych danych. PRÓBKA→ znakowanie znacznikami fluorescencyjnymi → hybrydyzacja → skanowanie → ilościowa analiza obrazu → SUROWE DANE LICZBOWE Źródło: http://neurophilosophy.wordpress.com/2006/08/21/ reseachers-will-use-dna-microarrays-to-probe-for-autism-genes/

2.1. Mikromacierze tilingowe

Mikromacierze tilingowe [3], czyli mikromacierze równomiernie pokrywające genom (zwane też mikromacierzami dachówkowymi), jest to rodzaj mikromacierzy DNA. Służą one do identyfikacji w organizmie nieznanych do tych czas produktów transkrypcji, takich jak: miejsca wiązania się czynników transkrypcyjnych, metylacji DNA, czy modyfikacji histonów. Dzięki ich budowie, za pomocą jednej płytki mikromacierzowej, możliwe jest zbadanie całego genomu m.in. takich organizmów jak *Arabidopsis thaliana*. Działanie mikromacierzy tilingowych jest bardzo podobne jak w innych mikromacierzach DNA, jednak są pewne różnice w rodzaju wykorzystywanych sond.

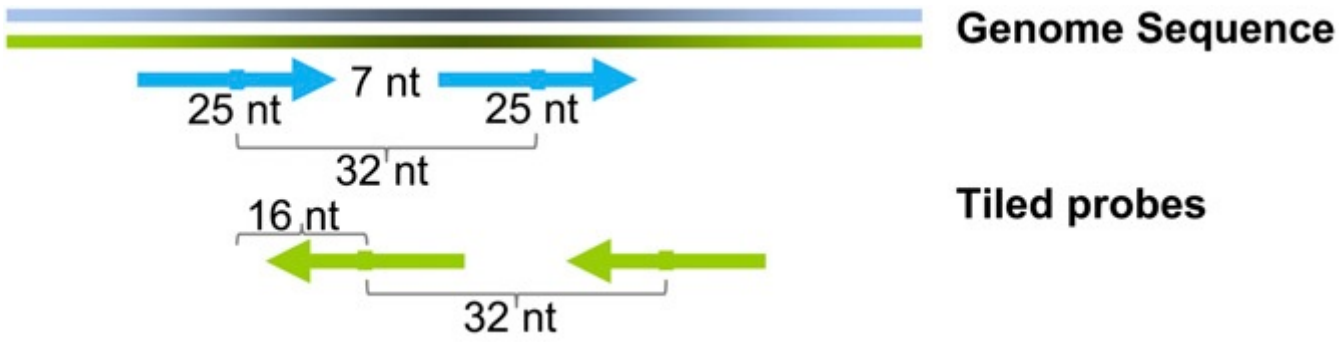


Rysunek 2.3: Porównanie sposobu rozłożenia sond w mikromacierzach tilingowych oraz mikromacierzach ekspresyjnych. a) mikromacierze ekspresyjne, b) - d) mikromacierze równomiernie pokrywające genom: b) małej rozdzielczości, c) rozdzielczości równej długości sond, d) wysokiej rozdzielczości. Źródło: [3] (Rys. 1)

W przeciwieństwie do mikromacierzy ekspresyjnych, za których pomocą możemy analizować jedynie znane transkrypty, w mikromacierzach tilingowych używa się sond komplementarnych do badanego fragmentu, bez względu na stan wiedzy o znaczeniu funkcjonalnym tego odcinka DNA. Obecnie używane mikromacierze tilingowe zawierają sondy o długości kilkudziesięciu nukleotydów. Jedynym charakterystycznym parametrem mikromacierzy jest jej rozdzielczość. Rozdzielczością (inaczej „krokiem") nazywa się odległość między środkami sekwencji sąsiadujących ze sobą sond na danej mikromacierzy. Jeżeli rozdzielczość mikromacierzy jest mniejsza lub równa długości sondy, otrzymujemy mikromacierze, które pokrywają genom w sposób ciągły, czyli na przykład mikromacierze tilingowe. Dzięki tak gęstemu rozłożeniu sond, zwiększa się ilość uzyskanych informacji podczas eksperymentu, co z kolei wpływa na dokładność wyników analiz.

2.2. Mikromacierze AGRONOMICS1

Najczęściej używanymi mikromacierzami do badania genomu *Arabidopsis thaliana* są mikromacierze Affymetrix ATH1. Jednak nie wykrywają one około 1/3 genów odnotowanych w szczepie referencyjnym. Alternatywą dla nich jest AGRONOMICS1 [4], mikromacierz, która pokrywa cały genom tego gatunku, z wyjątkiem powtarzalnych sekwencji, które mogłyby powodować cross hybrydyzacje. W sumie zostało wyłączone prawie 14 milionów nukleotydów. Oba rodzaje mikromacierzy dają wyniki podobnej jakości, jednak AGRONOMICS1 dzięki swojej budowie, umożliwia uzyskanie informacji dotyczącej ekspresji większej ilości genów. AGRONOMICS1 posiada sondy zawierające 25 nukleotydów, zaprojektowane na przemian nici, aby pokryć obie (dodatnią i ujemną) nici genomu. Dzięki temu możemy badać na raz dwie nici. Mediana odległości między środkami sond na tej samej nici genomu wynosi 32 nukleotydy, a na przeciwległych niciach - 16 nukleotydów. Dzięki temu, rozdzielczość mikromacierzy jest mniejsza od długości sondy i otrzymujemy mikromacierz pokrywającą genom w sposób ciągły. Mitochondrialny oraz chloroplastowy genom został w całości zaprezentowany na mikromacierzy. Ponieważ możliwe jest przeprowadzenie analizy jedynie na podstawie sond PM, AGRONOMICS1 nie posiada sond mismatch. Zawiera natomiast wszystkie sondy perfect match pochodzące z ATH1, co umożliwia łatwą integracje z istniejącymi bazami danych dla ATH1. W sumie mikromacierz ta zbudowana jest z 152 065 sond kontrolnych, ok. 5 894 000 sond tilingowych *Arabidopsis* (mniej więcej po tyle samo dla nici dodatniej i ujemnej) oraz ok. 250 100 sond PM pochodzących z ATH1.



Rysunek 2.4: Pokrycie obu nici genomu. Średnia odległość między środkami sekwencji sond sąsiadujących ze sobą na tej samej nici wynosi 32 nukleotydy. W przypadku sond na przeciwstawnych niciach odlgłość ta jest równa 16 nukleotydów. Średnia odległość między dwiema sondami wynosi 7 nukleotydów. Źródło: [4] (Figure 1)

AGRONOMICS1 jest używana między innymi do badania poziomu transkrypcji ponad 29 000 znanych genów *Arabidopsis*, profilu transkrypcji genów jeszcze nie poznanych, ekspresji genów i stanu chromatyny i wykrywania miejsc splicingowych, co wykorzystuje w mojej pracy.

Rozdział 3

Analiza danych

Po zeskanowaniu mikromacierzy skanerem fluorescencyjnym i otrzymaniu surowych danych liczbowych, należy poddać je różnego rodzaju analizom. Na samym początku należy pozbyć się tych wyników, które są efektem niedoskonałości technicznych takich jak: wpływ barwników fluorescencyjnych na hybrydyzacje, nierównomierne odmycie niektórych rejonów mikromacierzy. Następnie przechodzi się do analiz niższego i wyższego rzędu.

Rysunek 3.1: Opis procesu analizy danych. Źródło: Opracowanie własne na podstawie [6] (Rys. 2)

3.1. Analizy niższego rzędu - normalizacja RMA

Normalizacja RMA (Robust Multichip Average) jest to metoda służąca do wstępnej obróbki danych mikromacierzowych. Jej kolejnymi etapami jest: korekcja tła, normalizacja kwantylowa oraz sumowanie z wykorzystaniem algorytmu "wygładzania" mediany.

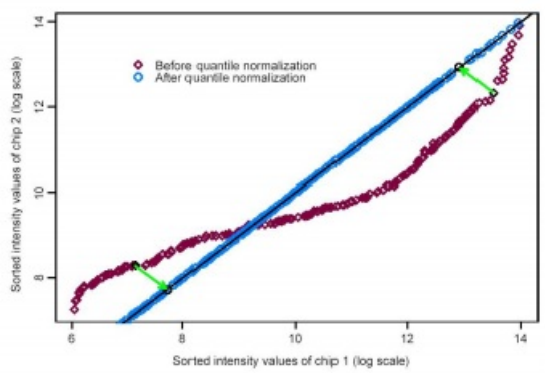
Korekcja tła, czy inaczej eliminacja sygnału tła odbywa się tu na podstawie globalnego modelu rozkładu wartości intensywności sond PM (perfect match), ignorując sondy MM

(mismatch). Wartość sygnału O składa się z sumy składnika tła N (o rozkładzie według krzywej Gaussa, średniej μ oraz odchyleniu standardowym σ) oraz składnika sygnału S (o średniej α i wykorzystaniu funkcji rozkładu normalnego i gęstości), czyli:

$$O = N + S, N \sim N(\mu, \sigma^2), S \sim Exp(\alpha) \tag{3.1}$$

Parametry μ, σ^2,α dla wszystkich sond PM są równe. Ujemne wartości rozkładu normalnego N są pomijane.[1]

Kolejnym etapem jest normalizacja kwantylowa, polegająca na przekształceniu danych tak, by było możliwe porównanie wartości pomiędzy różnymi mikromacierzami, wykorzystywanymi w danym eksperymencie. Jest to dość szybka metoda, oparta na niewielu założeniach. Bazuje na wykresie QQ-plot, rysowanym na podstawie posortowanych wartości intensywności jednej mikromacierzy na osi X oraz innej mikromacierzy na osi Y. Istotne jest tutaj, aby został zachowany rozkład intensywności sygnałów na każdej płytce mikromacierzowej. Zmian dokonuje się na podstawie empirycznego rozkładu intensywności na mikromacierzy, a także rozkładu uśrednionych odległości między tymi sygnałami (tzw. kwantyli).



Rysunek 3.2: Wartości intensywności dwóch mikromacierzy przed(czerwony wykres) i po normalizacji kwantylowej (niebieski wykres). Zielone strzałki pokazują funkcje normalizacji. Każdy punkt jest rzutowany na przekątnej, co jest równoważne wymianie każdego elementu wektora na średnią wektora. Źródło: [7] (Figure 3.3)

Na koniec wykonywana jest sumaryzacja polegająca na sumowaniu wartości intensywności sygnałów, pochodzących od zestawu sond przyporządkowanych każdemu z transkryptów, w celu przypisania każdemu genowi wartości ekspresji. Sumaryzacje można podzielić na dwa typy: obejmujące wiele mikromacierzy lub jedną mikromacierz. W obrębie jednej mikromacierzy wykorzystywana są m.in. średnia arytmetyczna logarytmów o podstawie 2 z wartości sygnałów dla zestawu sond, logarytm naturalny średniej wartości sygnałów, mediana wartości przedstawionych w skali logarytmicznej, logarytm naturalny mediany wyznaczonej dla zestawu sygnałów. W metodach obejmujących zestaw mikromacierzy wykorzystuje się głównie tworzenie modeli opartych na skali logarytmicznej liniowej lub na tzw. algorytmie „wygładzania" mediany. Właśnie ten ostatni algorytm wykorzystywany jest w przypadku RMA.

