POLITECHNIKA ŚLĄSKA

WYDZIAŁ AUTOMATYKI, ELEKTRONIKI I INFORMATYKI

KIERUNEK BIOTECHNOLOGIA

Projekt inżynierski

Aplikacja internetowa do analizy danych z eksperymentów real-time PCR

Autor: Jakub Porc

Kierujący pracą: Dr inż. Sebastian Student

Gliwice, styczeń 2017

Spis treści

Spis oznaczeń i symboli 4

Wprowadzenie 5

1. Cel projektu 6

2. Wstęp teoretyczny 6

2.1. Reakcja łańcuchowa polimerazy 6

2.2 Reakcja łańcuchowa polimerazy w czasie rzeczywistym (real-time PCR) 7

2.2.1 Metody pomiaru ilości amplikonu 8

2.2.2 Metody pomiaru poziomu ekspresji genów w eksperymencie real-time PCR 9

2.2.3 Metody obliczeń różnicy w poziomie ekspresji genów 10

2.2.3.1 Metoda ΔΔCt 10

2.2.3.2 Metoda Pfaffl 11

2.4 Wyznaczanie wydajności reakcji real-time PCR 12

2.5 Pakiet Shiny i struktura aplikacji 13

2.6 Przegląd programów do analizy wyników eksperymentu real-time PCR 14

3. Metodyka 16

3.1 Dane wejściowe 16

3.2 Wykorzystane dane testowe 16

3.3 Opis i struktura aplikacji 17

3.4 Interfejs graficzny 18

3.5 Wyniki działania programu 21

4. Podsumowanie 23

Bibliografia 24

Dodatek 1 – spis zawartości dołączonej płyty CD 25

Spis oznaczeń i symboli

DNA	–	kwas deoksyrybonukleinowy,
RNA	–	kwas rybonukleinowy,
PCR	–	reakcja łańcuchowa polimerazy,
Real-time PCR	–	reakcja łańcuchowa polimerazy w czasie rzeczywistym,
qPCR	–	ilościowa reakcja łańcuchowa polimerazy (inaczej real-time PCR),
RT-PCR	–	reakcja łańcuchowa polimerazy z odwrotną transkryptazą,
Ct	–	cykl progowy,
FAM	–	6-karboksyfluoresceina,
TAMRA	–	6-karboksy-tetrametylo-rodamina.

Wprowadzenie

Wraz z rozwojem wiedzy na temat biologii molekularnej oraz genetyki coraz większa grupa badaczy dąży do poznania mechanizmów działania istniejących chorób o podłożu genetycznym w nadziei na znalezienie sposobu na spowolnienie ich rozwoju lub całkowite wyleczenie. Do chorób tych należą m.in. nowotwory, których liczba przypadków ciągle wzrasta, mimo intensywnych prac mających na celu ich poznanie i wdrożenie nowych terapii. Mimo złożoności problemu tego typu schorzeń, do badań nad nimi często wykorzystuje się specyficzną grupę technik i analiz, które dostarczają specjalistom informacji na temat natury chorób, pozwalają na wyciągnięcie odpowiednich wniosków i podjęcie badań nad nową metodą leczenia. Wyniki takich eksperymentów często są jednak trudne do analizy bez zastosowania metod komputerowych i specjalnego oprogramowania. Nawet mało skomplikowane metody przetwarzania otrzymanych danych mogą sprawiać trudności i być bardzo czasochłonne dla badaczy nieposiadających doświadczenia w tej dziedzinie. Z tego powodu istnieje konieczność tworzenia oprogramowania i aplikacji, które pozwoliłyby na szybką, prostą i wszechstronną analizę wyników powstałych na drodze eksperymentów, w celu zwiększenia tępa pracy i zapewnienia odpowiednich metod analizy wyników.

Aby sprostać powyższym wyzwaniom podjęto się stworzenia prostej w obsłudze aplikacji internetowej pozwalającej na analizę danych z eksperymentu real-time PCR. Program ten jest skierowany m.in. do biologów i biotechnologów nieposiadających doświadczenia w komputerowej analizie danych, oraz do osób, którym zależy na szybkim uzyskaniu wyników.

W pierwszym rozdziale pracy przedstawiono cel tego projektu inżynierskiego. W rozdziale drugim zawarto wstęp merytoryczny oraz informacje potrzebne do zrozumienia założeń eksperymentu real-time PCR oraz metod analizy jego wyników. Rozdział trzeci zawiera opis danych wykorzystanych do testowania programu, zaimplementowane metody obliczeń, strukturę aplikacji i wyniki działania programu. W ostatnim rozdziale dokonano podsumowania wykonanej pracy, osiągniętych celów projektu oraz podano plany na ulepszenie aplikacji.

1. Cel projektu

Celem tego projektu inżynierskiego jest stworzenie interaktywnej aplikacji przetwarzającej dane wejściowe po stronie serwera, na której zainstalowane jest środowisko R z narzędziem o nazwie „Simple qPCR", której zadaniem jest analiza danych z eksperymentu real-time PCR. Aplikacja powinna być prosta w obsłudze i pozwalać na szybką i wygodną prezentację wyników. Do przetwarzania danych real-time PCR zaplanowano wykorzystanie i implementację metody Pfaffl, w której uwzględnia się wartości wydajności reakcji wyznaczone w ramach doświadczenia oraz możliwość wykorzystania wielu genów referencyjnych. Program powinien zawierać przykłady wczytywanych danych, wykonywać wszechstronną analizę wraz z wykresami i pozwalać na pobranie wyników wykonanych obliczeń. Jedną z głównych zalet programu powinna być możliwość wykorzystania go prze szerokie grono osób zainteresowanych analizą danych z eksperymentów real-time PCR, poprzez łatwy, darmowy dostęp, przejrzysty interfejs użytkownika i prostą obsługę, co pomogłoby mu wyróżnić się wśród innych programów tego typu.

2. Wstęp teoretyczny

2.1. Reakcja łańcuchowa polimerazy

Reakcja łańcuchowa polimerazy jest podstawową i zarazem jedną z najczęściej wykorzystywanych technik w laboratorium biologii molekularnej. Pozwala ona na powielenie wybranego fragmentu nici kwasu deoksyrybonukleinowego w niewielkim czasie, stosując proste metody bazujące na mechanizmie replikacji materiału genetycznego w żywych komórkach.

