AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE
WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI, INFORMATYKI
I INŻYNIERII BIOMEDYCZNEJ

ROZPRAWA DOKTORSKA

Model biocybernetyczny w analizie sekwencji genomu człowieka

mgr inż. Krzysztof Sarapata

Promotorzy pracy:
Prof. dr hab. med. Marek Sanak
Prof. dr hab. inż. Ryszard Tadeusiewicz

Kraków 2015

Tą pracę chciałbym poświęcić pamięci Karoliny Żyraldo,

profesora Jana Trąbki

oraz Marcina Usarza

Spis treści

Teza rozprawy

Możliwe jest rozszerzenie modelu TargetScore do postaci proponowanego w pracy modelu biocybernetycznego biTargetScore, zbudowanie takiego modelu oraz jego walidacja na podstawie danych eksperymentalnych dotyczących sekwencji genomu człowieka uzyskanych z materiału klinicznego

1 Wstęp

1.1 Omówienie podejmowanej tematyki rozprawy

Aktualny stopień zaawansowania nauki pozwala między innymi na eksplorację przyrody na poziomie molekularnym. Znaczącym i frapującym staje się przede wszystkim zagadnienie poznawania przyrody ożywionej, która po pierwsze stanowi nieskończenie bardziej skomplikowany, heterogeniczny system organizacji materii, a po drugie badacz odkrywając procesy życiowe niejako poznaje siebie samego. Jednak naukową informację, jaką uzyskujemy w wyniku przeprowadzanych eksperymentów biologicznych, charakteryzuje kolektywność, uśrednienie. Procedura eksperymentu najczęściej przebiega w kierunku analitycznym, od złożoności do uproszczenia. Odwrotny kierunek, czyli ogólny opis i wyjaśnianie poznanych zjawisk szczegółowych jest już domeną abstrakcyjną, wyobrażeniową. W jej wyniku powstaje synteza i wnioskowanie o funkcji całości.

Chronologicznie w toku ewolucji współczesnej nauki podejmowana w rozprawie tematyka mieściła się kolejno w obrębie przyrody, biologii, genetyki, genetyki molekularnej, a obecnie w dyscyplinie granicznej, określanej pojęciem *bioinformatyki*. Pojęcie to definiuje się, jako interdyscyplinarną dziedzinę nauki łączącą [77]:

- rozwój metod obliczeniowych służących do badania struktury, funkcji i ewolucji genów, białek i całych genomów,
- rozwój metod do zarządzania i analizy informacji biologicznej, gromadzonej w toku badań genomicznych, oraz badań prowadzonych z zastosowaniem wysokoprzepustowych technik eksperymentalnych.

Użyte w definicji bioinformatyki określenie *interdyscyplinarność* z jednej strony potwierdza tezę o wspólnym materialnym i organizacyjnym budulcu wszechświata i żywych organizmów, z drugiej interdyscyplinarność wymusza akceptację nowych idei przez dziedziny nauki na nią się składające. Przykładem może być wykorzystanie teorii informacji i teorii kodowania, czyli dyscyplin technicznych wiedzy w biologii, które obserwujemy w tzw. centralnym dogmacie biologii molekularnej demonstrującym przepływ informacji z wykorzystaniem kodu genetycznego. Odkrycie kodowania za pomocą genów nie dokonało się ad hoc. W 1869 roku F. Miescher dokonał pierwszej izolacji DNA z komórki, a dopiero po 100 latach R.W. Holley, H.G. Khorana, H. Matthaei i M.W. Nirenberg rozszyfrowali kod genetyczny, który wyjaśnia podstawowe reguły syntaktyczne i semantyczne interpretatora kodu źródłowego – sekwencji kodującej DNA. Miało to miejsce dopiero po zdefiniowaniu komputera przez Alana Turinga i John von Neumanna.

Jeżeli pojęcie bioinformatyki stanowi ogólne określenie technicznego podłoża realizacji modelowania, symulacji i obliczeń służących analizie informacji biologicznej, gromadzonej w toku badań genomicznych, to chronologicznie trochę starsze pojęcie cybernetyki precyzyjniej definiuje, jaki te realizacje mają charakter.

Teoria informacji, teoria maszyn cyfrowych, teoria kodowania i inne pokrewne dyscypliny naukowe połączone razem stanowią podstawę utworzonej w 1946 roku przez Norberta Wienera nowej interdyscyplinarnej dyscypliny naukowej – cybernetyki. Jej biologiczne aplikacje określa się mianem biocybernetyki.

Przepływ informacji biologicznej oraz jego modele nie byłyby pełne, gdyby nie uwzględnić problematyki regulacji, która stanowi formę organizacji informacji i jej znaczenia. Profesor Jan Trąbka w ten sposób komentował osiągnięcia N. Wienera [170]:

"Informacja występuje zawsze z regulacją. Cybernetyka uznawała rolę celu zadanego z góry i pochodzącego spoza układu, także ze świadomości ludzkiej. To odwołanie się do teleologicznej świadomości stanowi pierwszy wyłom w gmachu scjentyzmu i posiada wręcz nieprzewidywalne konsekwencje, ponieważ pozwala na stawianie pytań: *Dlaczego tak, a nie inaczej wybraliśmy cel?*, a zwłaszcza zapewnia odpowiedź etyczną na indagacje o *moralnym imperatywie*, w której pionierem był genialny Kant. Cybernetyka to cenny nabytek intelektu, skoro potrafiła tak perfekcyjnie sugerować i prowadzić postęp techniczno-cywilizacyjny, że wzbudziła obawy przed zastąpieniem umysłu ludzkiego przez maszyny matematyczne."

Na tle tego cytatu można sformułować następujące pytania: Czy modelowanie to naśladowanie, kopiowanie, czy też twórcze poszukiwanie nieznanych mechanizmów natury? Czy inspiracja technicznymi realizacjami cybernetyki pozwoli wykazać podobieństwo mechanizmów natury do systemów maszyn liczących, teorii informacji, kodowania czy komunikacji?

Treść tej pracy będzie pośrednio stanowiła próbę odpowiedzi na te pytania.