W rzeczywistości jest jeszcze wiele innych sposobów normalizacji które obejmują korekcje tła, normalizacje i sumaryzacje danych. Niektóre z nich korzystają z tych samych metod

[1] Źródło: [7] (wzór (3.1))

3.2. Analizy wyższego rzędu

Analiza wyższego rzędu umożliwia znalezienie biologicznych zależności wśród dużej ilości danych, otrzymanych po analizie niższego rzędu. Najczęściej opiera się ona na grupowaniu lub klasyfikacji genów, czy próbek w zespoły wykazujące wspólne cechy, na przykład podobny profil ekspresji, regulację przez ten sam czynnik, związek z procesem chorobowym.

Ważnym etapem analizy wyższego rzędu jest filtracja, czyli usunięcie z dalszych analiz genów, których poziom ekspresji lub jego zmiana zostanie uznana za zbyt niską. Upraszcza to analizę, ze względu na zmniejszenie ilości analizowanych danych, ale też zabezpiecza przed nadmiarem wyników fałszywie pozytywnych.

W celu wykonania filtracji wprowadza się progi zmiany poziomu ekspresji. Najczęściej wartości powyżej 1,75 lub 2 (co oznacza wzrost wartości ekspresji 1,75 lub 2 - krotnie) oraz poniżej 0,5 lub 0,75 (spadek ekspresji o 0,5 lub 0,75) są uważane za istotne, resztę sond odrzuca się z dalszych analiz.

Często dane poddaje się też różnego rodzaju testom statystycznym, aby zmniejszyć ilość błędów typu I oraz typu II.

Zamykającym elementem każdej analizy wyższego rzędu, powinna być konfrontacja uzyskanych wyników z informacjami zawartymi w internetowych bazach danych, w celu weryfikacji, czy otrzymane zależności mają swoje potwierdzenie na podłożu biologicznym i nie są jedynie sztucznym wytworem narzędzi matematyczno–statystycznych.

Rozdział 4

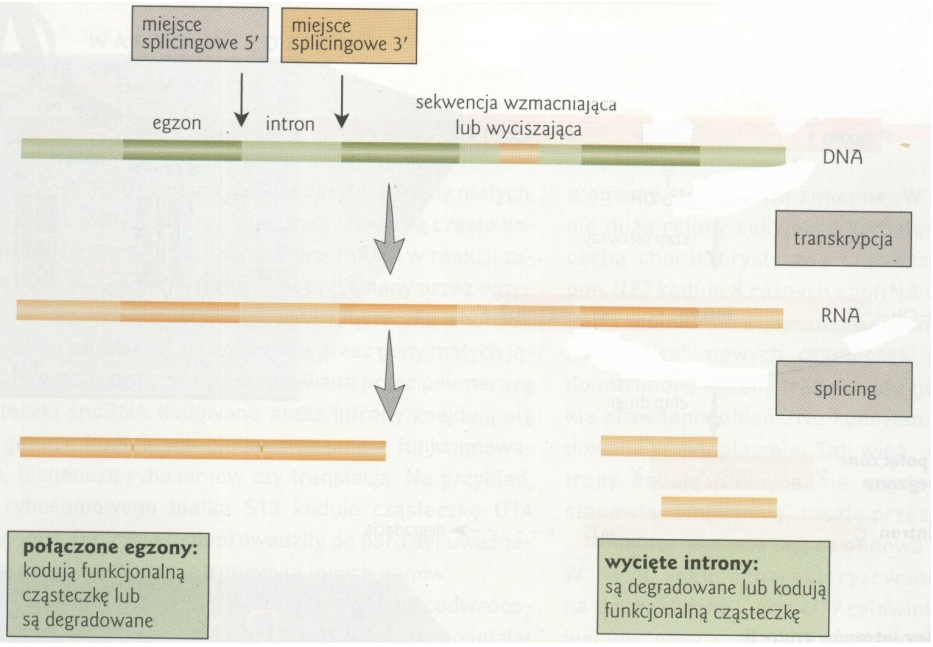
Splicing RNA

Na początku zostanie przedstawione kilka definicji [10], które ułatwią zrozumienie dalszej części pracy.

Intronami początkowo nazywano fragmenty genu o długości od mniej niż 100pz do nawet tysięcy par zasad, nie przenoszące żadnej informacji biologicznej, oddzielające egzony od siebie. Z kolei egzonami nazywano krótkie fragmenty genu o długości od 50 do 300 pz, zawierające informacje genetyczną. Uważano je za sekwencje funkcjonalne, ulegające ekspresji, wchodzące w skład dojrzałej cząsteczki RNA.

Jednak całkiem niedawno odkryto, że introny nie są to sekwencje nie niosące żadnej istotnej informacji biologicznej. Zauważono, że mogą one zawierać elementy regulatorowe transkrypcji, takie jak sekwencje wzmacniające, czy wyciszające, a także kodować cząsteczki małych RNA.

Odkryto także geny „odwrócone na drugą stronę", w których introny kodują sekwencje funk-



Rysunek 4.1: Splicing. DNA zostaje przepisane na RNA, by następnie poddać się procesowi splicingu. Introny są tutaj zaznaczone na jasnopomarańczowo, a egzony na ciemnopomarańczowo. Egzony są łączone razem, dzięki czemu mogą kodować funkcjonalną cząsteczkę. Z kolei introny na ogół są degradowane. Źródło: [10]

cjonalne, a egzony są usuwane. Dlatego zaczęto używać nowych definicji - introny definiuje się jako sekwencje, które zostały rozdzielone po wycięciu, a egzony jako sekwencje, które po wycięciu są złączane.

Teraz opisane zostanie meritum tego rozdziału, czyli splicing. Splicing, inaczej składanie genu lub wycinanie intronów jest to proces biologiczny, polegający na wycięciu w ściśle określonych miejscach splicingowych z prekursorowego RNA organizmów eukariotycznych, sekwencji niekodujących - intronów. Zachodzi podczas obróbki posttranskrypcyjnej po to, by dojrzały mRNA, przygotowany do translacji, kodował ciągły łańcuch polipeptydowy (od kodonu start do stop). Końce RNA, pozostałe po tym wycięciu, są łączone i tworzą ciągłą nić mRNA (RNA informacyjne), rRNA (RNA rybosomowe) lub tRNA (RNA transportujące).

Miejsca wycinania intronów, a następnie łączenia egzonów są wyznaczane przez specjalne sekwencje, nazywane miejscami splicingowymi. Miejsce splicingowe 5' oznacza połączenie egzon-intron na końcu 5' intronu. Na drugim końcu tego intronu miejsce splicingowe 3' wyznacza połączenie z następnym egzonem.

Ponieważ znane są różne mechanizmy splicingu, wyróżnia się pięć klas intronów. Są to: autokatalitycznie wycinające się introny grupy I i II, introny tRNA, introny archebakteryjne oraz introny spliceosomowe, występujące w cząsteczkach jądrowych pre-mRNA. Zaburzenia zachodzące w procesie splicingu mogą być przyczyną różnych chorób.

4.1. Mechanizmy splicingowe

Opisane zostaną teraz krótko wybrane mechanizmy splicingowe. [10] Pierwszym z nich będą autokatalitycznie wycinjące się introny z grupy I. Są to duże, katalityczne cząstki RNA, zachowujące się jak elementy ruchome, mogące włączać się w sekwencje genów bezintronowych. Introny z tej grupy często występują w genomach mitochondrialnych, chloroplastowych oraz jądrowych różnych eukariontów. Zdecydowana większość z nich została odnaleziona u grzybów, roślin i glonów, a u zwierząt jedynie w mitochondriach. Bardzo rzadko można je spotać u bakterii, czy wirusów bądź bakteriofagów.

Autokatalityczne wycinanie się intronów z tej grupy zachodzi w dwóch etapach. Aby proces ten mógł zajść, jako kofaktor wymagany jest zewnętrzny nukleotyd guanozynowy (G). Trzeciorzędowa struktura RNA intronu wygląda jak zwinięty stos przylegających do siebie bokami dwuniciowych helis. Sukces reakcji katalitycznej jest zależny od prawidłowego zwinięcia się intronu.

Introny grupy II podobnie jak grupy I są dużymi, katalitycznymi cząstkami RNA i tak samo potrafią włączać się w sekwencje bezintronowe. Występują znacznie rzadziej niż poprzednio opisywane introny. Są obecne w genomach mitochondrialnch i chloroplastowch niektórych pierwotniaków, grzybów, glonów oraz w DNA bakteryjnym.

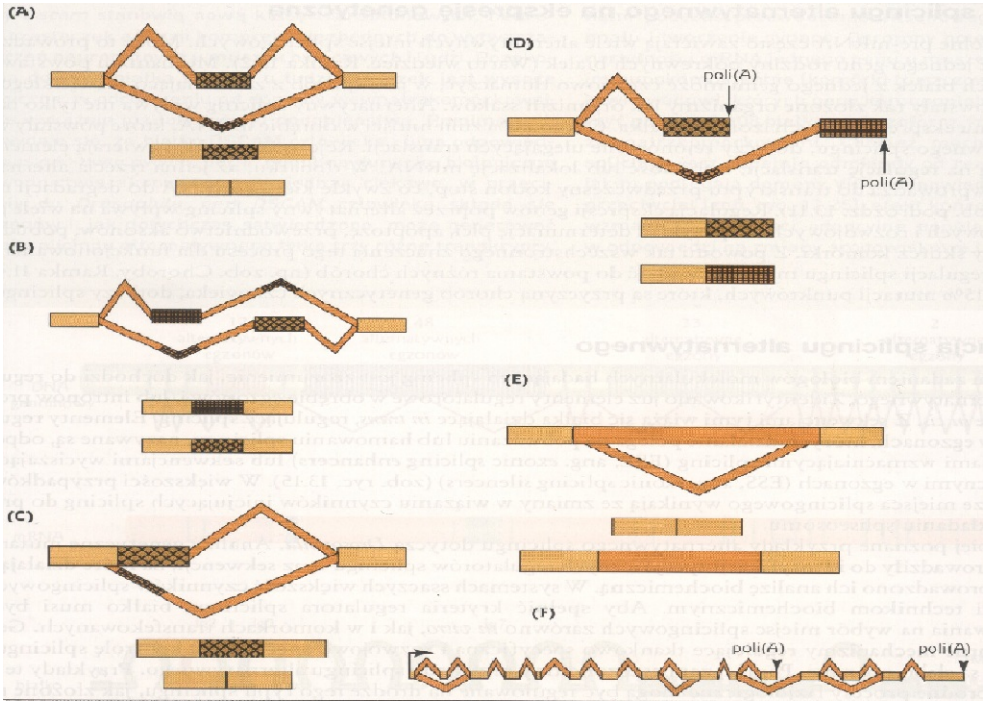
Autokatalityczne wycinanie w tym przypadku, przypomina mechanizm splicingu pre-mRNA z wykorzystaniem spiliceosomu. Drugorzędowa struktura intronu przypomina kwiat z 6 helikalnymi domenami odchodzącymi promieniście od centralnego koła. Podczas splicingu zachodzą dwie reakcje transestryfikacji, czyli procesu polegającego na otrzymaniu estrów przez reakcję chemiczną innych estrów z alkoholami, kwasami lub innymi estrami.