W pierwszej fazie eksperymentu zostają zaprojektowane startery, czyli krótkie jednoniciowe cząsteczki DNA, które są komplementarne do sekwencji matrycowego DNA oskrzydlających fragment, który ma zostać powielony. Wyznaczają one tym samym granice powielanej sekwencji. Wykorzystywane są startery *forward*, który wyznacza początek sekwencji oraz *reverse*, który wyznacza koniec sekwencji. Po zaprojektowaniu i syntezie starterów tworzy się mieszaninę reakcyjną, w skład której wchodzą matrycowe DNA, trifosforany deoksyrybonukleotydów, startery oraz termostabilna polimeraza DNA. Tak przygotowaną mieszaninę umieszcza się w termocyklerze – urządzeniu służącemu do amplifikacji (powielania)

materiału genetycznego poprzez zmiany temperatury w poszczególnych fazach reakcji. Cała reakcja zachodzi w kilkudziesięciu cyklach, które dzielą się na 3 etapy przebiegające w różnych temperaturach. W pierwszym etapie każdego z cykli mieszaninę podgrzewa się do temperatury ok. 95^{o}C, co skutkuje denaturacją DNA, czyli rozdzieleniem dwuniciowej cząsteczki DNA na pojedyncze łańcuchy. W kolejnym etapie temperaturę obniża się do wartości odpowiadającym temperaturze topnienia starterów, co skutkuje ich przyłączeniem się do komplementarnych miejsc na matrycy. W ostatnim etapie temperaturę zwiększa się do ok. 72^{o}C, co umożliwia związanie się termostabilnej polimerazy DNA ze starterami. Następnie za jej pomocą od startera syntetyzowana jest druga nić, poprzez jego wydłużanie i dołączanie do niego trifosforanów deoksyrybonukleotydów. Do tak utworzonej nici w kolejnych cyklach zostanie dołączony drugi starter i otrzymany zostanie docelowy fragment DNA. Po zakończeniu reakcji w próbce znajdują się miliony kopii powielanej sekwencji [2, 10, 12].

2.2 Reakcja łańcuchowa polimerazy w czasie rzeczywistym (real-time PCR)

Jednym z wariantów reakcji łańcuchowej polimerazy jest real-time PCR. Istotą tej techniki jest wykorzystanie jednej z metod śledzenia przebiegu reakcji w czasie rzeczywistym w celu określenia ilości powstałego produktu po każdym cyklu reakcji.

Aby przeprowadzić real-time PCR konieczne jest dodanie do mieszaniny reakcyjnej oprócz podstawowych substratów reakcji PCR interkalującego barwnika, czyli związku zdolnego do wiązania się wewnątrz dwuniciowej helisy DNA, lub sondy molekularnej [6, 7]. Podczas trwania reakcji związki te będą oddziaływać z dwuniciowym DNA obecnym w próbce i powodować fluorescencję, której poziom będzie proporcjonalny do ilości powielanego DNA, co daje informację na temat ilości amplikonu (produktu) w mieszaninie w każdym cyklu. Pozwala to również na określenie początkowego stężenia danej próbki (RNA lub DNA) i wynikającego z niego poziomu ekspresji poszczególnych genów. W pierwszych cyklach reakcji powielanie produktu zachodzi wolno, przez co poziom fluorescencji jest rejestrowany jako tło. Dopiero w kolejnych cyklach zostaje przekroczona wartość progowa, a cykl, w którym ją przekroczono, nazywa się cyklem progowym (Ct). Wraz ze wzrostem ilości początkowego materiału w próbce zmniejsza się czas potrzebny do osiągnięcia wartości progowej flu-

orescencji, przez co numer cyklu progowego jest niższy. Wartość Ct może być więc wykorzystywana do porównywania ilości fragmentów do których specyficzne są wykorzystane startery. Aby uzyskać dokładne informacje na temat liczby kopii w próbce wartość Ct porównywana jest z krzywą wzorcową, którą wyznacza się dla zestawu prób ze znanymi ilościami wzorca, lub dokonuje analizy zmian ilości obecnego w próbie transkryptu pomiędzy wybranymi genami i genami referencyjnymi [3, 12]. W celu wyznaczenia różnicy pomiędzy ekspresją tych genów wykorzystuje się różne metody obliczeniowe pozwalające na porównanie wartość Ct pomiędzy genami.

2.2.1 Metody pomiaru ilości amplikonu

W celu wykrycia i określenia ilości produktu reakcji PCR wykorzystuje się różnego rodzaju indywidua chemiczne. Ich zastosowanie pozwala na określenie poziomu fluorescencji, a mechanizm działania zależy od rodzaju zastosowanego reportera. Wyróżnić można dwie grupy związków: specyficzne do produktu reakcji i niespecyficzne.

Istotą detekcji dwuniciowych sekwencji DNA metodą niespecyficzną jest wykorzystanie fluorescencyjnych barwników wiążących się z DNA. Stosuje się takie związki jak SYBR Green I, EvaGreen, czy bromek etydyny. Każda z tych cząsteczek jest zdolna do fluorescencji, kiedy jest poddana działaniu światła o odpowiedniej długości fali. Wykorzystanie tych cząsteczek nie wymaga tyle wiedzy co projektowanie sond oligonukleotydowych specyficznych do określonego amplikonu, metoda ta jest też tańsza i może być wykorzystana w przypadku, gdy fragmenty sekwencji próbki różnią się między sobą. Istnieje jednak możliwość, iż startery w mieszaninie mogą wiązać się ze sobą tworząc dimery, czyli dwuniciowe cząsteczki, z którymi wiązać się mogą barwniki fluorescencyjne, przez co otrzymane wyniki mogą nie być miarodajne. Tworzy to konieczność zastosowania dodatkowej inkubacji w wysokiej temperaturze po etapie wydłużania starterów w reakcji PCR lub wykorzystania oprogramowania zdolnego do analizy krzywych topnienia fluorescencji. Dimery starterów to krótkie cząsteczki, których temperatura denaturacji jest niższa niż w przypadku amplikonu, dlatego możliwe jest rozpoznanie sygnału fluorescencyjnego pochodzącego od nich i dokonania korekcji [6]. Ważne jest również zastosowanie odpowiedniego stężenie barwników w reakcji, gdyż mogą one wpływać na jej przebieg. Zbyt wysokie stężenia barwników mogą doprowadzić do inhibicji reakcji PCR i konieczne może być wyznaczenie progu maksymalnego stężenia, który zależy od wielu czynników [8].

Jedną z głównych metod wykorzystujących cząsteczki specyficzne do określonej sekwencji DNA jest zastosowanie sond oligonukleotydowych. Dużą popularnością cieszą się sondy TaqMan, których mechanizm działania oparty jest na nukleolitycznym działaniu polimerazy DNA. Cząsteczki te to zmodyfikowane chemicznie fragmenty DNA, w których na jednym końcu znajduje się barwnik reporterowy (FAM), będący źródłem fluorescencji, natomiast na drugim końcu znajduje się barwnik tłumiący fluorescencję (TAMRA). Podczas reakcji PCR w każdym cyklu następuje denaturacja DNA, po której następuje przyłączanie się komplementarnych starterów na końcach sekwencji, która ma być amplifikowana. Jeżeli w mieszaninie obecne są sondy oligonukleotydowe komplementarne do fragmentu sekwencji produktu, to będą one również przyłączać się do DNA próbki. W ostatnim etapie cyklu następuje wydłużenie starterów i synteza produktu za pomocą polimerazy DNA. Jeżeli do próbki przyłączona jest sonda oligonukleotydowa, zostanie ona rozerwana przez polimerazę DNA dzięki jej nukleolitycznej aktywności w kierunku 5'-3', a barwnik reporterowy i cząsteczka wygaszająca fluorescencję zostaną rozdzielone, przez co możliwe będzie określenie poziomu fluorescencji pochodzącej od dokładnie określonego, specyficznego produktu [3, 6, 7].