Wiedza o molekularnych podstawach życia ujawniła system informacji genetycznej. System ten, rozważany w kolejności od najszerszych ram czasowych obejmuje kolejno: ewolucję (filogenezę), dziedziczenie cech (fenotypów), i ostatecznie rozwój osobniczy (ontogenezę). Każdy z tych poziomów został zamodelowany na poziomie molekularnym. System informacji genetycznej cechuje uporządkowanie, linearność i cyfryzacja. Idąc dalej tym technicznym tropem dostrzegamy w nim elementy uporządkowania hierarchicznego, wielopoziomowego czyli najlepiej byłoby określić jego złożoność (biologiczną) jako obiektowość (informatyczną). Przedstawiona w rozprawie tematyka generalnie koncentruje się na mechanizmach i procesach regulacji genów, a w szczególności na stosunkowo niedawno odkrytym procesie interferencji RNA (RNAi). Mechanizm ten został opisany na poziomie molekularnym i został poparty eksperymentami, które *a priori* zakładały nadrzędny sposób tej regulacji z jej podstawowym efektem łączenia się w pary cząsteczki miRNA i odpowiedniego komplementarnego odcinka transkryptu. To łączenie się, czyli hybrydyzacja, doprowadza różnymi metodami do degradacji transkryptu, realizując w ten sposób ujemne sprzężenie zwrotne korygujące ilość i stężenie odpowiednich transkryptów w cytoplazmie.

Jednak pewien niepokój powstał, gdy zaczęto dokładniej analizować ten proces. Przedstawiona idea tego procesu zakłada bowiem nieprecyzyjność działania czynników transkrypcyjnych w kwestii bilansowania podaży i popytu na transkrypty. Czynniki transkrypcyjne stanowią podstawowy element regulacji procesu transkrypcji. Mechanizm interferencji RNA nie rozstrzyga, czy owa nadprodukcja transkryptów podyktowana została czynnikiem patogennym, czy jest niedoskonałością samą w sobie wcześniejszych etapów regulacji genów. Zakłócenia – pojęcie techniczne, pojawiające się w torze – ścieżce transmisji informacji genetycznej na jej nośniku – transkrypcie w takiej sytuacji powinno się zakwalifikować do czynników patogennych. Znane są pewne mechanizmy detekcji i kontroli tych zakłóceń, ale główny ciężar oceny jakości transkryptu przypisuje się organelli komórkowej - retikulum endoplazmatycznemu. Dopiero tam, po translacji i etapach fałdowania białka, odpowiedni mechanizm kontroluje jakość dostarczonego sygnału - transkryptu. Naturalnym wydaje się, że każdy rodzaj regulacji genów występujący w torze transmisji i znajdujący się dalej od matrycy DNA (źródła) stanowi tak naprawdę mechanizm

kontroli jakości sygnału. A oczywistym jest, że jakość tego sygnału pogarsza się wraz odległością od źródła (matrycy DNA).

Kolejna refleksja potwierdzająca niejasność mechanizmu interferencji RNA związana jest z zaskakującymi rezultatami analiz przeprowadzonych w ramach tzw. dynamiki molekularnej badającej stabilność dupleksów miRNA/transkrypt. Wykazano, że nie jest preferowana wysoka komplementarność sekwencji cząsteczek dupleksu, jak również nie nadrabia tego deficytu relatywnie niska wartość jego energii swobodnej.

Przedstawione dylematy dają podstawę do weryfikacji i modyfikacji obecnego modelu regulacji, jaka ma miejsce w ramach procesu interferencji RNA. Taką próbę weryfikacji i modyfikacji stanowi ta praca.

Uzasadnienie prowadzenia badań w tym kierunku stanowi chęć poszerzenia możliwości sterowania i manipulacji fenotypowej za pomocą kodujących sekwencji. Trudność w ocenie uwarunkowań genetycznych organizmu, czy w zdefiniowaniu poligeniczności uzasadnia potrzebę poszerzenia znaczenia i interpretacji sekwencji genomu zlokalizowanych poza samymi rejonami kodującymi.

W zaproponowanym modelu typowania par miRNA/mRNA wykorzystano bibliotekę TargetScore implementującą probabilistyczny model Bayesa połączony z wnioskowaniem wariacyjnym występujący pod nazwą *Variational Bayesian-Gaussian Mixture Model* (VB-GMM)[15].

1.2 Cel i zakres pracy

Podczas analizowania publikacji dotyczących zagadnienia regulacji w strukturach żywych najbardziej uderza kontrast rozwiniętej, uporządkowanej i zastosowanej wiedzy abstrakcyjno-matematycznej w porównaniu do sprawiającej wrażenie chaotycznie zorganizowanej wiedzy dotyczącej tej materii w żywych strukturach, trzymanej jedynie w ryzach z wykorzystaniem kilku poznanych reguł. Nie można jednak zapominać, że ta rozwinięta uporządkowana abstrakcja matematyczna jako taka jest wytworem owej poznawanej, pozornie chaotycznej, struktury biologicznej.

Wykorzystując wiedzę na temat struktury cząsteczki nie jesteśmy w stanie przewidzieć jej funkcjonalności. W rozwiązaniu problemu pomaga podejście biocybernetyczne, które proponuje szersze spojrzenie na proces kontroli i regulacji. Dlatego celem rozprawy jest:

1. Interpretacja wybranego mechanizmu regulacji genów w ujęciu biocybernetycznym.
2. Rozwinięcie modelu TargetScore predykcji par interakcji miRNA/mRNA.

Autor niniejszego opracowania uważa, że poprawa jakości predykcji par interakcji miRNA/mRNA metodami obliczeniowymi, poprzez rozwinięcie jego biocybernetycznego modelu, zbliży nas do poznania faktycznego mechanizmu interferencji RNA.

Celem pracy jest zarówno teoretyczne opracowanie jak i praktyczna realizacja narzędzia bioinformatycznego. Dysertacja ma za zadanie zebranie aktualnej wiedzy na temat procesu regulacji genów. Wiedzę tę będziemy zbierać zwracając szczególną uwagę na najnowsze osiągnięcia w tej dziedzinie, ale także na stawianie odpowiednich pytań. Natomiast projekt i realizacja informatyczna modelu będą miały charakter praktyczny. Można o nich mówić jako o korzyści z przeprowadzonych naukowych eksploracji. Model ma wspierać identyfikację

(rozpoznawanie) dupleksów: transkryptów oraz miRNA - cząsteczek nukleotydowych w konkretnym biologicznym procesie regulacji genów na podstawie danych eksperymentalnych pochodzących z wysokoprzepustowych technik eksperymentalnych.