Tak więc głównymi cechami odróżniającymi te dwa mechanizmy splicingowe od innych jest ich zdolność do autokatalitycznego wycinania się oraz możliwość przemieszczania się w obrębie genomu. Dzięki temu mogą one między innymi efektywnie zasiedlać allele homologiczne, nie posiadające intronów.

Introny, przy których wycinaniu bierze udział spliceosom, stanowią największy udział pośród wszystkich intronów. Występują przede wszystkim w jądrowym pre-mRNA wielokomórkowych organizmów eukariotycznych.

Mechanizm wycinania jest identyczny jak w przypadku intronów autokatalitycznie wycinających się z grupy II, jednak introny pre-mRNA nie zwijają się w zachowawcze ewolucyjnie struktury drugo- czy trzeciorzędowe. W tej sytuacji, trzeciorzędowa struktura, umożliwiająca przeprowadzanie splicingu, powstaje w obrębie spliceosomu, czyli wielkiego kompleksu RNA i białek, który ma naturę dynamiczną. Jego skład zmienia się w trakcie procesu splicingu - w zależności od katalizowanej reakcji niektóre czynniki są usuwane, a inne dodawane.

4.2. Splicing alternatywny



Rysunek 4.2: Wzory splicingu alternatywnego. Na jasnopomarańczowo zaznaczone są egzony konstytutywne. Zakreskowane prostokąty są to egzony alternatywne. Ciemnopomarańczowe linie pokazują sposób wycinania intronów. (A) egzon może zostać włączony do mRNA lub wycięty, (B) sytuacja gdzie tylko jeden z dwóch, sąsiadujących egzonów może być włączony w danym momencie do mRNA, (C) alternatywne miejsca splicingowe 3' lub 5', (D) alternatywne miejsce poliA zmienia egzony na końcu 3', (E) intron może być wycięty lub pozostawiony w mRNA, (F) pre-mRNA może zawierać wiele miejsc splicingu alternatywnego. Źródło: [10]

Splicing alternatywny [10] [12] jest to proces zachodzący w trakcie splicingu pre-mRNA. Polega na pomijaniu niektórych egzonów lub pozostawieniu niektórych intronów, co za tym idzie łączeniu egzonów ze sobą na różne sposoby, nie koniecznie w kolejności ich występowania w sekwencji. Zdecydowana większość egzonów jest określana mianem konstytutywnych, czyli takich, które zawsze są włączane do dojrzałego mRNA. Jednak występują też takie egzony, których splicing jest regulowany, nazywane egzonami alternatywnymi. Oznacza to, że

mogą być one włączane lub wycinane z cząsteczki mRNA. Zdarza się też tak, że występują alternatywne miejsca splicingowe 5' lub 3', które wydłużają lub skracają określony egzon. Dzięki temu, splicing alternatywny umożliwia uzyskanie z jednego genu więcej niż tylko jednej cząstki mRNA, zwiększając tym samym zmienność białek, a także dostarcza wiele możliwości dotyczących regulacji ekspresji genów.

4.3. Splicing index

Splicing index (SI) [9] jest to wartość wyliczana dla konkretnego egzonu w genie. Umożliwia porównanie ekspresji egzonu między grupami. Dzięki tej wartości można zidentyfikować i zbadać jak różnią się warianty splicingowe wśród osobników z jednego gatunku,wykazujących odmienne fenotypy w tej samej sytuacji, czy dana reakcja organizmu jest wynikiem różnego rozłożenia miejsc splicingowych w genie, czy może nie ma z tym nic wspólnego. Wyliczenie splicing indexu dla każdego egzonu, który wykazał ekspresje na mikromacierzy zmniejsza nam ilość danych w dalszej analizie, pozostawiając do kolejnych badań jedynie te egzony, których różnice w ekspresji są rzeczywiście istotne, czyli te gdzie moduł ze splicing index w grupach jest największy. Skraca to czas potrzebny na przeprowadzenie następnych etapów analizy.

Metoda wyliczania tej wartości jest stosunkowo prosta. Jest to logarytm ze stosunku N_{p1} i N_{p2}, gdzie N_{p1} jest to średnia ze wszystkich N_{pcr} w jednej grupie, a N_{p2} jest średnią wszystkich N_{pcr} z grupy drugiej. Wzór [1] ten ma postać:

$$SI = log(\frac{N_{p1}}{N_{p2}}) \tag{4.1}$$

Npcr z kolei, wyliczamy z poniższego wzoru [2]:

$$N_{pcr} = \frac{P_{cr}}{G_{cr}} \tag{4.2}$$

gdzie P_{cr} jest definiowane jako poziom ekspresji danego egzonu dla konkretnej mikromacierzy, a G_{cr} jako średnia wartość ekspresji wszystkich egzonów w genie dla jednej mikromacierzy, nazywana poziomem genu. Indeks dolny p oznacza konkretny zestaw sond, g - gen, c - grupę, a r - kolejną replikę. Przy wyliczaniu wielkości G_{cr} zamiast średniej ze wszystkich egzonicznych próbek, można używać mediany. Sposób wyliczania, z którego się korzysta, zależy już od indywidualnych upodobań.

Jak widać wyliczanie tej wartości z punktu widzenia matematyki nie jest trudne i nie wymaga wykonywania skomplikowanych obliczeń. Dlatego też, implementacja w języku R funkcji, która liczyłaby splicing index dla zadanego przez użytkownika zbioru danych jest stosunkowo prosta, a to z kolei w dużej mierze, ułatwia i przyspiesza prace nad dużymi zbiorami danych. Jest to istotne, ponieważ przy tak dużej ilości danych otrzymanych z różnych mikromacierzy, ważne jest minimalizowanie czasu, który należy poświęcić na przeprowadzenie poszczególnych etapów analizy.

[1] Źródło: [9] (wzór (1))
[2] Źródło: [9] (wzór (2))

Rozdział 5

Wyniki i wnioski

Przeprowadzana w tej pracy analiza opierała się na danych uzyskanych z eksperymentu mikromacierzowego z użyciem 6 mikromacierzy. Trzy pierwsze zestawy (K_06, K_07, K_08) są to dane z próbek pobranych od *Arabidopsis thaliana* typu dzikiego (wt), czyli bez zmian w genotypie. Kolejne (K_13, K_14, K_15) pochodzą od organizmów knock out, czyli z mutacją w dowolnym genie. (w tym przypadku jest to mutacja w białku histonowym). Wszystkie wykonywane etapy analizy oparte są o język R - gotowe moduły udostępnianie na serwisach internetowych oraz własne funkcje zaimplementowane w tym języku. [16][17][18]

Na początku wykonana została wstępna obróbka danych za pomocą normalizacji RMA. [19] Dzięki korekcji tła, normalizacji i sumowaniu wchodzącym w skład RMA, których działanie opisane zostało we wcześniejszych rozdziałach pracy, uzyskano 116199 próbek dla każdej mikromacierzy, na których można było przeprowadzić analizę wyższego rzędu.

	K_06	K_07	K_08	K_13	K_14	K_15
AT1G03190.2.Chr1.plus.779538.779646	4.609612	4.740236	3.980820	4.172943	4.250951	4.266472
AT1G03190.2.Chr1.plus.779744.780062	5.563554	5.946954	5.689601	5.158247	5.730096	6.302240
AT1G03210.1.Chr1.plus.782948.783175	7.919246	7.845558	8.066065	7.787577	7.274504	7.412121
AT1G03210.1.Chr1.plus.783318.783596	7.624561	7.437945	7.814272	7.230866	7.261865	7.079762
AT1G03210.1.Chr1.plus.783681.783842	7.286983	7.181449	7.060529	6.776125	7.166030	6.838110
AT1G03220.1.Chr1.plus.787122.788651	8.142176	8.267345	8.167486	7.883901	8.173599	7.979921
AT1G03230.1.Chr1.plus.790091.791592	6.695250	6.620779	6.608811	6.842991	6.584787	6.574833
AT1G03270.1.Chr1.plus.799191.799431	6.348637	6.168892	6.357696	6.381046	6.237861	6.093988
AT1G03270.1.Chr1.plus.799999.800112	4.291558	5.219265	4.007244	4.534132	4.009779	4.701920
AT1G03270.1.Chr1.plus.801517.801889	6.061904	5.834056	5.454441	6.244207	5.973181	5.218552
AT1G03270.1.Chr1.plus.802320.802436	2.880140	5.806187	4.569058	5.208401	5.374161	5.315301

Rysunek 5.1: Przykładowe dane po normalizacji RMA. Źródło: Opracowanie własne na podstawie [19]

Rysunek 5.1 przedstawia przykładowe wyniki otrzymane po normalizacji. Pierwsza kolumna odpowiada nazwie sondy na mikromacierzy. Zawiera kilka informacji, które mogą być przydatne w dalszych analizach. Składa się kolejno z oddzielonych kropkami: informacji o ge-

nie, chromosomie, nici, pozycji na której zaczyna się i kończy dany egzon.[4] Kolejne kolumny to wyliczone wartości ekspresji dla poszczególnych płytek mikromacierzowych.

	sonda	dlugosc
AT5G28263.1.Chr5.plus.10241337.10256531		15194
AT1G43060.1.Chr1.plus.16189210.16203832		14622
AT5G30269.1.Chr5.plus.11596127.11610717		14590
AT1G40101.1.Chr1.minus.15063543.15076496		12953
AT1G41930.1.Chr1.plus.15685411.15695693		10282
AT2G01029.1.Chr2.plus.28465.38652		10187
AT5G36935.1.Chr5.plus.14574600.14584700		10100
AT3G60965.1.Chr3.minus.22541491.22551393		9902
AT5G32345.1.Chr5.plus.11977548.11987264		9716
AT1G47860.1.Chr1.minus.17620677.17630270		9593