2.2.2 Metody pomiaru poziomu ekspresji genów w eksperymencie real-time PCR

W celu określenia poziomu ekspresji genów stosuje się dwa główne podejścia – metodę ilościową (absolutną) oraz względną (relatywną). Obie te metody wymagają innych procedur laboratoryjnych, mają różne zastosowania oraz wady i zalety. Wybór pomiędzy nimi zależy przede wszystkim od celu eksperymentu.

Istotą metody absolutnej jest pomiar ilości na podstawie interpolacji danych z krzywej standardowej, która przedstawia zależność pomiędzy początkowym stężeniem próbki a wartością cyklu progowego. Krzywą standardową wyznacza się na podstawie znanych ilości produktu, którego rodzaj zależy od materiału badawczego i spodziewanej ilości próbki. Stosuje się m.in. rekombinowane plazmidowe DNA, genomowe DNA oraz produkty RT-PCR. Zastosowanie tego podejścia wymaga założenia, iż wydajności wszystkich reakcji, na pod-

stawie których wyznaczono krzywą, były bardzo podobne. W celu wykrycia możliwych inhibitorów reakcji PCR w próbie, związki te dodaje się również w określonej ilości do próby badanej w celu sprawdzenia, czy reakcja zachodzi prawidłowo. Zaletą tej metody jest możliwość dokładnego określenia ilości badanego materiału w określonych jednostkach (np. ng/µl, liczba kopii), natomiast wadą konieczność stworzenia krzywych kalibracyjnych, których jakość zależy od poprawności procedury laboratoryjnej, spełnienia założeń oraz stosowanych związków [12, 13].

Metoda względna opiera się na porównaniu ilości produktu pomiędzy dwoma próbkami – dla genu testowego oraz genu referencyjnego. Jako geny referencyjne stosuje się przede wszystkim geny *house-keeping* (geny podstawowego metabolizmu), których ekspresja jest względnie stała w komórce. Pozwala to na określenie zmian ekspresji w komórce oraz jej zmian fizjologicznych. Istnieje kilka modeli matematycznych wykorzystywanych do obliczeń zmian ekspresji, które można podzielić na te, które zakładają wydajność reakcji PCR oraz te, które wykorzystują rzeczywistą wydajność wyznaczoną na podstawie danych z eksperymentu. Zaletą tej metody jest brak konieczności sporządzania krzywych standardowych, jednakże ekspresja genów *house-keeping* może się różnić znacząco pomiędzy eksperymentami [6, 12].

2.2.3 Metody obliczeń różnicy w poziomie ekspresji genów

2.2.3.1 Metoda ΔΔCt

Jedną z popularniejszych metod służących do wyznaczania różnicy w ekspresji genów jest metoda komparatywna (**ΔΔ**Ct). Podejście to opiera się na stworzeniu krzywej standardowej na podstawie serii rozcieńczeń próby. Ważne jest, aby warunki, w których wykonuje się reakcję real-time PCR w celu wykonania krzywej, były zbliżone do tych w reakcji dla próby badanej, a wydajności reakcji były podobne. Metoda ta wykorzystuje model matematyczny pozwalający na wyznaczenie różnicy pomiędzy próbą badaną a genem referencyjnym. Potrzebne są 4 wartości Ct: wartości testowe i kontrolne dla genu będącego obiektem badań i genu referencyjnego. W pierwszym etapie obliczeń dokonuje się normalizacji wartości Ct próby badanej względem genu referencyjnego.

$$\Delta Ct = Ct(gen) - Ct(ref) \quad (1.1)$$

gdzie:

ΔCt – *znormalizowana wartość dla kalibratora lub badanej próby,*

$Ct(gen)$ – *wartość Ct dla genu będącego obiektem badań,*

$Ct(ref)$ – *wartość Ct dla genu referencyjnego.*

W kolejnym etapie dokonuje się normalizacji wartości Ct każdej próby względem kalibratora według poniższego wzoru:

$$\Delta\Delta Ct = \Delta Ct(test) - \Delta Ct(cal) \quad (1.2)$$

gdzie:

$\Delta\Delta Ct$ – *różnica wartości* ΔCt *badanej próby i kalibratora,*

$\Delta Ct(cal)$ – *wartość* ΔCt *dla kalibratora,*

$\Delta Ct(test)$ – *wartość* ΔCt *dla badanej próby.*

W ostatnim etapie obliczeń wyznacza się stosunek ekspresji badanej próby względem kalibratora [5, 10]:

$$Różnica = 2^{-\Delta\Delta Ct} \quad (1.3)$$

gdzie:

$\Delta\Delta Ct$ – *różnica wartości* ΔCt *badanej próby i kalibratora.*

2.2.3.2 Metoda Pfaffl

Kolejną z metod obliczeniowych wykorzystywanych we względnej analizie ekspresji genów jest metoda Pfaffl. Jej zastosowanie wykorzystuje wydajność reakcji w obliczeniach, przez co nie ma konieczności osiągnięcia bardzo zbliżonych wydajności w rzeczywistym eksperymencie. Część etapów obliczeń tej metody jest podobna do innych sposobów obliczeń, jednakże główną różnicą jest właśnie wykorzystanie wydajności wyznaczonej na podstawie eksperymentu. Aby wyznaczyć zmiany w ekspresji genów również potrzebne są 4 wartości Ct, a mianowicie wartości testowe i kontrolne dla genu będącego obiektem badań i genu referencyjnego. W pierwszej części obliczeń dochodzi do normalizacji wartości Ct kalibratora względem wartości Ct próby testowej zgodnie z poniższym wzorem:

$$\Delta Ct = Ct(cal) - Ct(test) \quad (2.1)$$

gdzie:

ΔCt – *znormalizowana wartość dla genu docelowego lub referencyjnego,*

$Ct(cal)$ – *wartość Ct dla kalibratora,*

$Ct(test)$ – *wartość Ct dla badanej próby.*

W drugim etapie wyznacza się różnicę w ekspresji genów wykorzystując poniższą formułę:

$$Różnica = \frac{E(gen)^{\Delta Ct(gen)}}{E(ref)^{\Delta Ct(ref)}} \quad (2.2)$$

gdzie:

$E(gen)$ – *wydajność przyłączania starterów dla genu będącego obiektem badań,*

$E(ref)$ – *wydajność przyłączania starterów dla genu referencyjnego,*

$\Delta Ct(gen)$ – *wartość ΔCt dla genu będącego obiektem badań,*

$\Delta Ct(ref)$ – *wartość ΔCt dla genu referencyjnego.*

Na podstawie tego wyniku można stwierdzić, ilu-krotnie zmieniła się ekspresja genu testowego względem genu referencyjnego [9].