Do analizowanych danych eksperymentalnych zastosowano dwa podejścia: sekwencyjne i biocybernetyczne. Podejście sekwencyjne polega na przeprowadzeniu a priori zdefiniowanej sekwencji obliczeniowej. Zakłada ona, że rozpoznanie dupleksów można dokonać jedynie na podstawie informacji uzyskanej z wcześniejszego etapu, startując od informacji o sekwencji nukleotydowej. Drugie podejście, nazwane biocybernetycznym, próbuje wykorzystać sygnał wyjściowy, czyli w tym przypadku poziom ekspresji regulowanych genów, do rozbudowy istniejącego modelu rozpoznawania dupleksów.

Praca ma również wykazać możliwości adaptacji narzędzi i metod informatyczno-matematycznych w oryginalny sposób pozwalający na realizację założonego celu. Zakłada się również poszerzenie wiedzy z zakresu stosowania metod bayesowskich, szczególnie *Variational Bayesian-Gaussian Mixture Model* (VB-GMM), który został wykorzystany w modelu.

Część eksperymentalna projektu ma na celu dostarczenie danych do weryfikacji opracowanego modelu. Przeprowadzenie analizy tych danych pozwoli na zapoznanie się z metodami obliczeniowymi operowania na surowych danych, bezpośrednio uzyskanych z mikromacierzy DNA.

Teza rozprawy została sformułowana następująco: *Możliwe jest rozszerzenie modelu TargetScore do postaci proponowanego w pracy modelu biocybernetycznego biTargetScore, zbudowanie takiego modelu oraz jego walidacja na podstawie danych eksperymentalnych dotyczących sekwencji genomu człowieka uzyskanych z materiału klinicznego*.

1.3 Struktura rozprawy

Rozprawa wychodząc od przedstawienia zagadnienia biologicznego w rozdziale 2 "Biologiczne podstawy regulacji genów", zmierza do przedstawienia jednej z eksperymentalnych technik wysokoprzepustowych oraz wskazania nowych kierunków poszukiwań. Owo poszukiwanie będzie wiodło poprzez demonstrację przykładowych, istniejących rozwiązań bioinformatycznych służących do analizy zagadnienia, a zmierzać będzie do zastosowania wnioskowania Bayesa. W rozdziale 3 "Modele w analizie interakcji miRNA/mRNA" dochodzi do prezentacji autorskiego rozwiązania, poprawiającego jakość predykcji procesu regulacji. Owo autorskie rozwiązanie przedstawione będzie od strony koncepcyjnej w rozdziale 4 "Definiowanie modelu biocybernetycznego". Z kolei rozdział 5 "Opis implementacji" zawiera zaplecze informatyczne wykorzystane w analizie i przeprowadzonym wnioskowaniu. Rozdział 6 "Walidacja opracowanego modelu" opisuje testowanie wprowadzonego rozwiązania i porównanie go z innymi narzędziami. Rozdział 7 "Podsumowanie" zawiera wnioski z przeprowadzonych badań i perspektywy rozwoju poruszanego w rozprawie tematu.

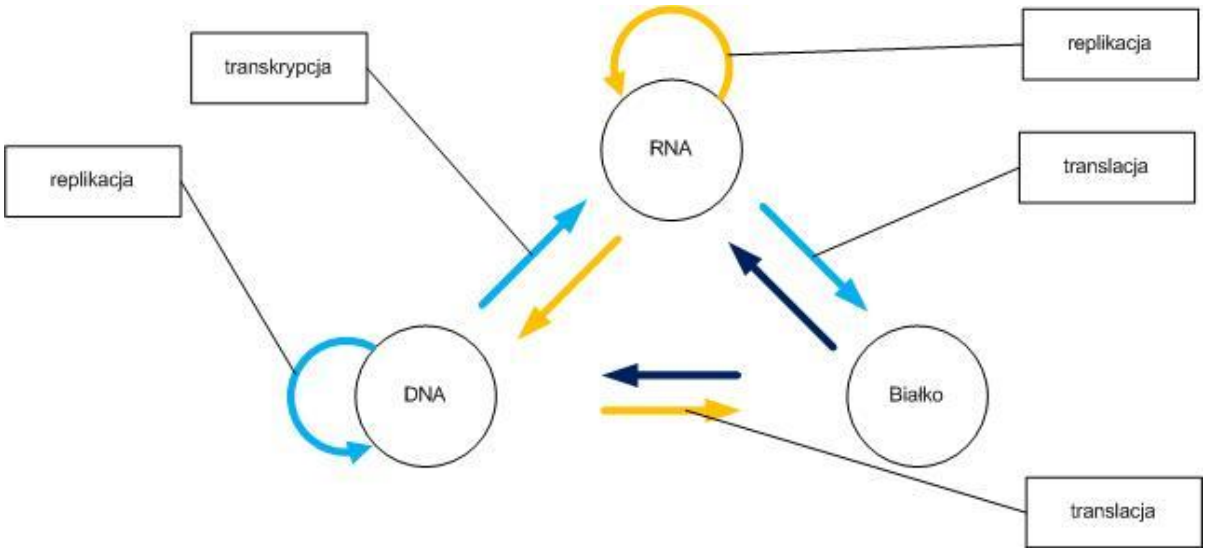
2 Biologiczne podstawy regulacji genów

2.1 Wprowadzenie

Biologia molekularna, jako poddziedzina biologii, poszukuje molekularnych podstaw życia. W biochemii poznajemy funkcje białek składających się na zespół cech organizmu –fenotyp, a dzięki genetyce poznajemy związek genów z fenotypem. W takim schemacie nie wolno zapominać o pośredniku tych zależności - cząsteczkach RNA, których rola jest informacyjna i regulacyjna, a być może (jak zakładają niektóre teorie) nawet pierwotna w interakcji żywego tworu ze środowiskiem [4].