Rysunek 5.2: Lista najdłuższych egzonów. Źródło: Opracowanie własne

Dzięki takiemu nazewnictwu sond, jakie zostało uzyskane po RMA można dowiedzieć się jakiej długości są najkrótsze, czy najdłuższe egzony oraz w jakim miejscu na chromosomie występuje dany egzon. Taka prosta analiza, wzbogaca naszą wiedzę o informacje, które mogą w późniejszych badaniach okazać się przydatne. Na rysunkach 5.2 oraz 5.3 pokazana jest lista 10 odpowiednio najdłuższych i najkrótszych egzonów z analizowanego w tej pracy zbioru danych. Średnia długość odcinka kodującego u *Arabidopsis thaliana* wynosi ok. 304 bp [13]. Patrząc na wyniki przeprowadzonej analizy widać, że zdarzają się egzony znacznie dłuższe. Kolejną informacją, którą możemy uzyskać, jest to z ilu egzonów składa się każdy gen *Arabidopsis thaliana*. Na rysunku 5.4 przedstawione jest 10 genów złożonych z największej liczby egzonów. Aby uzyskać, pojęcie na temat tego, czy przedstawione na rysunku 5.4 wartości są istotnie większe od przeciętnej ilości egzonów w innych genach tego gatunku, czy zbliżone , wyliczono medianę oraz średnią z listy przedstawiającej liczbę egzonów w każdym genie. Okazuje się, że mediana jest równa 2, a średnia 3, tak więc wyraźnie widać, że wybrane geny składają się ze znacznie większej ilości egzonów, niż przeciętny gen *Arabidopsis thaliana*. Takie informacje, które są stosunkowo łatwe do wyliczenia, dają nam lepsze pojęcie na temat genomu badanego organizmu.

	sonda	dlugosc
AT1G45160.1.Chr1.minus.17085065.17085153		88
AT1G45160.2.Chr1.minus.17085065.17085153		88
AT2G23093.1.Chr2.plus.9833369.9833457		88
AT2G34010.1.Chr2.plus.14369562.14369650		88
AT2G22955.1.Chr2.minus.9771465.9771553		88
AT3G63400.1.Chr3.plus.23412770.23412858		88
AT3G63400.2.Chr3.plus.23412770.23412858		88
AT3G15040.1.Chr3.minus.5066602.5066690		88
AT4G16650.1.Chr4.plus.9373707.9373795		88
AT4G36920.1.Chr4.plus.17402340.17402428		88

Rysunek 5.3: Lista najkrótszych egzonów. Źródło: Opracowanie własne

Następnie porównano między sobą wartości ekspresji egzonów z każdej mikromacierzy, po to by odnaleźć te sondy, których wartości znacznie się różnią. Jako próg przyjęto co najmniej 2-krotną różnice między eksperymentami (mikromacierzami). Tak więc, wykonano łącznie 6 porównań, których wyniki przedstawiono w tabeli (Rysunek 5.5). Miało to na celu pokazać,

nazwa genu	ilosc egzonow
AT1G67120.1.Chr1.minus	53
AT1G48090.1.Chr1.minus	48
AT4G17140.2.Chr4.minus	46
AT4G17140.1.Chr4.minus	45
AT1G64790.1.Chr1.minus	40
AT1G02080.1.Chr1.plus	39
AT3G48190.1.Chr3.plus	39
AT1G48090.2.Chr1.minus	38
AT1G50030.1.Chr1.minus	35
AT1G50030.2.Chr1.minus	34

Rysunek 5.4: Lista 10 genów z największą ilością egzonów. Źródło: Opracowanie własne

że mimio iż zrobiono porównania w grupach, gdzie spodziewano się wyników podobnych, nieznacznie różniących się, mogą występować też wyniki które różnią się bardziej znacząco. Przyczyną może być niedoskonałość takiego eksperymentu oraz fakt, że przeprowadzane analizy są narażone na pewien margines błędu, który w przypadku statystyki zawsze występuje i jest nie do uniknięcia. Jednak dzięki dużej ilości danych, nie zaburza to naszych analiz.

Porównanie	Ilość sond o wartościach różniących się co najmniej 2-krotnie
K06-K07	477
K07-K08	315
K06-K08	401
K13-K14	393
K14-K15	557
K13-K15	459

Rysunek 5.5: Źródło: Opracowanie własne

Aby wyszukać różne warianty splicingowe u *Arabidopsis thaliana* niezbędna jest identyfikacja tych egzonów, których wartość ekspresji istotnie różni się między grupami. Mogłoby się wydawać, że aby to zrobić można jedynie wyliczyć średnią z wartości ekspresji dla każdej mikromacierzy w grupie, a następnie tak uśrednione wartości porównać i uznać za istotne te, które różnią się o zadaną przez nas wartość. Po wykonaniu takiej analizy, z założeniem, że istotne są tylko te egzony, których uśrednione wartości ekspresji między grupami różnią się co najmniej 2 - krotnie, uzyskano listę jedynie 10 takich sond.

Można więc sądzić, że jest to dobra metoda, która doskonale zawęża zbiór danych przeznaczonych do dalszej analizy. Jednak sposób ten, ma jedną podstawową wadę, która wyklucza jego poprawność. Przy tak przeprowadzanej analizie, nie uwzględniany jest fakt, że zdecydowanie wyższy poziom ekspresji może mieć cały gen, co za tym idzie, rośnie poziom ekspresji każdego konstytutywnego egzonu w tym genie. Na przykład, jeżeli gen A charakteryzuje się dwukrotnie większa ekspresją niż gen B, to każdy egzon genu A może mieć 2 - krotnie wyższy poziom ekspresji od egzonów genu B.

Widzimy, że zwykłe porównanie, nie jest poprawne i daje nam fałszywe pojęcie o różnicy w ekspresji danego egzonu. W związku z czym została wymyślona inna wartość, umożliwiająca przeprowadzenie takiej analizy, ale już uwzględniająca poziom ekspresji genu, a nie tylko poszczególnego egzonu. Jest to już opisany wcześniejszym rozdziale tej pracy splicing index

"AFFX-r2-At-U12639_st"	"3.65106771783015"	"8.3960689523365"
"AT2G20170.1.Chr2.plus.8701013.8701168"	"3.01663686740974"	"1.49447770096815"
"AT2G20170.2.Chr2.plus.8701013.8701168"	"3.01663686740974"	"1.49447770096815"
"AT2G25470.1.Chr2.plus.10838861.10839016"	"3.57747383794729"	"1.71865314170504"
"AT2G41970.1.Chr2.minus.17522546.17522756"	"1.42996916452451"	"3.0001294156447"
"AT3G18240.1.Chr3.plus.6257792.6258090"	"1.65690717370481"	"3.45747388484944"
"AT3G42190.1.Chr3.plus.14370582.14370690"	"0.979624763974313"	"2.05203051907"
"AT3G14185.1.Chr3.minus.4709272.4709475"	"3.53490843484151"	"1.46480985997558"
"AT4G21680.1.Chr4.minus.11519459.11519765"	"2.18228705656032"	"4.42143229102499"
"AT5G64541.1.Chr5.plus.25801314.25801451"	"2.67120236504384"	"1.3263468365116"

Rysunek 5.6: Sondy, których uśrednione wartości ekspresji w grupach różnią się co najmniej 2 - krotnie. Źródło: Opracowanie własne

(Kod funkcji w języku R przedstawiony w: Dodatek A).

Dlatego też, kolejnym krokiem wykonanym w tej analizie, było wyliczenie splicing indexu dla każdego egzonu. [9] Aby to zrobić, trzeba było pogrupować wszystkie egzony należące do jednego genu, by móc wyliczyć średnią wartość ekspresji genu. Następnie, gdy były już każdemu genowi przypisane egzony, można było przejść do meritum i policzyć splicing index dla danych z tej pracy, wybierając następnie kilka o największym module wartości tego parametru. W przypadku gdy wartość SI jest równa 0, oznacza to że dany egzon zachowuje się w obu grupach tak samo, dodatnie wartości SI informują o wyższym poziomie w grupie 1 i analogicznie, ujemne wartości oznaczają wyższą ekspresje w grupie 2.[8]

	SI
AT2G20170.1.Chr2.plus.8701013.8701168	1.895653
AT2G20170.2.Chr2.plus.8701013.8701168	1.891788
AT2G25470.1.Chr2.plus.10838861.10839016	1.855773
AT2G22950.1.Chr2.plus.9769527.9769766	1.771466
AT4G27550.1.Chr4.plus.13759507.13759740	1.746176
AT2G05410.1.Chr2.plus.1977186.1977383	1.734821
AT5G37478.1.Chr5.plus.14883679.14883989	1.695214
AT4G33790.1.Chr4.minus.16204589.16204716	1.691532
AT4G34060.1.Chr4.plus.16315532.16315852	1.681012
AT4G34060.2.Chr4.plus.16315532.16315852	1.671960

Rysunek 5.7: Lista egzonów o największym module z wartości splicing indexu liczonego z wykorzystaniem średniej. (Kod funkcji w języku R przedstawiony w: Dodatek A) Źródło: Opracowanie własne

Na rysunkach 5.7 i 5.8 widać posortowaną listę 10 egzonów, których moduł ze splicing indexu liczonego dwiema różnymi metodami jest największy. Rys 5.7 ukazuje wyniki dla sposobu opartego o średnią, a rys 5.8 o medianę, oba przedstawiają podobne listy egzonów. Widać, że 6 z nich zostało zidentyfikowane przez oba przypadki (rys 5.9), jednak różna jest ich kolejność na liście. Wartości ekspresji tych egzonów przedstawiono na wykresach, aby móc porównać, które egzony w danym genie, wykazują największą różnice w ekspresji między grupami osobników wild type i knock out. Poniżej zamieszczono 6 wykresów dla każdego genu, w którym zidentyfikowano egzon o dużej wartości modułu spicing index.