2.4 Wyznaczanie wydajności reakcji real-time PCR

Wydajność reakcji PCR zależy w głównym stopniu od jej warunków oraz ilości substratów, dlatego stężenie i jakość matrycy, starterów i chlorku magnezu w mieszaninie ma znaczny wpływ na przebieg procesu i powinny być ściśle kontrolowane. Pozwala to na uzyskanie wysokiej dokładności w eksperymencie oraz zapewnia jego wysoką powtarzalność. Istnieją różne podejścia wyznaczania wydajności reakcji - metoda oparta na krzywej kalibracyjnej i metoda wykorzystująca krzywą przyrostu fluorescencji.

W celu wyznaczenia krzywej kalibracyjnej wydajności wykorzystuje się podstawowe równania charakteryzujące kinetykę reakcji PCR. Po wykonaniu serii reakcji real-time PCR dla wzorców nanosi się odpowiednie punkty na wykres zależności początkowego stężenia próbki od wartości Ct. Możliwe jest wyznaczenie prostej na podstawie tego zestawu prób, której nachylenie może być wykorzystane do obliczenia wydajności zgodnie ze wzorem:

$$\mathrm{E} = 10^{\frac{-1}{nachylenie}} - 1 \tag{3.1}$$

Zastosowanie tej metody wymaga wykonania szeregu rozcieńczeń dla matrycy i kilku powtórzeń, w celu osiągnięcia jak najlepszego wyniku. Każda z tych reakcji może jednak przebiegać z inną wydajnością, co w tym podejściu nie jest brane pod uwagę i jest wadą tej metody [9, 12, 13].

Metoda Lui i Sainta wyznaczania wydajności bazuje na krzywej przyrostu fluorescencji. Ponieważ wydajności reakcji dla genów badanych i referencyjnych nie zawsze są do siebie zbliżone, opracowano nową metodę, w której wyznacza się zakres wzrostu wykładniczego z danych, a następnie dobiera dowolne punkty z tego zakresu. Pozwala to na wyznaczenie wydajności reakcji stosując prosty wzór [5]:

$$E = \frac{Rna}{Rnb}^{-(Cta-Ctb)} \tag{3.2}$$

gdzie:

Rna – *wartość fluorescencji z punktu a,*

Rnb – *wartość fluorescencji z punktu b,*

Cta – *cykl, w którym osiągnięto wartość fluorescencji z punktu a,*

Ctb – *cykl, w którym osiągnięto wartość fluorescencji z punktu b.*

2.5 Pakiet Shiny i struktura aplikacji

Shiny jest pakietem dla środowiska R, dającym możliwość prostego tworzenie aplikacji internetowych, które pozwalają na interaktywną analizę danych. Dostępnych jest wiele gotowych elementów do interfejsu użytkownika, co umożliwia proste i wszechstronne jego tworzenie.

Aplikacja tworzona w Shiny składa się z 2 głównych skryptów: serwera oraz ui (user interface). Ui reprezentuje interfejs użytkownika oraz pozwala na wprowadzanie danych, natomiast serwer przyjmuje dane wejściowe i w nim wykonywane są wszystkie zadania i obliczenia, których wyniki następnie przedstawiane są w interfejsie. Jedną z głównych cech Shiny

jest reaktywność – istnieje możliwość zmiany wartości danych wejściowych w czasie rzeczywistym, które są przez program na nowo wykorzystywane do wykonania określonych czynności. Mechanizm działania programu opiera się na wejściach (*inputs*), funkcji renderowania oraz wyjściach (*outputs*). W pierwszym etapie użytkownik dokonuje wyboru danych, które są zapisywane jako wartości wejściowe. Następnie wykonywane są określone czynności i tworzone są wartości wyjściowe, które następnie wyświetlane są w interfejsie użytkownika. Aby stworzyć wartość wyjściową konieczne jest zastosowanie funkcji renderowania, które tworzą elementy interfejsu użytkownika bazując na wartościach wejściowych. Funkcje te mają wiele zastosowań, służą m.in. do tworzenia wykresów, tabel, diagramów itp. [1].

Gotowe aplikacje można udostępniać on-line poprzez umieszczenie ich na serwerze. Jednym z głównych i najbardziej rozpowszechnionych serwerów służących do dzielenia sią aplikacjami jest shinyapps.io. Jest to bezpłatna usługa, pozwalająca na szybkie i proste umieszczenie aplikacji na portalu, a każda dodana aplikacja otrzymuje swój własny adres URL, przez co dostęp do niej jest znacząco ułatwiony.

2.6 Przegląd programów do analizy wyników eksperymentu real-time PCR

Dotychczas stworzono wiele narzędzi zdolnych do przetwarzania wyników reakcji real-time PCR. Różnią się one zastosowanymi metodami obliczeniowymi, rodzajem wprowadzanych danych, rodzajem podawanych wyników i platformą, na której są dostępne. Część z nich wymaga od użytkownika podstawowej wiedzy z dziedziny programowania (np. pakiety do R), natomiast część to aplikacje z interfejsem użytkownika, które mogą być wykorzystane przez znacznie szersze grono użytkowników. W tym rozdziale skupiono się na programach darmowych, dostępnych dla największej grupy odbiorców, czyli narzędzi internetowych i dostępnych na system Windows.

Dużą część dostępnych narzędzi tworzą programy działające na systemie Windows. Aplikacje te mogą wymagać instalacji dodatkowego oprogramowania i wymagać od użytkownika podania wielu ustawień. Konieczne może być również założenie konta na portalu oraz uzyskanie licencji.

Program CopyCaller wykorzystuje metodę relatywną do wyznaczania różnic w ilości kopii pomiędzy sekwencjami bez normalizacji względem znanego kalibratora. Wynikiem działania programu jest liczba kopii próby oraz poziom ufności dla wyznaczonych wartości.

DART-PCR to program bazujący na Excelu, przyjmujący surowe dane z eksperymentu i obliczający wartości Ct. Program wykorzystuje metodę relatywną i potrafi wyznaczyć wydajność na podstawie danych, nie ma jednak możliwości dokonania normalizacji względem genów referencyjnych.

LinRegPCR to program z graficznym interfejsem służący do analizy surowych danych. Pozwala on na wyznaczenie linii bazowej fluorescencji i odjęcie jej od danych, a następnie wyznaczenie wydajności dla każdej próbki, wartości Ct i początkowego stężenia. Możliwe jest również stworzenie wykresów w celu porównania wydajności pomiędzy próbkami. Program nie pozwala jednak na analizę ilości badanej sekwencji.

qBase to program, który bazuje na Excelu i wykorzystuje względną metodę pomiaru ilości. Wykonywana jest propagacja błędów i istnieje możliwość zastosowania wielu genów referencyjnych. Istnieje też możliwość wyświetlania krzywych wykorzystywanych do obliczania wydajności oraz względnej ilości próbek. Stworzony został inny program qBase^{+}, który posiada znacznie większą funkcjonalność, jednakże dostęp do niego jest płatny.