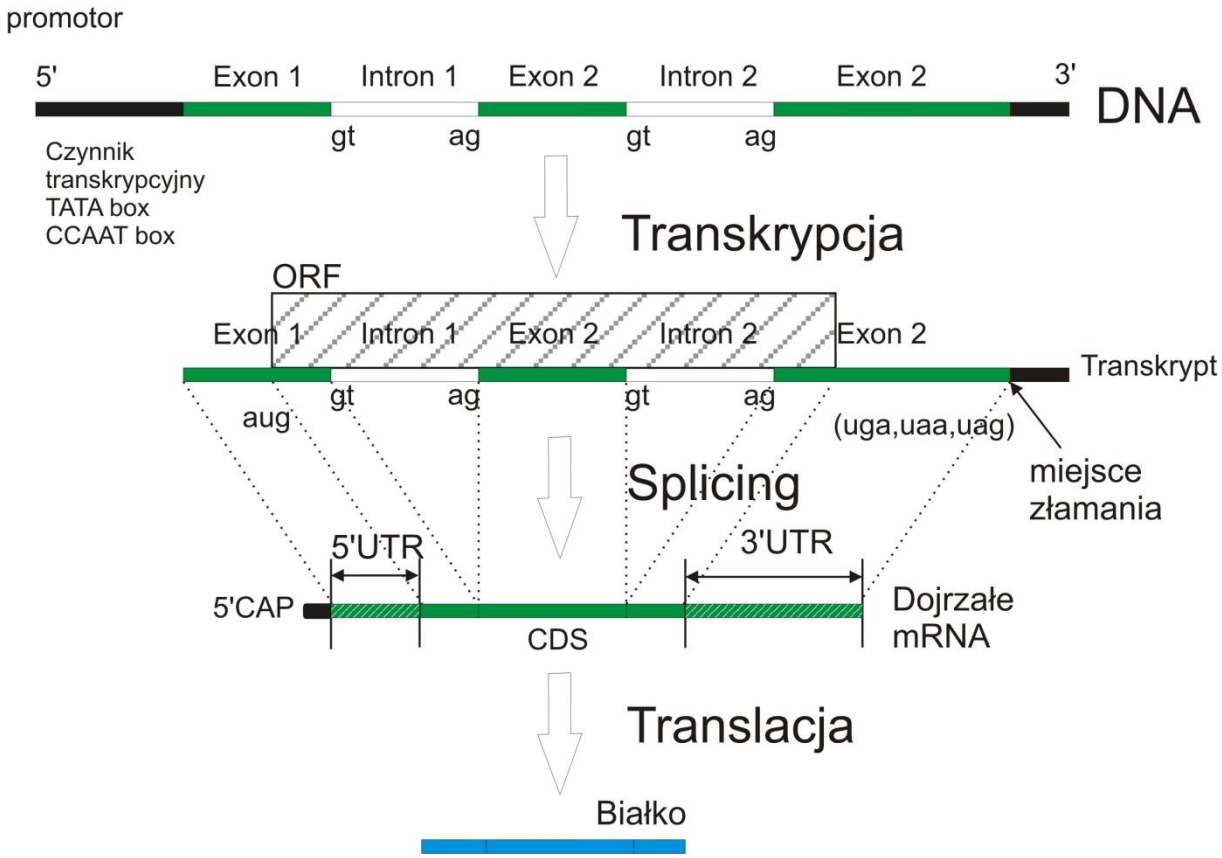
Jak ogólnie wiadomo, informacja genetyczna przechowywana jest w jądrze komórkowym na nośniku w postaci podwójnej helisy DNA. Informacja została tam zakodowana posługując się alfabetem czteroliterowym, gdzie każdej literze odpowiada jeden z czterech nukleotydów. Dowolność przechowywania informacji i jej odporność na zakłócenia gwarantuje dwuniciowa cząsteczka DNA powstała przez połączenie (hybrydyzację) dwóch nić na zasadzie komplementarności odpowiednich nukleotydów. Znając sekwencję jednego łańcucha można określić sekwencję drugiego przez zastosowanie reguł tworzenia par zasad azotowych. Właśnie ta własność została wykorzystana w procesach regulacji i kontroli procesów przetwarzania informacji biologicznej.

Przepływ informacji biologicznej (Rys. 2.1.) opisuje tak zwany centralny dogmat biologii. Przedstawia on sformalizowany schemat transmisji informacji między jej nośnikiem - źródłem i jej interpretatorem. Pierwszy zarys tego schematu ustanowił Francis Crick w 1956 r. [32]. Wraz z rozwojem wiedzy jest on stale uzupełniany. Obecnie oprócz kanonicznego kierunku przepływu informacji od DNA przez RNA do białka, uzupełniono go o niekanoniczne elementy np. retrotranskrypcję oraz przede wszystkim o elementy regulacji i kontroli, o czym będzie mowa w dalszej części. Najprostszy schemat przepływu informacji stanowi model jednokierunkowego kanału transmisji danych z wyróżnionymi procesami transkrypcji i translacji.



Rys. 2.1. Centralny dogmat biologii molekularnej. Niebieski kolor – kanoniczny przepływ informacji, żółty - niekanoniczny i czarny – przepływ informacji uzyskany jedynie laboratoryjnie.

Genetyka molekularna wprowadza szereg terminów dotyczących nazw przedmiotów, procesów czy funkcjonalności biologicznych. Najważniejsze pojęcie – **gen** - zidentyfikowano jako informację o budowie odpowiedniego białka zapisaną na polimerze kwasów nukleinowych (DNA), który ze względu na swoje właściwości fizyko-chemiczne funkcjonuje jako nośnik informacji. Analiza odpowiedniego regionu DNA zawierającego gen dostarczyła informacji o jego strukturze sekwencyjnej, która, jak wykazano, ma istotny związek z procesami odczytu tej informacji. Strukturę genu najlepiej wyjaśnia schemat przetwarzania informacji genetycznej, przedstawiony na rysunku Rys. 2.2.