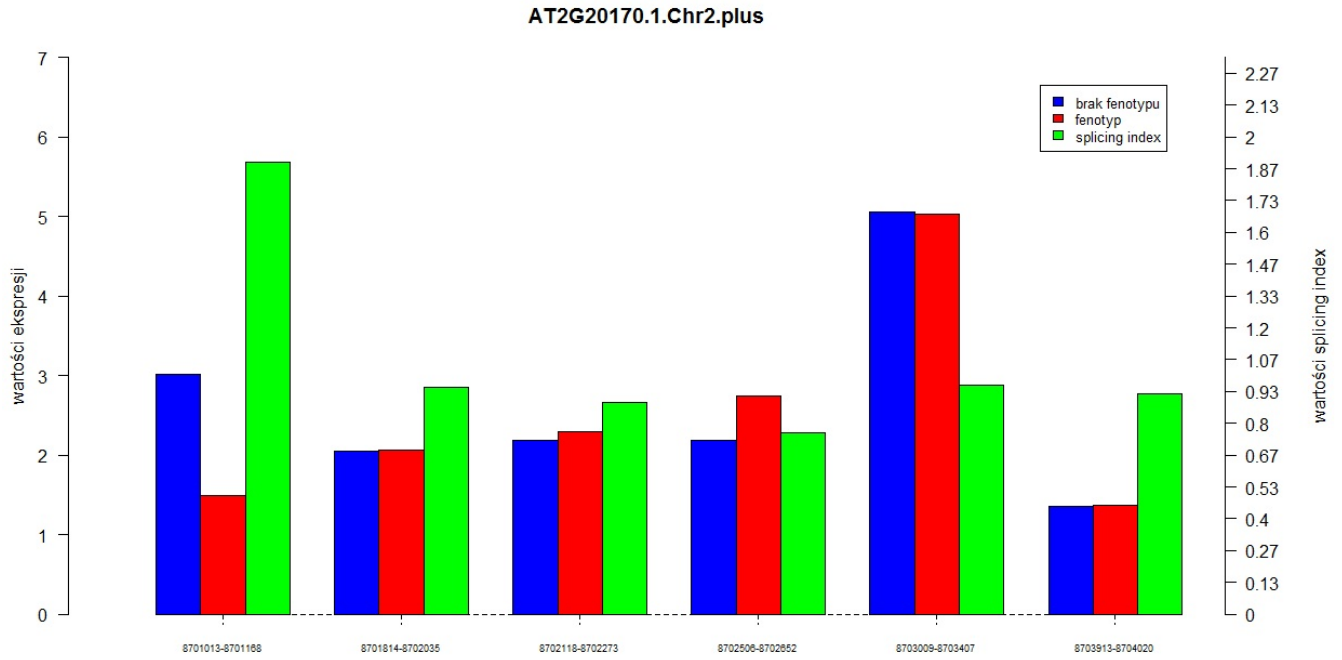
Widzimy na wykresach jak zachowują się egzony w poszczególnych genach, które zawierają chociaż jeden z egzonów pochodzący z listy na rysunku 5.9. Dzięki temu możemy zidentyfikować geny, które są dla nas interesujące i zacząć od nich analizy biologiczne. Na przykład jeżeli, któryś gen będzie składał się z więcej niż jednego egzonu o wysokim module

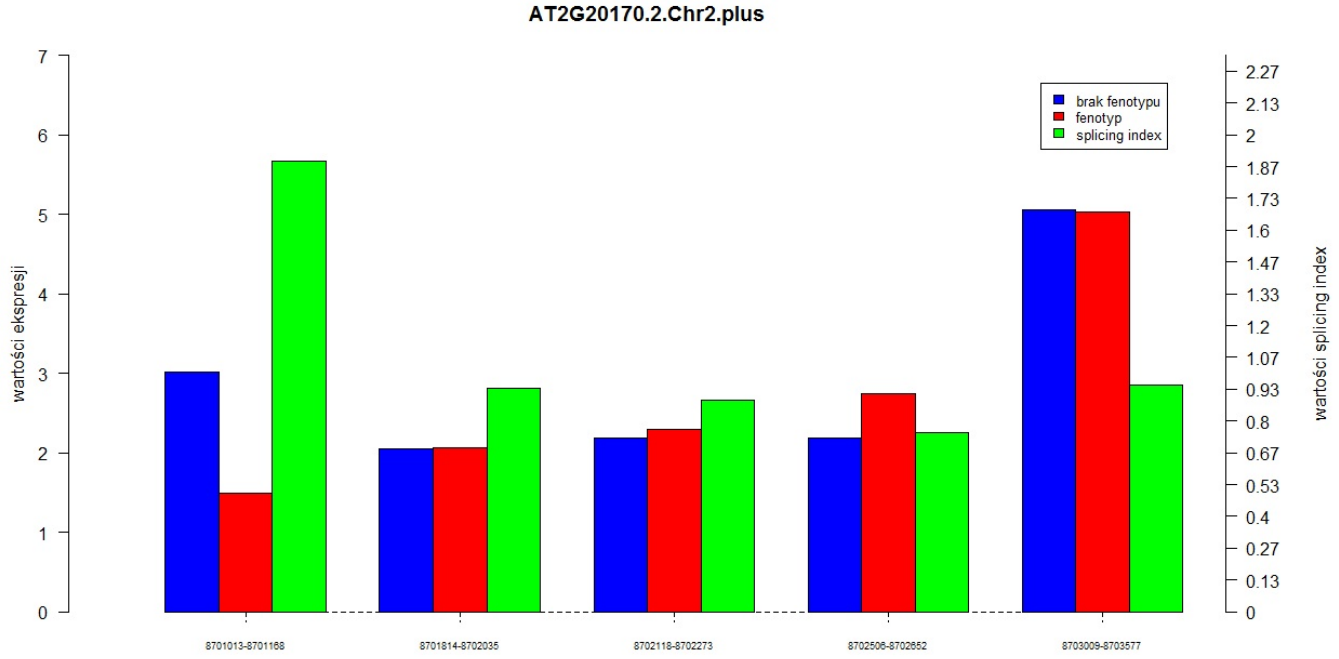
```
                                                  SI
AT2G25470.1.Chr2.plus.10838861.10839016 2.092859
AT2G20170.2.Chr2.plus.8701013.8701168   1.917317
AT5G37478.1.Chr5.plus.14883679.14883989 1.908000
AT2G20170.1.Chr2.plus.8701013.8701168   1.857309
AT2G22950.1.Chr2.plus.9769527.9769766   1.837343
AT1G78160.1.Chr1.plus.29410080.29410184 1.820427
AT5G20450.1.Chr5.minus.6912487.6912654  1.800309
AT4G27550.1.Chr4.plus.13759507.13759740 1.776588
AT1G07540.1.Chr1.minus.2319874.2320038  1.772851
AT5G43110.1.Chr5.plus.17310810.17310975 1.761235
```

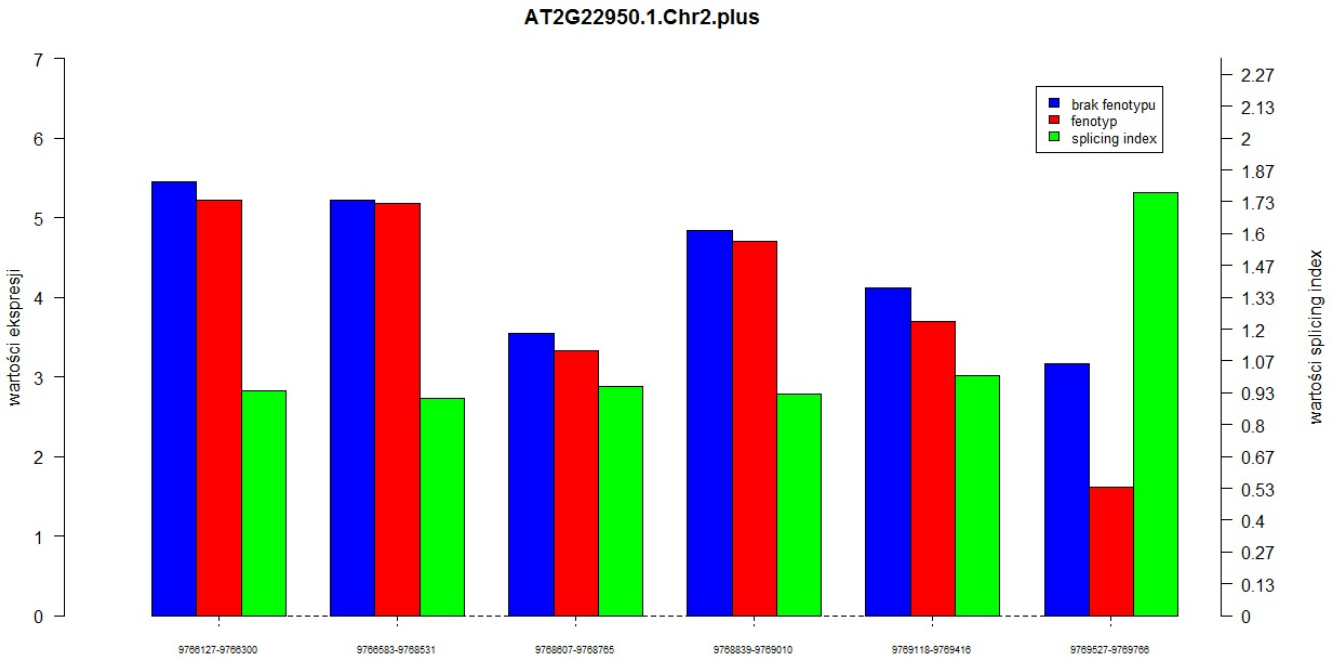
Rysunek 5.8: Lista egzonów o największym module z wartości splicing indexu liczonego z wykorzystaniem mediany. (Dodatek A) Źródło: Opracowanie własne

```
"AT2G20170.1.Chr2.plus.8701013.8701168"
"AT2G20170.2.Chr2.plus.8701013.8701168"
"AT2G22950.1.Chr2.plus.9769527.9769766"
"AT2G25470.1.Chr2.plus.10838861.10839016"
"AT4G27550.1.Chr4.plus.13759507.13759740"
"AT5G37478.1.Chr5.plus.14883679.14883989"
```