Relative Expression Software Tool (REST) to aplikacja z interfejsem użytkownika dokonująca względnej analizy ilości próbek. Obsługuje on pliki tekstowe z wartościami Ct, pozwala na uwzględnienie różnych wartości wydajności dla genów i potrafi dokonać normalizacji względem wielu genów referencyjnych. Możliwe jest również wyświetlanie wykresów pudełkowych dla wyników.

Narzędzia internetowe pozwalają na prostszą analizę danych i nie wymagają zwykle instalacji programów, jednakże często istnieje konieczność wykonania kilku dodatkowych czynności. Duża część narzędzi opisana w dostępnej literaturze jest już niestety niedostępna.

MAKERAGUL to aplikacja wykorzystująca surową fluorescencję w celu wyznaczenia ilości próby bazując na modelu MAK2. Metoda ta może być użyta do obliczeń bez wykorzystania krzywych kalibracyjnych i genów referencyjnych. Narzędzie to musi być jednak zainstalowane i skonfigurowane na lokalnym serwerze, co może być źródłem wielu trudności.

QPCR to internetowe narzędzie zdolne do analizy, przechowywania i zarządzania wynikami z eksperymentu real-time PCR. Program przyjmuje surowe dane i może wykorzystać wiele metod w celu wyznaczenia wartości Ct i wydajności. Narzędzie posiada szeroką funkcjonalność, np. zdolność wykorzystania wielu genów referencyjnych, propagację błędów, obliczanie krotności zmian w ekspresji i wyświetlanie wykresów pudełkowych dla wyników. Konieczne jest jednak założenie konta na portalu, jak również podanie wielu ustawień dla wykonywanej analizy [11].

3. Metodyka

3.1 Dane wejściowe

Stworzona aplikacja wykorzystuje dane na bazie systemu ABI 7900HT w postaci SDS 2.3. Pliki w formacie tekstowym zawierają przede wszystkim wartości Ct dla każdej z próbek, jak również inne informacje na temat przeprowadzonego eksperymentu, takie jak zastosowany reporter, nazwa detektora itp. W przyszłości program będzie umożliwiał wykorzystanie danych w innych formatach.

3.2 Wykorzystane dane testowe

Wszystkie testy aplikacji przeprowadzono na danych z eksperymentu real-time PCR wykonanego w Centrum Onkologii – Instytucie im. Marii Skłodowskiej-Curie i Instytucie Onkologii w Warszawie. Próbki pochodzą z tkanek raka jajnika od pacjentów, u których nie przeprowadzono chemioterapii przedoperacyjnej i próbek, w których zanieczyszczenie komórkami zrębowymi było mniejsze niż 15%. Eksperyment został przeprowadzony z wykorzystaniem sond TaqMan [4], których mechanizm działania opisano w rozdziale 2.2.1. Dzięki temu można było zastosować dokładny pomiar ilości produktów specyficznych dla zaprojektowanych sond. Łącznie do testu aplikacji wykorzystano 10 plików z danymi dla następujących genów: ACTB, ATP6V1E1, BIRC3, GPRC5B, HADHA, IL1A, MFAP4,

POSTN i TNFAIP3. Każdy plik zawiera nazwy próbek i odpowiadające im wartości Ct, w tym wartość Ct dla kalibratorów.

3.3 Opis i struktura aplikacji

Aplikacja została wykonana w środowisku R za pomocą pakietu „Shiny". Posiada ona interfejs graficzny, w którym użytkownik może dokonać wczytania plików, wyświetlić wyniki analizy oraz obejrzeć wygenerowane wykresy. Zgodnie z wymogami „Shiny", aplikacja składa się z 2 głównych skryptów: server.R oraz ui.R, a także skryptu o nazwie funkcja1.R oraz obrazków wyświetlanych w narzędziu. Narzędzie zostało umieszczone na serwerze i dostępne jest pod adresem http://157.158.14.220:3838/simple_qpcr/.

Skrypt *ui.R* zawiera fragment kodu odpowiedzialny za stworzenie interfejsu użytkownika i wszystkich jego elementów. Poprzez niego użytkownik może: dokonać wczytania listy z wydajnościami dla poszczególnych genów oraz plików z danymi z eksperymentów, jak również dokonać wyboru genów referencyjnych z wczytanej listy, które są zapisywane jako wejścia (*inputs*). Przedstawia on również wygląd całego interfejsu wraz z umiejscowieniem wszystkich paneli, zakładek i przycisków.

server.R służy do tworzenia wszystkich elementów będących wynikiem analizy (*outputs*). W nim zdefiniowane są: wybrane geny referencyjne z listy, dane znajdujące się w tabeli wynikowej, wykresy pudełkowe i struktura pliku do pobrania. Są one generowane za pomocą funkcji renderowania na podstawie danych wejściowych.

funkcja1.R to główna funkcja, w której zaimplementowano metodę Pfaffl i w której wykonywane są wszystkie obliczenia. W pierwszym etapie geny dzielone są na dwie grupy: geny badane i referencyjne, dokonuje się odczytu wartości *Ct* i obliczenie średniej z wartości dla kalibratorów. Jeśli występują braki wartości *Ct* dla próbek, program omija je. Następnie dla wszystkich próbek obliczane są kolejno wartości ΔCt oraz Q $(E^{\Delta Ct})$ zgodnie z protokołem metody Pfaffl. Jeżeli dokonano wyboru więcej niż jednego genu referencyjnego, obliczana jest średnia geometryczna wartości Q wszystkich genów referencyjnych. Po wykonaniu powyższych obliczeń dla genów badanych i referencyjnych obliczona zostaje wartość *Fold difference* (zmiana ekspresji) zgodnie z poniższym wzorem:

$$Fd = \frac{Q(\text{gen})}{Q(ref)} \quad (4)$$

gdzie:

Fd – *fold difference,*

Q(gen) – *wartość Q dla genu badanego,*

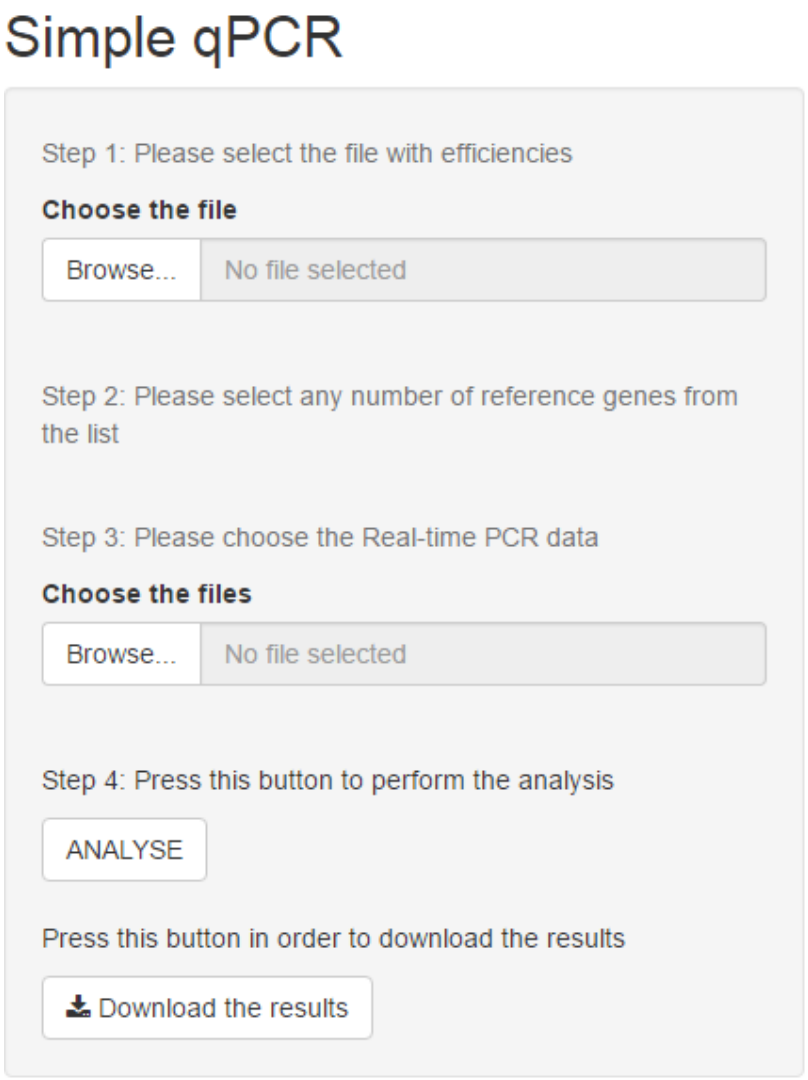
Q(ref) – *wartość Q dla genu referencyjnego.*

3.4 Interfejs graficzny

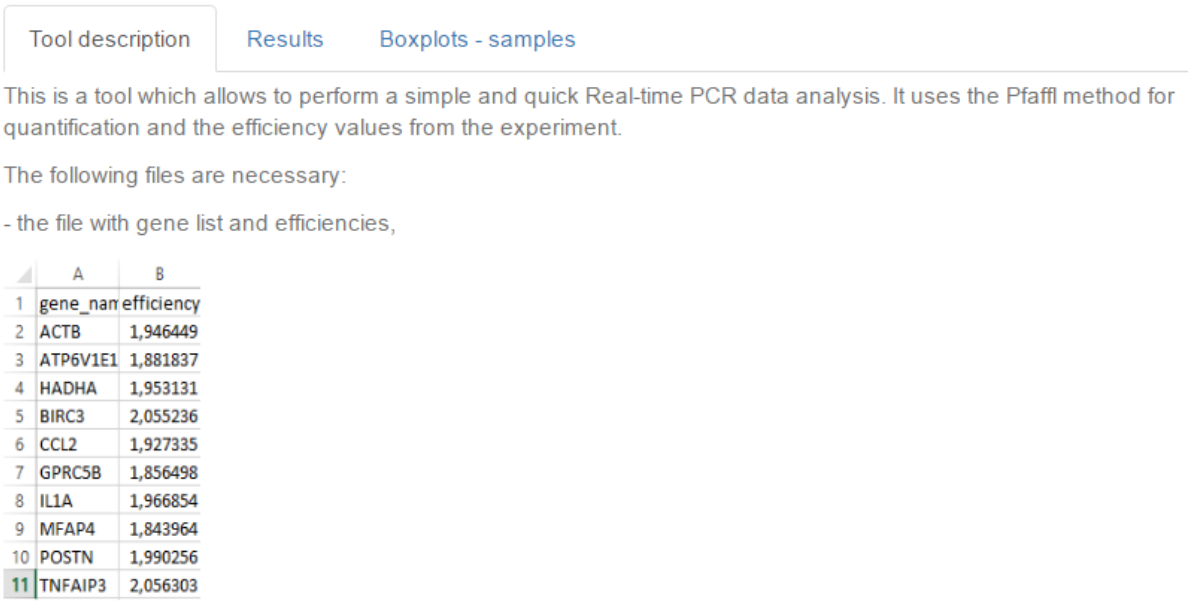
Interfejs graficzny stworzonej aplikacji składa się z 2 głównych części: panelu bocznego oraz panelu głównego. Zawartość każdego z nich jest wyświetlana w taki sposób, aby jak najlepiej dostosować się do rozmiaru okna.

W panelu bocznym znajdują się elementy interfejsu związane z wczytywaniem i wyborem danych. W pierwszym kroku użytkownik ma możliwość wczytania pliku z listą nazw genów referencyjnych. Następnie użytkownik dokonuje wyboru genów z wczytanej listy, które zostaną wykorzystane do dalszej analizy. W trzecim kroku użytkownik wczytuje zestaw danych z eksperymentu. Po wykonaniu powyższych czynności można rozpocząć wykonywanie obliczeń poprzez naciśnięcie przycisku „ANALYSE”. Istnieje również możliwość pobrania wyników wykonanej analizy poprzez naciśnięcie przycisku „Download the results”.

W panelu głównym znajdują się trzy zakładki: Tool description, Results, Boxplot – samples. W pierwszej z nich znajduje się krótki opis narzędzia, instrukcja obsługi oraz lista potrzebnych plików wraz z obrazkami przedstawiającymi ich wymaganą strukturę, jak również wymagany format nazwy plików. W drugiej znajduje się tabela z wynikiem analizy, a w ostatniej wykresy pudełkowe dla każdej próbki.



Obraz 1 Panel boczny aplikacji



Obraz 2 Panel główny: zakładki oraz część opisu działania aplikacji

- the Real-time PCR data in SDS 2.3 format.

```
SDS 2.3 AQ Results      1.0
Filename        pl2_07_10_2014_KK_cDNA_384H
PlateID
Assay Type      Absolute Quantification
Run DateTime    10/7/14 4:29:22 PM
Operator
ThermalCycleParams

Sample Information

Well    Sample Name     Detector Name   Reporter        Task    Ct      Quantity        Qty Mean        Qty
StdDev  Ct Median       Ct Mean Ct StdDev       Baseline Type   Baseline Start  Baseline Stop   Threshold Type
Threshold       FOS     HMD     LME     EW      BPR     NAW     HNS     HRN     EAF     BAF     TAF     CAF
55      121+122p1       Stanley FAM     Unknown 18.144453
Automatic                       Manual  0.112289794

79      121+122p2       Stanley FAM     Unknown 18.063736
Automatic                       Manual  0.112289794

103     124+125p1       Stanley FAM     Unknown 18.07071
Automatic                       Manual  0.112289794

127     124+125p2       Stanley FAM     Unknown 18.172913
Automatic                       Manual  0.112289794

151     127+128p1       Stanley FAM     Unknown 18.267763
Automatic                       Manual  0.112289794

175     127+128p2       Stanley FAM     Unknown 18.27802
Automatic                       Manual  0.112289794

102     14p1    Stanley FAM     Unknown 18.310541
Automatic                       Manual  0.112289794
```

The file names of the data must have the following format: gene_other

Where:

gene - gene name,

other - any additional information.