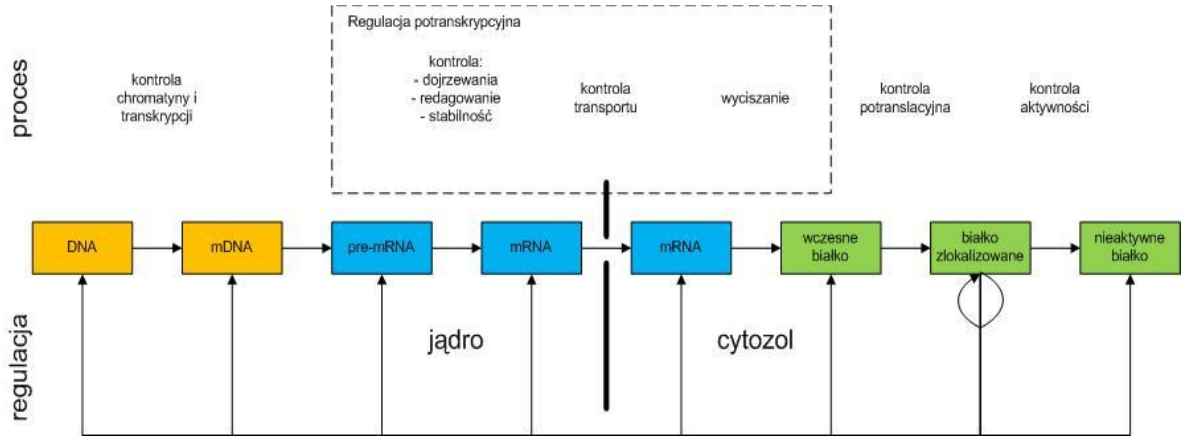
Rys. 2.2. Struktura genu na schemacie przepływu informacji biologicznej

Wyróżniamy tutaj trzy podstawowe procesy: **transkrypcja** – odczyt informacji z DNA, **splicing** – złożenie informacji o sekwencji aminokwasowej zakodowanego w genie białka, **translacja** – synteza na matrycy RNA cząsteczki białka. Na rysunku pojawiają się szczegółowe informację o strukturze i terminologii genu w kontekście jego przetwarzania: **promotor, czynniki transkrypcyjne, ekson, intron, ORF, mRNA, 5'UTR, CDS, 3'UTR.**

Pojęciem "regulacja genów" określa się czasową i lokalizacyjną kontrolę ekspresji i jakości genów. Z kolei pojęcie "kontroli genów" wykorzystuje się wobec procesów rozpoznawania błędów zaistniałych w procesach przetwarzania i transmisji informacji biologicznej. Analogicznie jak w rozwiązaniach technicznych występuje pewna prawidłowość – najłatwiej można kontrolować jakość informacji na podstawie jej finalnego produktu. Można też powiedzieć, że jakość regulacji jest tym większa im bliżej źródła zakłócenia następuje korekta informacji. Na przykład weryfikacja niepożądanych transkryptów na etapie potranslacyjnym powinna wpływać regulująco na etapy wcześniejsze, a nie tylko ograniczać regulacji do destrukcji gotowego produktu.

Umiejscowienie procesów regulacyjnych na schemacie przepływu informacji biologicznej (Rys. 2.3) pozwala wyróżnić regulację **transkrypcyjną** i **potranskrypcyjną** genów.

Regulacja i kontrola ekspresji genów może być realizowana na każdym etapie transmisji informacji biologicznej. Czas, miejsce, intensywność wykorzystania danego genu - kontrola transkrypcji, składanie (*splicing*) i dojrzewania (*processing*) pierwotnego transkryptu, redagowanie [10], kontrola transportu do cytoplazmy, system degradacji mRNA (*mRNA surveillance*) [83]: kontrola wierności i jakości transkryptów - kontrola potranskrypcyjnej, kontrola translacji i kontrola potranslacyjna, kontrola aktywności białka. Pomimo wielu poziomów i etapów regulacji podstawowa regulacja genów dotyczy kontroli transkrypcji przez czynniki transkrypcyjne, ponieważ jakość tego procesu niweluje zbyteczność nadmiarowych lub w ogóle niepożądanych transkryptów.



Rys. 2.3. Schemat przepływu informacji i regulacji genów. Kolor żółty – regulacja transkrypcji, kolor niebieski – regulacja potranskrypcyjna, zielony – regulacja potranslacyjna.

Przedstawiony rysunek (Rys. 2.3) dotyczy regulacji genu kodującego, czyli takiego, którego finalnym produktem jest białko. Symbolicznie przedstawiona strzałkami regulacja zawiera sprzężenia zwrotne zarówno dodatnie jak i ujemne. Ujemne pozwala utrzymywać odpowiedni poziom np. ekspresji genu. Przykładem dodatniego sprzężenia zwrotnego jest to, jak białko regulatorowe aktywuje transkrypcję własnego genu.

Regulacje i kontrole, które występują w dalszych etapach transmisji informacji biologicznej, stają się bardziej zrozumiałe, jeśli porównamy przedstawiony schemat do modelu transmisji informacji w telekomunikacji. Wymienione powyżej potranskrypcyjne metody kontroli przez analogię mogą być porównane do technicznych rozwiązań służących do weryfikacji informacji w odbiorniku. Przesył informacji i tej biologicznej i technicznej wiąże się z wpływem zakłóceń i przekłamań toru transmisji lub wręcz z pojawieniem się patologicznego sygnału. W tym kontekście rozpatrzona zostanie tzw. regulacja potranskrypcyjna, której niektóre elementy zostały już poznane. Zakłócenia biologiczne ogólnie zdefiniowane dotyczą reakcji niedostosowania się organizmu wobec zmieniających się warunków środowiskowych. Regulacja potranskrypcyjna obejmuje: kontrolę dojrzewania transkryptu, jego redagowanie, zapewnienie jego stabilności oraz system degradacji patologicznego transkryptu (mRNA surveillance), który z kolei obejmuje: wykrywanie przedwczesnego kodonu stopu (*Nonsense Mediated mRNA decay pathway*), detekcję braku kodonu stopu (*Nonstop Mediated mRNA decay pathways*) oraz blokowanie translacji (*No-go Mediated mRNA decay pathway*)[44]. Wreszcie ostatni element regulacji potranskrypcyjnej, będący przedmiotem dokładniejszej analizy w tej rozprawie, stanowi interferencja RNA.