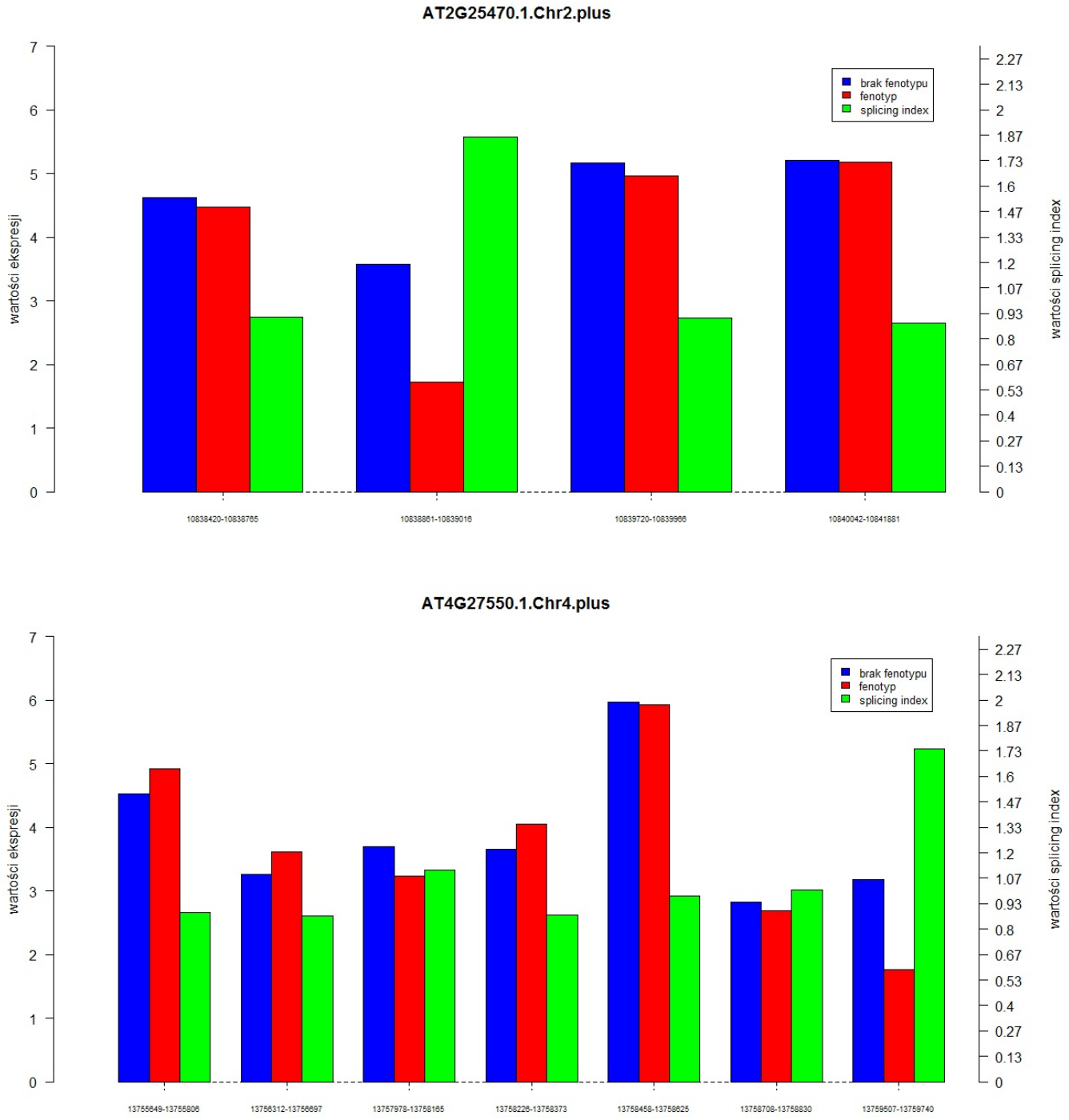
Rysunek 5.9: Lista egzonów zidentyfikowanych przez obie metody wyliczające splicing index. Źródło: Opracowanie własne



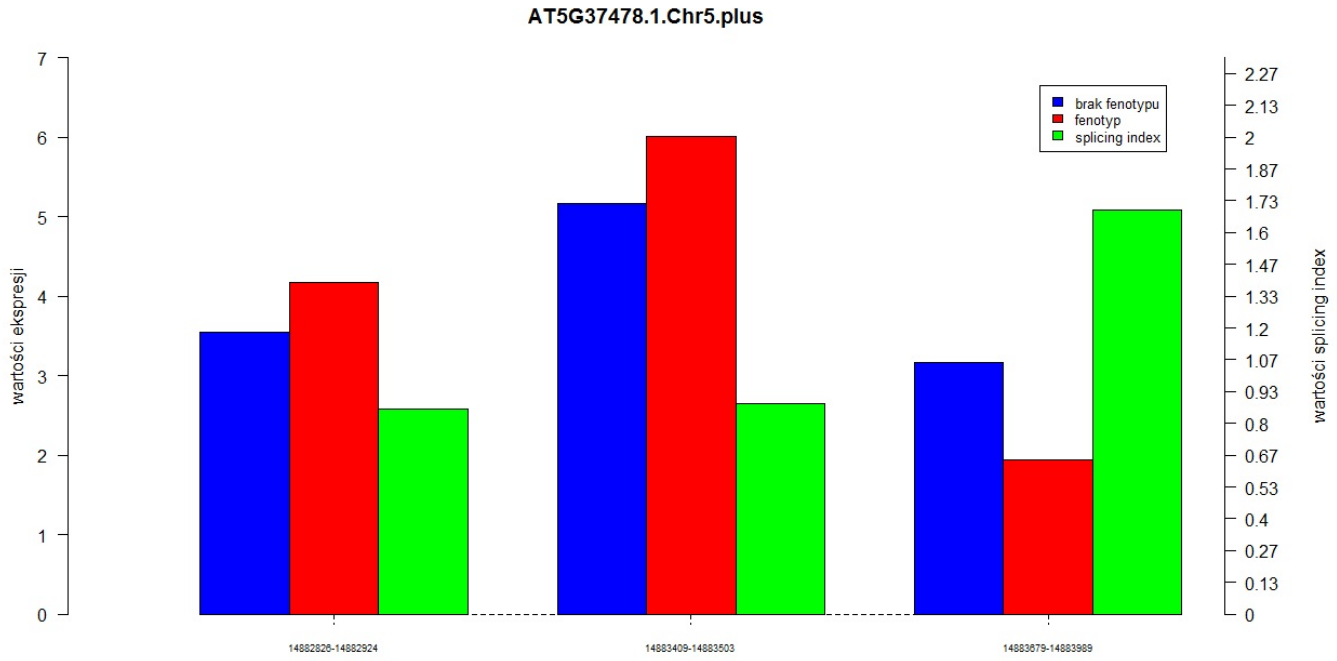




splicing index, bardziej skupimy na nim swoją uwagę podczas badań biologicznych, niż na genie posiadającym jedynie egzony o niskiej wartości modułu splicing index, gdyż może on mieć duże znaczenie w procesach biologicznych badanego gatunku. Można zauważyć, że im większa jest różnica w ekspresji między grupami, tym bardziej rośnie moduł z wartości splicing index tego egzonu. W większości przypadków wartości splicing index dla poszczególnych egzonów są bardzo małe, co oznacza że egzon nie zmienił swojej aktywności i występują tylko nieznaczne różnice w ekspresji między grupami.



Dane otrzymane przez obie metody, podzielono przez ich średnią wartość, aby je znormalizować, a następnie policzono wariancje. Miało to na celu pokazanie, w której metodzie jest większe zróżnicowanie. Otrzymano dwie wartości, 0.0029 dla danych wyliczanych za pomocą średniej oraz 0.0036 dla wyliczanych z użyciem mediany. Widać, że wariancja w obu przypadkach jest niska i różniąca się jedynie nieznacznie między metodami. Tak więc dla obu sposobów, zróżnicowanie nie powinno być duże, a wyniki skupiają się wokół średniej, która w obu przypadkach jest w przybliżeniu taka sama i wynosi 1.001.



Rysunek 5.10: Wykresy przedstawiające uśrednione wartości ekspresji egzonów dla obu grup oraz moduły wartości splicing index dla wszystkich egzonów w genach, złożonych z egzonów o dużym module z wartości splicing indexu. Na osi y umieszczonej z lewej strony znajdują się wartości liczbowe ekspresji, z kolei na prawej osi y wartości splcing index. Oś x przedstawia konkretne egzony z których składa się dany gen, z podanymi miejscami ich początku i końca na nici chromosomu. Źródło: Opracowanie własne

Następnie, korzystając z serwisu TAIR [13] sprawdzono, jakie funkcje spełniają u *Arabidopsis thaliana* geny, w których skład wchodzą egzony, zidentyfikowane jako te z największym splicing index, przez oba sposoby wyliczania tej wartości. Informacje na ten temat zawarto w tabelce (Rysunek 5.11)

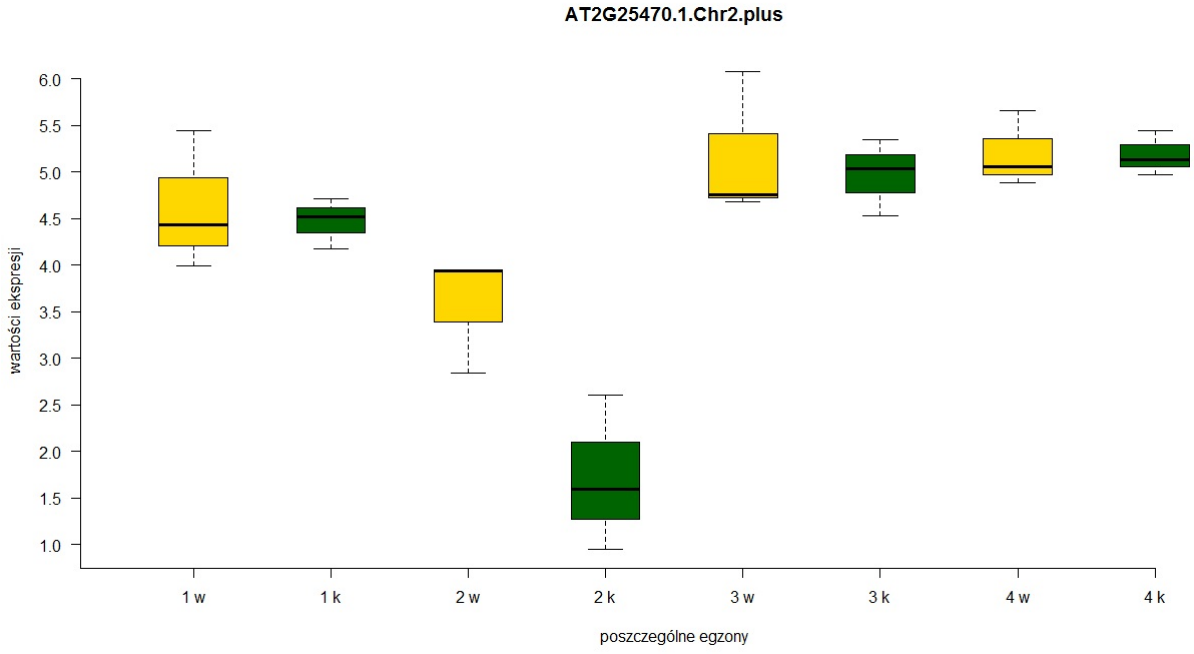
GEN	FUNKCJA
AT5G37478.1.Chr5.plus	Rodzina białek TPX2 (kierowanie białka do XklP2)
AT4G27550.1.Chr4.plus	Koduje enzym biorący udział w biosyntezie trehalozy
AT2G25470.1.Chr2.plus	Transdukcja sygnału
AT2G22950.1.Chr2.plus	Koduje auto-regulowaną Ca2+-ATPaze
AT2G20170.2.Chr2.plus	Funkcja nieznana
AT2G20170.1.Chr2.plus	Funkcja nieznana

Rysunek 5.11: Geny w raz z funkcjami, które spełniają u organizmów gatunku *Arabidopsis thaliana* Źródło: Opracowanie własne. Na podstawie danych z [13]

Po zapoznaniu się z rolami danych genów [13], najbardziej istotny dla naszej analizy, wydaje się być AT2G25470.1.Chr2.plus, odpowiedzialny za transdukcje sygnału. Na wykresach (Rysunek 5.10) pokazane są między innymi, uśrednione po wszystkich eksperymentach, eks-

presje egzonów w tym genie, zarówno dla osobników wt, jak i tych z mutacją. Pokazano także (Rysunek 5.12), zmienność wartości ekspresji poszczególnych sond dla każdego eksperymentu mikromacierzowego. Pozioma kreska znajdująca się wewnątrz prostokąta odpowiada medianie. [21] Między górą, a dołem "skrzynki" zawiera się 50 % obserwacji. "Wąsy" to różnica pierwszego i trzeciego kwartyla. Boxploty pokazują wartości z uwzględnieniem podziału na grupy wt i knock out.