The output table will be visible in the Results tab, while the boxplots will be available in the boxplots tab.

If you wish to save the results, press the 'Download the results' button.

Note: The program assumes that each plate has the same samples.

Obraz 3 Dalsza część opisu aplikacji wraz z przykładowymi danymi

3.5 Wyniki działania programu

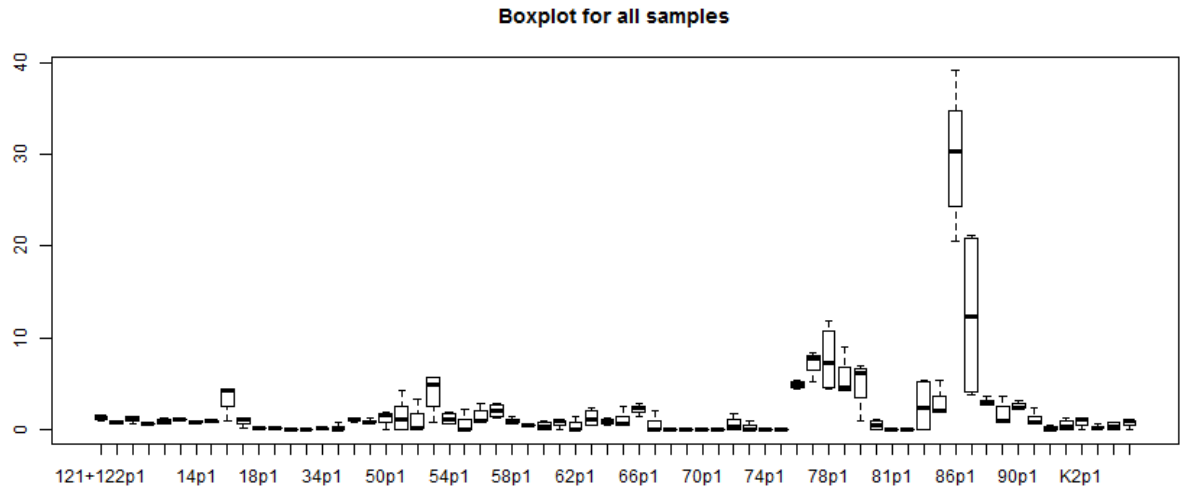
Po wykonaniu obliczeń dla danych wykorzystanych do testów program generuje:

- tabelę z wartościami *fold difference*,

Sample Name	ACTB	BIRC3	CCL2	GPRC5B	IL1A
121+122p1	1.13847554938083	4.09670805212449	1.18658422822053	2.54323732167115	0.499772081750035
121+122p2	1.17492074198917	3.98931574054315	1.21664734804807	2.6109333684399	0.590748490415339
124+125p1	1.15624821596073	4.08076147659764	1.05310806074133	2.68936421291295	0.526463895631909
124+125p2	1.16899323750131	4.15685565304566	1.02679096044858	2.64778986812624	0.522335499928422
127+128p1	1.13380471169414	3.77652259983747	1.13019464399269	2.24209951431513	0.58064349514203
127+128p2	1.12167578948294	3.7032727385796	1.07165961457179	2.2455573286825	0.562865457314627
14p1	0.886077555670707	1.14388113290061	0.746996738378139	1.51665997559285	2.31788570510448
14p2	0.862868715957436	1.07418552375432	0.788032480846823	1.50331124226183	2.5052788997112
16p1	0.951143023320845	1.2242287450095	0.770420125119262	1.34913718715238	2.31901693902567
16p2	0.926639576219997	1.21601389414855	0.800804933908945	1.39555196823034	2.48688980783748
18p1	0.937292173201797	1.29847163892502	0.749144710430388	1.38246445296051	2.39455016927
18p2	0.915937582199359	1.22992234260313	0.762602679217617	1.45301827375883	2.83393756695742
32p1	0.942202500292248	0.13454058491008	2.08120740608717	0.896084904398088	2.69103735493894
32p2	0.961066118128227	0.142567262331688	2.00691759096615	0.876222850790427	2.3728610920583
34p1	1.1594165783702	0.16376502100224	1.98502646781579	0.92967343939727	1.8253123198375

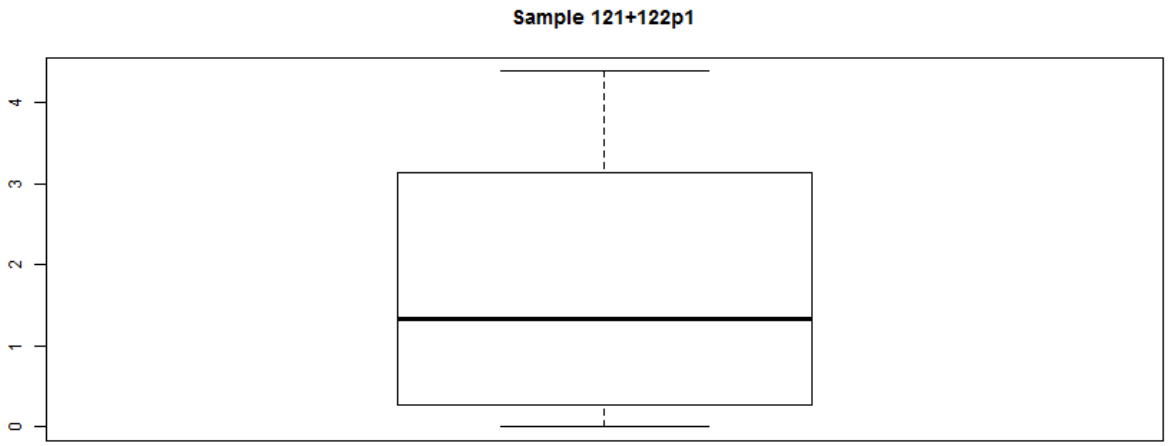
Obraz 4 Fragment przykładowej tabeli wynikowej

- wykresy pudełkowe dla wszystkich próbek,



Obraz 5 Zbiorczy wykres pudełkowy dla wszystkich próbek

- wykresy pudełkowe dla pojedynczych próbek,



Obraz 6 Przykładowy wykres pudełkowy dla jednej z próbek

- plik z wynikami w formacie XLSX.