Organizmy proste - w porównaniu do złożonych - mają podobną liczbę genów kodujących. Różnicę między nimi wykazuje natomiast złożoność układów regulacji, które pozwalają w bardziej różnorodny sposób wykorzystać tą podobną pulę genów. Elementy tego układu regulacji stanowią także cząsteczki mikro-RNA (miRNAs). Liczba wykrytych do tej pory miRNAs koreluje ze złożonością organizmu [92]. Wykorzystując dostępne publiczne zasoby zarejestrowanych cząsteczek RNA przeprowadzono próbę oszacowania całkowitej liczby miRNAs u człowieka [57]. Zrealizowano ten cel selekcjonując z całkowitej puli krótkich RNA odpowiednie pary komplementarne stanowiące fragmenty odpowiedniej struktury zgodnie z klasycznym przebiegiem biogenezy miRNAs. Odnaleziono około 1000 nowych miRNAs w uzupełnieniu do 1500 uprzednio poznanych. Część z nich, pomimo małego poziomu uśrednionej ekspresji, może podlegać dużej ekspresji, ale na poziomie pojedynczej komórki.

2.2 Mechanizm interferencji RNA

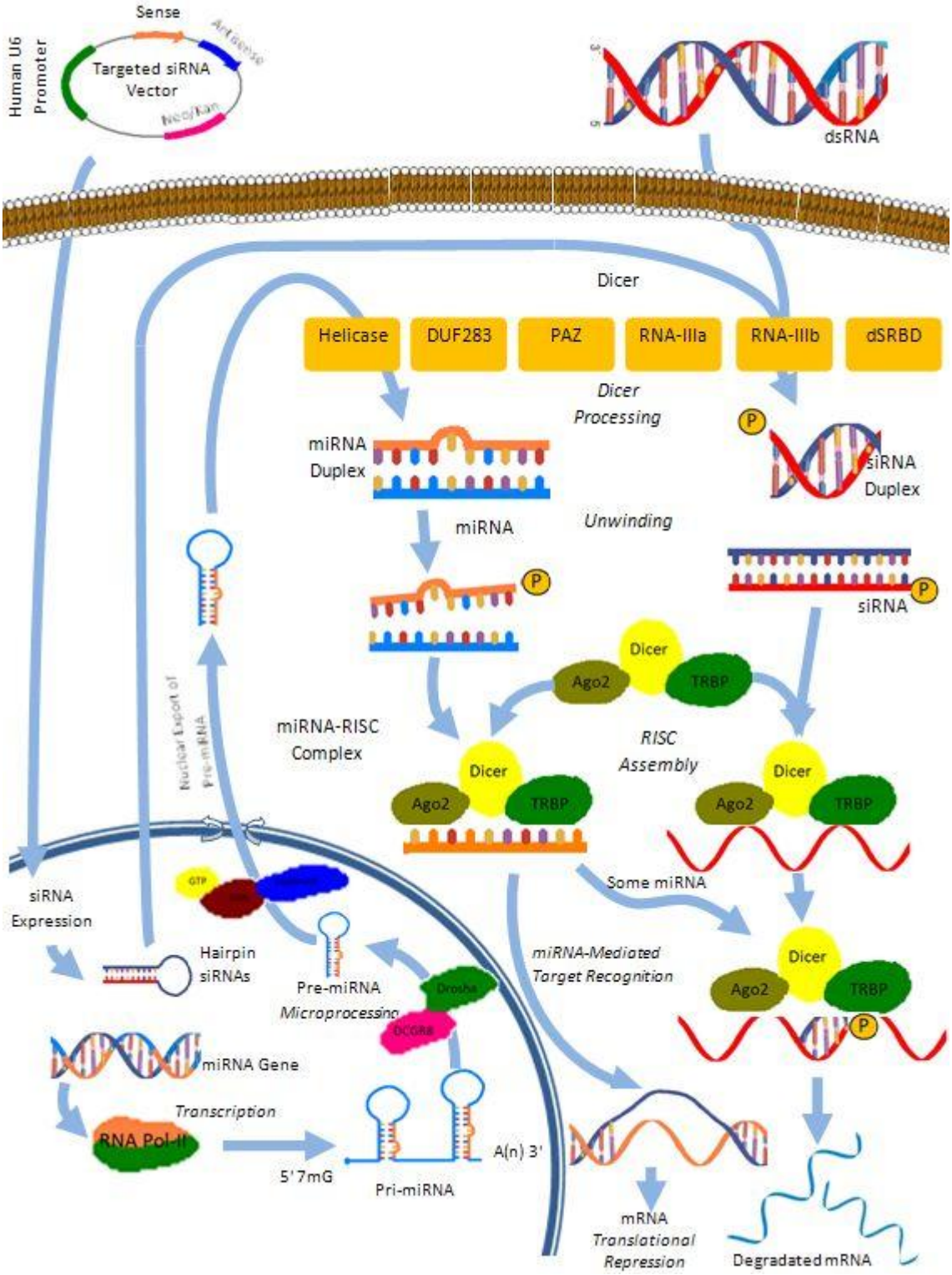
Interferencja RNA[74] (RNAi) to jeden z nowo poznanych procesów regulacji, zachodzących w żywej komórce. Proces ten nazwano tak, ze względu na jego efekt modulacji poziomu ekspresji genu, przez analogię do wzmacniania i osłabiania fal w fizyce. Do celu rozpoznania i tym samym regulacji konkretnego transkryptu wykorzystywane są krótkie cząsteczki RNA, wykazujące komplementarność względem docelowego transkryptu, z którym łączą się parując komplementarne zasady.

Przyjmuje się, że mechanizm RNAi występuje we wszystkich systemach biologicznych, chociaż są pewne wyjątki [143][2]. Powszechność tego mechanizmu potwierdza zatem jego istotność ewolucyjną. W porównaniu do eukariotycznego RNAi również i prokarioty posiadają funkcjonalny analog nazwany *Clustered regularly interspaced short palindromic repeats*, który najprawdopodobniej ewoluował niezależnie [152]. Jego podobieństwo do eukariotycznego RNAi polega na specyficzności sekwencyjnej krótkiej cząsteczki RNA oraz efekcie działania wyciszania transkryptów.