Na Rysunku 5.12 można zaobserwować, że w większości, wartości ekspresji oraz ich zróżnicowanie w grupach jest podobne. Nie ma znaczących różnic, które zwracałyby naszą uwagę. Wyjątkiem są egzony, których splicing index jest dość duży. W tym przypadku widzimy zmiany, zachodzące w wartości ekspresji między grupami. Splicing index wykryje takie różnice i pozwoli nam zidentyfikować takie egzony, jednak nie działa to w obie strony. Nie ma możliwości stwierdzenia, który egzon będzie posiadał wysoką wartość splicing index, jedynie na podstawie wykresów, ukazujących ekspresje. Nie możemy polegać jedynie na wykresach i porównywaniu wartości ekspresji oraz ich zróżnicowania między grupami, pomijając splicing index.



Rysunek 5.12: Zmienność wartości ekspresji egzonów, z których skład się gen AT2G25470.1.Chr2.plus dla każdego przeprowadzonego eksperymentu mikromacierzowego. Na osi x przedstawione są kolejno egzony, które zajmują następujące pozycje na chromosomie: 1 - 10838420.10838765, 2 - 10838861.10839016, 3 - 10839720.10839966, 4 - 10840042.10841881. Litery w i k oznaczają kolejno wild type oraz knock out. Źródło: Opracowanie własne

Egzon AT2G25470.1.Chr2.plus jest przypadkiem, gdzie zarówno dla każdego egzonu, jak i całego genu zmniejsza się wartość ekspresji. Największy spadek, występuje w egzonie, którego splicing index został wykryty jako jeden z 10 największych, w AT2G25470.1.Chr2.plus. 10838861.10839016. Wartość ekspresji zmienia się z 3.58 do 1.72, czyli ze względu na to, że są to logarytmy wartości ekspresji, prawie czterokrotnie. W tabeli (Rysunek 5.13) przedstawione są już uśrednione po wszystkich eksperymentach wartości ekspresji dla poszczególnych egzonów AT2G25470.1.Chr2.plus. Takie zachowanie genu odpowiedzialnego za transdukcje

sygnału może wskazywać, że mutanty których próbki badamy i wykazują fenotyp w sytuacji braku światła są mutantami fitochromowymi i mają pewnie defekty w łańcuchu transdukcji sygnału świetlnego. [20] Może być to efektem braku fitochromu w badanych roślinach. Takie osobniki, charakteryzują się odmienną morfologią oraz anomaliami rozwojowymi w stosunku do osobników wt. Dochodzi do zaburzeń w przebiegu procesów fotomorfogenetycznych, indukowanych przez fitochrom. Następuje ich inicjacja, co w przypadku form dzikich, wymaga stymulacji światłem. Dlatego też, można zaobserwować dobrze rozwinięte liścienie oraz krótkie hypokotyle u siewek roślin. Mutacje wpływają także, na wewnątrzkomórkowy metabolizm rośliny, co wzmaga ekspresje niektórych genów jądrowych.

EGZON	Wt	MUTANT
AT2G25470.1.Chr2.plus.10838420.10838765	4.61	4.46
AT2G25470.1.Chr2.plus.10838861.10839016	3.57	1.71
AT2G25470.1.Chr2.plus.10839720.10839966	5.17	4.96
AT2G25470.1.Chr2.plus.10840042.10841881	5.20	5.18
ŚREDNIA DLA CAŁEGO GENU	4.64	4.08

Rysunek 5.13: Uśrednione wartości ekspresji egzonów, z których skład się gen AT2G25470.1.Chr2.plus. Źródło: Opracowanie własne

Ponieważ w przeprowadzonej analizie, jako interesujących nas genów, nie zidentyfikowano tych, które były by związane z fitochromem, możemy stwierdzić, że badane przez nas mutanty, raczej nie posiadają zmian w budowie tego fotoreceptora. Nie należą więc do grupy najbardziej pospolitych mutacji fitochromowych wśród roślin.

Tabela (Rysunek 5.11) zawiera także 2 geny, których funkcja jak na razie, nie została poznana. Warto by było, przeprowadzić badania nad tymi genami, w kierunku odpowiedzi rośliny na światło. Być może udałoby się poznać ich funkcje i uzyskać interesujące wyniki.

Pokazano więc, że splicing index daje nam całkiem dobre pojęcie o różnicach wartości ekspresji egzonu między poszczególnymi grupami. Uwzględnia ekspresje genu, co nie było wykonywane przy bezpośrednim porównywaniu średniej ekspresji egzonów w grupach. Jednak wartości najbardziej zgodne z rzeczywistością, otrzymamy jeżeli w genie, z którego składa się badany przez nas egzon, znajduje się jak najwięcej egzonów konstytutywnych, a niewielka ilość alternatywnych. Jak każda wartość wyliczana za pomocą metod statystycznych, splicing index jest narażony na błędy statystyczne, a także posiada pewne wady.

Jedną z nich jest to, że nie uwzględniana jest nigdzie ilość egzonów w danym genie. W momencie gdy gen w którym jeden egzon ma bardzo wysoką wartość ekspresji, składa się z dużej ilości egzonów o przeciętnych wartościach, splicing index jest wyższy, niż w analogicznym przypadku, ale dla genu z małą ilością egzonów. Przykładowo (w przykładzie posłużono się bardzo uproszczonymi i wyidealizowanymi wartościami) mamy dwa geny, A składający się z 3 egzonów oraz B złożony z 6 egzonów. W genie A pierwszy egzon ma wartość ekspresji równą 2,2 w próbce wt a w przypadku knock out 1,1. Pozostałe egzony, w obu przypadkach, mają wartość 1. Dla tego egzonu wartość splicing indexu wynosi 0,39. W genie B mamy sytuacje analogiczną, czyli jeden gen o ekspresji podwyższonej - 2,2 u wt i 1,1 u knock out, reszta z ekspresją 1. Splicing index jest wtedy równy 0,76. Widać więc, jak duża jest różnica w wartości splicingu indexu mimo, że w obu przypadkach jest tylko jeden gen wyróżniający się tak samo spośród innych.

Kolejnym probemem jest to, że nawet najmniejsza zmiana we wzorze splicingowym, która nie zostanie wykryta lub zostanie uznana za nieistotną, może być bardzo ważna dla biologii. Tak więc, mimo wielu metod statystycznych, które w ogromnym stopniu są w stanie ułatwić

pracę biologom, zawężając zbiór danych i wskazując najistotniejsze elementy od których należało by rozpocząć badanie, nie możemy obejść się bez badań w laboratorium. Biologia jest nauką w której bardzo rzadko istnieją konkretne wzory, bądź reguły, a nawet jeśli istnieją, to ilość istniejących wyjątków uniemożliwia przeprowadzenie analizy, która dawałaby pewność, że uzyskane dane są faktycznie istotne biologicznie. Każde dane, które otrzymamy po analizach statystycznych wymagają konfrontacji z rzeczywistością, po to aby potwierdzić ich poprawność lub obalić ich istotność.

Istnieją także inne metody, które można wykorzystać do tego samego celu co splicing index. Jedną z nich jest MiDAS [8]. Koncepcyjnie jest to metoda podobna do wyliczania splicng index, jednak pozwala na porównania pomiędzy wieloma, różnymi grupami na raz. Nie jest to możliwe przy splicing index, który daje nam możliwość badania jedynie dwóch grup. Można także korzystać z VFC (Variation of reliability weighted Fold Changes) [9], opartego na wartościach ważonych fold changes i jest najbardziej wrażliwa na dane znajdujące się w ekstremach. Mimo istnienia innych metod, w tej pracy skupiono się na splicing index i jedynie zwrócono uwagę na alternatywne sposoby, umożliwiające wyszukiwanie wariantów splicingowych genów.

Podsumowując, przeprowadzono analizę danych uzyskanych z 6 eksperymentów mikromacierzowych, opartych na mikromacierzach tilingowych. 3 z nich zwierały dane otrzymane od osobników typu dzikiego, kolejne pochodziły od organizmów z mutacją w białku histonowym. Na początku przeprowadzono normalizację RMA, po której otrzymano 116199 wyników dla każdej mikromacierzy. Kolejnym istotnym krokiem, było wyliczenie wartości splicing index dla każdego egzonu, z użyciem dwóch różnych metod wyliczania poziomu genu. (za pomocą mediany, bądź średniej ze wszystkich ekspresji egzonów w danym genie). Gdy uzyskano wyniki, wybrano 10 egzonów o najwyższym module z wartości splicing index dla obu metod, a następnie wyodrębniono te z nich, które zostały zidentyfikowane przez obie metody. Otrzymano listę 6 takich egzonów. Dla genów, które zawierały jeden z nich, utworzono wykresy, przedstawiające uśrednioną wartość ekspresji w grupach oraz odpowiadającą mu wartość splicng index, dla każdego egzonu, wchodzącego w skład tego genu. Kolejnym krokiem było zapoznanie się z biologiczną funkcją wszystkich 6 genów i wybranie tych, które wydają się być najbardziej istotne dla naszych analiz. Wyodrębniono gen, mający wpływ na szlak transdukcji sygnału u *Arabidopsis thaliana* oraz przeprowadzono jego dokładniejszą analizę i wysnuto wnioski, do których doprowadziła nas powyższa analiza danych, oparta na metodach wykorzystywanych w statystyce, a następnie skonfrontowana z wiedzą biologiczną.