	A	B	C	D	E	F	G	H	I	J
1	Sample.N	ATP6V1E1	BIRC3	CCL2	GPRC5B	IL1A	MFAP4	POSTN	TNFAIP3	
2	121+122p1	1.3961328	4.3849976	1.2700853	2.7222075	0.5349415	0.0030382	0.0009840	3.53297607380287	
3	121+122p2	1.4383226	4.2675733	1.3015093	2.7930478	0.6319536	0.0034087	0.0006793	3.67620023272936	
4	124+125p1	1.4336964	4.3840609	1.1313795	2.8892491	0.5655929	0.0074589	0.0011641	3.2412022103936	
5	124+125p2	1.4224908	4.4378992	1.0962119	2.8268060	0.5576504	0.0070125	0.0013299	2.90374070739887	
6	127+128p1	1.4953068	4.1414851	1.2394164	2.4587756	0.6367567	0.0035687	0.0020593	2.76505415638895	
7	127+128p2	1.4584193	4.0419480	1.1696660	2.4509202	0.6143411	0.0045538	0.0022192	2.48977604879883	
8	14p1	0.9547064	1.1726820	0.7658047	1.5548467	2.3762459	No data	No data	2.43392080312937	
9	14p2	0.9564937	1.1117106	0.8155612	1.5558272	2.5927971	No data	4.8531563	2.37161204819159	
10	16p1	0.9008285	1.2022497	0.7565884	1.3249156	2.2773827	No data	6.6368120	2.27310188463342	
11	16p2	0.9653676	1.2327239	0.8118093	1.4147292	2.5210638	No data	7.1865804	2.25609437406945	
12	18p1	0.9724208	1.3144948	0.7583891	1.3995241	2.4240990	No data	No data	2.05936987744992	
13	18p2	0.9662550	1.2520441	0.7763191	1.4791527	2.8849098	No data	5.5482597	2.33543698612568	
14	32p1	0.7651876	0.1255243	1.9417348	0.8360335	2.5106969	0.0004711	No data	0.832178908060378	
15	32p2	0.7838721	0.1332039	1.8751104	0.8186756	2.2170200	No data	No data	0.776757584506958	
16	34p1	0.6639746	0.1359956	1.6484292	0.7720304	1.5157974	0.0004166	No data	0.709231311104286	
17	34p2	0.6756958	0.1235301	1.6420250	0.7237864	1.7481487	0.0003999	No data	0.674678311639217	
18	36p1	0.6828213	0.1297885	1.7174490	0.7383558	1.9335169	No data	No data	0.750699830218832	
19	36p2	0.7507111	0.1492206	1.6989795	0.8050282	2.0812329	0.0006530	No data	0.881335980976727	
20	50p1	1.3093554	0.1730477	0.0002449	0.4618129	0.0157290	0.0014531	0.0025581	0.0446902853582789	
21	50p2	1.3127560	0.1674377	0.0001327	0.5106494	0.0183910	0.0007995	0.0011069	0.0451215737184783	

Obraz 7 Pobrane wyniki w formacie XLSX

4. Podsumowanie

Analiza danych z eksperymentów będących źródłem bardzo dużej ilości danych, do których należy real-time PCR, jest bardzo trudna bez zastosowania zaawansowanych metod komputerowych. Brak biegłości w tej dziedzinie może być dla osób zajmujących się tymi badaniami barierą uniemożliwiającą utrzymanie tempa prowadzonych badań i może wymagać pomocy ze strony innych specjalistów lub zakupu dodatkowego oprogramowania. Realizacja celów niniejszego projektu inżynierskiego miała za zadanie rozwiązanie tego problemu, co próbowano osiągnąć tworząc opisaną w tej pracy aplikację i dbając o jej funkcjonalność. Stworzony interfejs użytkownika jest prosty w obsłudze, program pozwala na prostą i szybką analizę danych, wykorzystując metodę Pfaffl, która pozwala na osiągnięcie dokładniejszych wyników dzięki wykorzystaniu wydajności reakcji. Narzędzie zawiera też informację, jak krok po kroku wykonać analizę i jak przygotować do niej dane. Tabela oraz wykresy pudełkowe dla genów i pojedynczych próbek pozwalają na porównanie wyników pomiędzy sobą oraz pomagają wskazać czynniki, które mogą mieć wpływ na różnicę w ekspresji poszczególnych genów.

Narzędzie to wymaga jednak od użytkownika, aby z surowych danych zostały wyznaczone wartości Ct oraz wydajności reakcji, a pliki wejściowe muszą być w określonym formacie. Program zakłada również, że na wszystkich płytkach zostały umieszczone takie same próbki w takiej samej kolejności, nie dokonuje się również żadnej statystycznej analizy wyników. W celu zwiększenia funkcjonalności, do programu w przyszłości zostanie dodana możliwość wyznaczania wartości Ct oraz wydajności z surowych danych, obsługa innych typów plików z danymi, jak również możliwość dodania pliku konfiguracyjnego z informacjami na temat próbek obecnych na płytce oraz ich rozmieszczenia.

Bibliografia

[1] Chris Beeley, *Web application development with R using Shiny*, Packt Publishing Ltd. 2013.

[2] Garibyan L., Avashia N., *Research Techniques Made Simple: Polymerase Chain Reaction (PCR)*, J Invest Dermatol. 133 2013, str. 1-4.

[3] Heid C, Stevens J, Livak K, Williams P, *Real time quantitative PCR*, Genome Research 6 1996, str. 986-994.

[4] Lisowska, K.M., Olbryt, M., Student, S. et al. *J Cancer Res Clin Oncol* 142 2016, str. 1239- 1252.

[5] Livak, K., Schmittgen, T., *Analysis of relative gene expression data using real-time quantitative PCR and the 2^[-delta delta C(T)] Method*, Methods 25 2001, str. 402-408.

[6] Mackay M., Ardem K., Nitsche A., *Real-time PCR in virology*, Nucleic Acids Res (30) 2002, str. 1292–1305.

[7] Mackay M., *Real-time PCR in the microbiology laboratory*, CMI 10 2004, str. 190-212.

[8] Mao F, Leung W-Y, Xin X. *Characterization of EvaGreen and the implication of its physicochemical properties for qPCR applications*. BMC Biotechnology 7:76, 2007.

[9] Michael W. Pfaffl, *A-Z of quantitative PCR*, International University Line 2004, str. 87-112.

[10] Mullis K., Faloona F., Scharf S., Saiki R., Horn G., Erlich H., *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction*, Cold Spring Harbor Symposia on Quantitative Biology, Cold Spring Harbor, NY 1986, str. 263-273.

[11] Pabinger S., Rödiger S., Kriegner A., Vierlinger K., Weinhäusel A., *A survey of tools for the analysis of quantitative PCR (qPCR) data*, Biomolecular Detection and Quantification 1 2014, str. 23–33.

[12] Studzińska A., Tyburski J., Daca P., Tretyn A., *PCR w czasie rzeczywistym. Istota metody i strategie monitorowania przebiegu reakcji*, Biotechnologia 1 (80) 2008, str. 71-96.

[13] Wong M., Medrano J., *Real-time PCR for mRNA quantitation*, BioTechniques 39 2005, str. 75-85.

Dodatek 1 – spis zawartości dołączonej płyty CD

- Praca inżynierska w formacie PDF