W publikacjach przedstawiających proces interferencji RNA u zwierząt możemy wyróżnić dwa zasadnicze podejścia. Pierwsze koncentruje się na tezie przedstawiającej proces interferencji RNA, jako metodę fizjologicznej regulacji genów wtórnej wobec regulacji transkrypcyjnej, jaka się odbywa przed lub w trakcie procesu translacji. Jej głównym celem jest kontrola poziomu ekspresji transkryptów. Drugie podejście traktuje ten proces, jako poszerzenie mechanizmów immunologicznych w organizmie na procesy wewnątrzkomórkowe, umożliwiające rozpoznanie patologicznej cząsteczki transkryptu. Są to więc procesy zmierzające w kierunku kontroli jakości i wierności transkryptu, nawiązujące tym samym do takich procesów zachodzących u roślin i u prokariotów, jak również do pierwotnego znaczenia ewolucyjnego RNAi.

Podstawowy mechanizm RNAi (Rys. 2.4) polega na wykorzystaniu krótkiej cząsteczki RNA do identyfikacji - na zasadzie komplementarności - docelowego transkryptu. Krótkie cząsteczki RNA o charakterystycznej sekwencji stanowią element rozpoznawczy w układach regulacji. Interakcja funkcyjnej cząsteczki RNA z transkryptem odbywa się zgodnie regułami termodynamiki. Dwie cząsteczki polinukleotydowe, jednoniciowe odnajdą takie wzajemne położenie, które gwarantuje największą ilość tworzących się tam par Watsona-Cricka. Zatem krótka cząsteczka polinukleotydowa „sama" odnajduje miejsce wiązania (*target site*) w długim polimerze nukleotydowym, tworząc z nim dupleks miRNA/mRNA. Funkcjonalność tej cząsteczki RNA

wspomagana jest przez kompleks białkowy *RNA-induced silencing complex* (RISC). Niemniej nieznany jest sam mechanizm lokalizacji targetu przez ten kompleks [150].



Rys. 2.4. Mechanizm interferencji RNA. (Rysunek zaczerpnięty z ilustracji RNAi Pathway (www.qiagen.com))

Zarówno w obrębie jednego organizmu, jaki i w różnych gatunkach, występuje wiele wariantów procesu RNAi. Ten fakt jest związany z pochodzeniem danej funkcyjnej cząsteczki RNA oraz różnych metod degradacji powstającego dupleksu RNA/mRNA. Ze względu na pochodzenie tych

cząsteczek endogenne lub egzogenne komórkowe RNAi wykazuje odmienne działanie wynikające bezpośrednio ze sposobu ich pozyskania. Wyróżniono następujące cząsteczki RNA uczestniczące w mechanizmie RNAi:

- mikroRNA (miRNA),
- *small interfering* RNA(siRNA),
- piwi-interacting (piRNA).

Przetwarzanie, stabilność i trwałość tych cząsteczek w cytoplazmie zapewnia kompleks białkowy *RNA-induced silencing complex* (RISC). Kompleks białkowo-nukleotydowy zawierający cząsteczkę miRNA określa się skrótem: miRISC (*miRNA-RISC*) lub miRNP (*microribonucleoprotein*). Przykładem egzogennego pochodzenia cząsteczki RNA jest penetracja komórki przez wirusa (w wyniku fuzji, wiropeksji, bezpośredniej penetracji) w procesie replikacji wirusa.

Endogenne cząsteczki RNA posiadają własne geny w jądrze lub w mitochondriach komórki gospodarza. Endogenne geny także mogą być przyczyną patogenezy (wbudowane w genom ludzki geny wirusowe, nieprawidłowe sygnały ekspresji genów, etc.).

Endogenna lub egzogenna dwuniciowa cząsteczka double – strand RNA (dsRNA) poddana zostaje przecięciu (przełamaniu) na odpowiedniej długości cząsteczkę (od 16- 28 nt) i następnie rozdzieleniu (rozwinięciu) przez kompleks białkowy DICER. Następnie, wiodąca nić, która zostaje rozpoznana na podstawie słabiej związanego końca [94], zostaje ulokowana w kompleksie białek RISC. Mocujące cząsteczkę RNA domeny białka Argonaut narzucają przyjęcie pewnych konformacji, które powstają po związaniu się cząsteczki z transkryptem.

RISC jest kompleksem rybonukleoproteinowym z zamocowaną jednoniciową cząsteczką RNA. Zamocowane w RISC'u służą jako wzorzec do rozpoznania specyficznego komplementarnego miejsca w rejonie transkryptu [75]. Pełny skład i warianty tego kompleksu nie są jeszcze poznane [156]. Podstawowe białka w kompleksie to Dicer – o aktywności rybonukleazy przełamuje dsRNA tworząc około 21 nukleotydowy jego fragment [185], białko Argonaut - aktywuje kompleks rozplatając dsRNA i wybierając jedną z cząsteczek tzw. wiodącą (sensowną): jego domena PAZ wiąże koniec 3' miRNA, domena PIWI – wiąże 5' miRNA.

Analiza inhibicji działania kompleksu RISC poprzez sprowokowane niedopasowania końców cząsteczki RNA czy to 5' czy 3' wykazała, że koniec 5' jest odpowiedzialny za dopasowanie i związanie targetu, podczas gdy koniec 3' jest odpowiedzialny za fizyczną aranżację targetu ułatwiającą przełamanie dupleksu w RISC'u [73].

Grupa fosforanowa występująca od strony 5' miRNA jest istotna przy pozycjonowaniu (wiązaniu) dojrzałego miRNA w białku Argonaut kompleksu RISC [56]. Stwierdzono jednak, że w obecności fosfatazy następuje defosforylacja końca 5' miRNA, która prowadzi do występowania w tej samej komórce mieszaniny obu postaci miRNA: ufosforylowanej i pozbawionej 5’-fosforanu [98]. Od strony 5' miRNA pojawiają się zatem różne izoformy jako konsekwencja różnic fosforylacji. Fakt ten jest istotny ze względu na to, że większość technologii określania profilu miRNAs w komórce nie jest wrażliwa na obecność lub brak tej właśnie grupy fosforanowej i jednakowo traktuje różne formy miRNAs. Zatem wśród metod określających profile miRNA trzeba zwrócić uwagę na te, które uwzględniają ten czynnik poprzez wprowadzenie etapu ligacji, który jest wrażliwy na występowanie tej grupy fosforanowej na końcu 5' miRNA.