Dodatek A

Kody funkcji w języku R

FUNKCJA LICZĄCA SPLICING INDEX Z UŻYCIEM ŚREDNIEJ[1]

```
spliceIndexS <- function(T,DobreKolumny,ZleKolumny){

## KROK 1: Pogrupowac geny /dataFrame

     print("Wywolanie funkcji")
     flush.console()

     nazwy <- strsplit(row.names(T),"\\.")
     print("Podzielono nazwy na kawalki")
     flush.console()

## KROK 2 srednia dla kazdego genu w kazdej kazdej grupie w kazdej mikromacierzy

     for(i in 1:length(names(T))){
          print("Liczenie intensynowsci")
          flush.console()

          if(i==1) {naz_d=c()}

          war_d=c()
          S = c(T[1,i])

           for(j in 2:length(nazwy)){
               if(nazwy[[j]][1] == nazwy[[j-1]][1]&&nazwy[[j]][2] == nazwy[[j-1]][2]&&
               nazwy[[j]][3] == nazwy[[j-1]][3]&&nazwy[[j]][4] == nazwy[[j-1]][4] ){
                     S = c(S ,T[j,i])
               }
               else{
                       if(i==1){
                            naz_d = c(naz_d,paste(nazwy[[j-1]][1],nazwy[[j-1]][2],
                                    nazwy[[j-1]][3],nazwy[[j-1]][4],sep="."))
```

[1]Źródło: Opracowanie własne

```
                    }
                    SN = rowMeans(matrix(data = S, nrow=1))
                    war_d = c(war_d,SN)
                    S = c(T[j,i])
              }
          }
          if(i==1){
             naz_d = c(naz_d,paste(nazwy[[j]][1],nazwy[[j]][2],nazwy[[j]][3],nazwy[[j]][4],
             sep=".")[length(paste(nazwy[[j]][1],nazwy[[j]][2],nazwy[[j]][3],nazwy[[j]][4],
             sep="."))])
           }
           SN = rowMeans(matrix(data = S, nrow=1))
           war_d = c(war_d,SN)

# KROK 3: Dla kazdego exonu policzenie jego intensywnosci dla kazdej grupy/mikromacierzy

          intensywnosc=c()
          for(j in 1:dim(T)[1]){
            poz = match(paste(nazwy[[j]][1],nazwy[[j]][2],nazwy[[j]][3],nazwy[[j]][4],
            sep="."),naz_d)
            intensywnosc = c(intensywnosc,T[j,i]/war_d[poz])
          }

           if(i == 1){
                   IntenDF = data.frame(intensywnosc)
                   row.names(IntenDF) <- row.names(T)
            }
            else{
                    IntenDF = data.frame(IntenDF,intensywnosc)
            }
      }

      print("Policzono intensywnosci. Liczenie srednich intensywnosci grupami")
      flush.console()

## KROK 4:  SREDNIA DLA EXONOW DOBRYCH, SREDNIA DLA EXONOW ZLYCH

      dobreExony = c()
      zleExony = c()

       for(i in 1:dim(IntenDF)[1]){
           S = 0

           for(j in DobreKolumny)
                 S = S+IntenDF[i,j]

           S = S/length(DobreKolumny)
          dobreExony = c(dobreExony,S)
          S = 0
```

```
          for(j in ZleKolumny)
                S = S+IntenDF[i,j]

          S = S/length(ZleKolumny)
          zleExony = c(zleExony, S)
        }

     print("Policzno. Liczenie splicing indeksu")
     flush.console()
     DF = data.frame(GR1 = dobreExony,GR2 = zleExony)
     row.names(DF) = row.names(IntenDF)
     splicingIndex = data.frame(log2(DF[1]/DF[2]))
     row.names(splicingIndex) = row.names(T)
     names(splicingIndex) = c("SI")
     print("Skonczono.")
     flush.console()
     splicingIndex
}
```

FUNKCJA LICZĄCA SPLICING INDEX Z UŻYCIEM MEDIANY[2]

```
spliceIndexM <- function(T,DobreKolumny,ZleKolumny){

## KROK 1: Pogrupowac geny /dataFrame

     print("Wywolanie funkcji")
     flush.console()

     nazwy <- strsplit(row.names(T),"\\.")
     print("Podzielono nazwy na kawalki")
     flush.console()

## KROK 2 srednia dla kazdego genu w kazdej kazdej grupie w kazdej mikromacierzy

      for(i in 1:length(names(T))){
            print("Liczenie intensynowsci")
            flush.console()

            if(i==1) {naz_d=c()}

            war_d=c()
            S = c(T[1,i])
```

[2]Źródło: Opracowanie własne

```
            for(j in 2:length(nazwy)){

               if(nazwy[[j]][1] == nazwy[[j-1]][1]&&nazwy[[j]][2] == nazwy[[j-1]][2]&&
                nazwy[[j]][3] == nazwy[[j-1]][3] && nazwy[[j]][4] == nazwy[[j-1]][4] ){
                        S = c(S ,T[j,i])
               }
               else{
                  if(i==1){
                     naz_d = c(naz_d,paste(nazwy[[j-1]][1],nazwy[[j-1]][2],nazwy[[j-1]][3]
                     ,nazwy[[j-1]][4],sep="."))
                  }
                 SN = rowMedians(matrix(data = S, nrow=1))
                 war_d = c(war_d,SN)
                 S = c(T[j,i])
               }
             }

            if(i==1){
             naz_d = c(naz_d,paste(nazwy[[j]][1],nazwy[[j]][2],nazwy[[j]][3],nazwy[[j]][4],
             sep=".")[length(paste(nazwy[[j]][1],nazwy[[j]][2],nazwy[[j]][3],nazwy[[j]][4],
             sep="."))])
            }

              SN = rowMedians(matrix(data = S, nrow=1))
              war_d = c(war_d,SN)

# KROK 3: Dla kazdego exonu policzenie jego intensywnosci dla kazdej grupy/mikromacierzy

               intensywnosc=c()
               for(j in 1:dim(T)[1]){
                 poz = match(paste(nazwy[[j]][1],nazwy[[j]][2],nazwy[[j]][3],nazwy[[j]][4],
                 sep="."),naz_d)
                  intensywnosc = c(intensywnosc,T[j,i]/war_d[poz])
                }

               if(i == 1){
                  IntenDF = data.frame(intensywnosc)
                  row.names(IntenDF) <- row.names(T)
               }
               else{
                    IntenDF = data.frame(IntenDF,intensywnosc)
               }
     }
     print("Policzono intensywnosci. Liczenie srednich intensywnosci grupami")
     flush.console()

## KROK 4:  SREDNIA DLA EXONOW DOBRYCH, SREDNIA DLA EXONOW ZLYCH

      dobreExony = c()
```

```
zleExony = c()

for(i in 1:dim(IntenDF)[1]){
     S = 0

     for(j in DobreKolumny)
         S = S+IntenDF[i,j]

      S = S/length(DobreKolumny)
      dobreExony = c(dobreExony,S)
      S = 0

      for(j in ZleKolumny)
          S = S+IntenDF[i,j]

      S = S/length(ZleKolumny)
     zleExony = c(zleExony, S)
}

print("Policzno. Liczenie splicing indeksu")
flush.console()
DF = data.frame(GR1 = dobreExony,GR2 = zleExony)
row.names(DF) = row.names(IntenDF)
splicingIndex = data.frame(log2(DF[1]/DF[2]))
row.names(splicingIndex) = row.names(T)
names(splicingIndex) = c("SI")
print("Skonczono.")
flush.console()
splicingIndex
```

Bibliografia

[1] **Wiktorek-Smagur Aneta, Hnatuszko-Konka Katarzyna, Gerszberg Aneta, Łuchniak Piotr, Kononowicz Andrzej**: *"Arabidopsis thaliana - metody genetycznej transformacji"* Postępy Biologii Komórki Tom 36 s.177 - 187 Polskie Towarzystwo Anatomiczne, Polskie Towarzystwo Biologii Komórki, Fundacja Biologii Komórki i Biologii Molekularnej 2009r

[2] **Murray W. Nabors**: *"Introduction to Botany"* s. 270 Universtity of Mississippi 2004r

[3] **Żemieńko Agnieszka, Guzowska-Nowowiejska Magdalena, Pląder Wojciech, Figlerowicz Marek**: *"Analiza aktywności transkrypcyjnej genomu przy zastoowaniu mikromacierzy dachówkowych"* Biotechnologia 4 s. 101-114 Polska Federacja Biotechnologii 2008r

[4] **Rehrauer Hubert i inni**: *"AGRONOMICS1: A New Resource for Arabidopsis Transcriptome Profiling"* Plant Physiology Tom 152 s. 487-499 American Society of Plant Biologists 2010r

[5] **Arabidopsis GROwth Network integrating OMICS technologies:** http://www.agron-omics.eu/ [Zacytowano dnia: 10-08-2012 r]

[6] **Stępniak Piotr, Handschuh Luiza, Figlerowicz Marek**: *"Mikromacierze DNA - analiza danych"* Biotechnologia 4 s. 68-87 Polska Federacja Biotechnologii 2008r

[7] **Freudenberg Joannes**: *"Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays"* Leipzig Bioinformatics Working Paper 3 Interdisciplinary Centre for Bioinformatics 2005r

[8] **Identifying and Validating Alternative Splicing Events:** http://media.affymetrix.com/support/technical/technotes/id_altsplicingevents _technote.pdf [Zacytowano dnia: 05-08-2012 r]

[9] **Moller-Levet Carla i inni**: *"Exon Array Analysis of Head and Neck Cancers Identifies a Hypoxia Related Splice Variant of LAMA3 Associated with a Poor Prognosis"* PLoS Comput Biol 5(11) 2009r

[10] **Allison Lizabeth**: *"Podstawy biologii molekularnej"* Wydawnictwa Uniwersytetu Warszawskiego 2007r

[11] **Ludwików Agnieszka, Krasowski Krzysztof, Misztal Lucyna, Sadowski Jan**: *"Mikromacierze DNA i ich zastosowanie w badaniach ekspresji genów u roślin"* Biotechnologia 4 s. 131-143 Polska Federacja Biotechnologii 2008r

[12] **Brown Terry**: *"Genomy"* Wydawnictwo Naukowe PWN 2001r

[13] **The Arabidopsis Information Resource:** http://www.arabidopsis.org/ [Zacytowano dnia: 01-08-2012 r]

[14] **Stolc Viktor i inni**: *"Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays"* PNAS vol.102 no.12 s.4453–4458 2005r

[15] **Pracownia Analiz Mikromacierzy *corelab*:** http://www.microarrays.pl/ [Zacytowano dnia: 28-07-2012 r]

[16] **The R Project for Statistical Computing:** http://www.r-project.org/

[17] **Komsta Łukasz** *"Wprowadzenie do środowiska R"*
http://cran.r-project.org/doc/contrib/Komsta-Wprowadzenie.pdf

[18] **Biecek Przemysław** *"Przewodnik po pakiecie R"*
http://cran.r-project.org/doc/contrib/Biecek-R-basics.pdf

[19] **Arabidopsis GROwth Network integrating OMICS technologies:** http://www.agron-omics.eu/index.php/resource_center/tiling-array/tools-and-protocols [Zacytowano dnia: 13-08-2012 r]

[20] **Nowakowska Agnieszka, Tretyn Andrzej**: *"Mutanty fotomorfogenetyczne"* Wiadomości botaniczne 40(1) s. 37-51 1996r

[21] **Wykres pudełkowy:** http://pl.wikipedia.org/wiki/Wykres_pudełkowy [Zacytowano dnia: 23-08-2012 r]