Związanie się cząsteczki RNA z transkryptem, dokładnie - w jego części nazwanej *target site* wywołuje jedną z kilku poznanych metod represji procesu translacji. Począwszy od degradacji transkryptu a skończywszy na blokowaniu translacji. Degradacja transkryptu następuje po przełamaniu transkryptu przez AGO2 [85] – jedno z najważniejszych białek kompleksu RISC. Przełamanie zachodzi w przypadkach pełnej komplementarności między cząsteczką RNA a transkryptem. W przypadku niepełnej komplementarności dupleksu zachodzi blokowanie translacji [109]. Niepełna komplementarność może doprowadzać do deadenylacji końca transkryptu i w ten sposób przyśpieszać jego degradację [50]. Inny mechanizm doprowadza do zakłócenie inicjacji translacji [42].

Podaje się następujące metody wyciszania transkryptów [126]:

1. Hamownie transkrypcji przez reorganizację chromatyny zależnej od miRNA
2. Przełamywanie mRNA
3. Hamowanie inicjacji translacji
4. Hamowanie tworzenia kompleksu rybosomalnego
5. Hamowanie elongacji
6. Izolacja transkryptów w ciałkach P
7. Odłączenie rybosomu
8. Współudział w degradacji białka
9. Destabilizacja mRNA

Uznaje się, że podstawową funkcją RNAi u przodków ewolucyjnych była ochrona immunologiczna przed egzogennym materiałem genetycznym takim jak transpozony i genomy wirusów [25][20]. Przypuszcza się, że modyfikacja histonów również była obecna u przodków, ale regulację rozwoju organizmu przez RNAi uznaje się za nowy ewolucyjny nabytek [25].

Rola RNAi u ssaków w odporności wrodzonej jest słabo poznana. Hipoteza o zależności RNAi i nieswoistej odpowiedzi immunologicznej jest wspierana poprzez wskazanie genów wirusowych, których aktywność tłumi mechanizm RNAi [13][149][118][105][147]. Ta hipoteza znajduje także swoich przeciwników [33].

Największą różnice w funkcji biologicznej RNAi zaobserwowano między roślinami a zwierzętami. U roślin mechanizm ten stanowi istotny element odpowiedzi immunologicznej; reakcja wewnątrzkomórkowa na obcy materiał genetyczny – np. pochodzenia wirusowego [17]. U roślin miRNAs zasadniczo regulują czynniki transkrypcyjne [187]. Natomiast u zwierząt wykazano, że wyciszanie genów jako podstawowa funkcję miRNAs zachodzi w regulacji morfogenezy w tym także w komórkach macierzystych [158][24], różnicowaniu [154], namnażaniu komórek [21] i apoptozie [86].

Znane są przykłady zwierząt, u których RNAi stanowi przykład reakcji antywirusowej. Na przykład u muszki owocowej mechanizm RNAi stanowi element nieswoistej odpowiedzi immunologicznej, która zostaje aktywowana przez patogeny [184]. Podobnie u nicieni stwierdzono nadmiar białek Argonaut w odpowiedzi na infekcję. Wtedy następuje nadekspresja składników szlaku RNAi [177].

RNAi jako komponent antywirusowego wrodzonego systemu immunologicznego u wielu eukariontów koewoluował razem z wirusami. Niektóre wirusy rozwinęły mechanizm supresji odpowiedzi RNAi w komórkach gospodarza, szczególnie u wirusów roślinnych [115].

Porównanie funkcjonowania RNAi u zwierząt i roślin pozwala lepiej go zrozumieć. Najważniejsza różnica dotyczy znacząco różnego stopnia komplementarności cząsteczek miRNAs względem targetów. U zwierząt stwierdzono tylko częściową, około 7 nukleotydową komplementarność z miejscem wiązania w transkrypcie. Te siedem nukleotydów jest zlokalizowane na początku 5' miRNA. Znaczenie pozostałych ok 15 nukleotydów nadal nie jest poznane. Międzygatunkowa analiza porównawcza w obrębie królestwa zwierząt wskazuje na konserwatywność całej cząsteczki miRNA: części komplementarnej i tej nie-komplementarnej [89]. Badanie konserwatywności przeprowadzano także dla prekursorów miRNA (pre-miRNA) jak i pierwotnej formy miRNA (pri-miRNA). Wspólny mechanizm RNAi dla każdego miRNA, oraz różnorodność sekwencyjna w zbiorze miRNAs stanowi podstawę także dla statystycznego i całościowego opisu zbioru miRNA z uwzględnieniem relacji między elementami tego zbioru.

Niejasność mechanizmów RNAi staje się podstawą wielu hipotez. Przykładowe hipotezy wyjaśniające mechanizm interferencji RNA u zwierząt są następujące:

1. teza o regulacji tropizmu tkankowego patogennych ludzkich wirusów poprzez cząsteczki miRNA i wpływie miRNA na ewolucję wirusów [35], maksymalizacja replikacji wirionu [34];
2. cząsteczki miRNA miały znaczący wpływ na ewolucję transkryptów mRNA [157][55];
3. wirusy RNA, jak również pokswirusy (rodzina cytoplazmatycznych wirusów DNA) prawdopodobnie nie koduje wirusowych miRNA [35];
4. występowanie aktywującego mechanizmu regulacji potranskrypcyjnej komplementarnego do mechanizmu interferencji RNA [183];
5. częściową komplementarność miRNA z regulowanym transkryptem powoduje: sekwestracja transkryptów w ciałkach P (*processing bodies*), obniżony współczynnik elongacji podczas translacji, wyzwalanie deadenylacji docelowego transkryptu [134].

2.3 Funkcyjne cząsteczki RNA

Zaledwie 2% ludzkiego genomu poddane zostaje przepisaniu na RNA kodujące białko, podczas gdy 60-70% DNA jest przepisywane na niekodujące RNA [121]. Około 3% wszystkich ludzkich genów koduje miRNA i szacuje się, że ok 60% ludzkich genów jest targetami miRNAs [64].

W ramach interferencji RNA wyróżniono kilka typów cząsteczek funkcyjnych RNA. Są to *small interfering RNA* (siRNA), *micro RNA* (miRNA) i *piwi-interacting RNA* (piRNA). Ich porównanie [117] pozwala wykazać różnice w ich funkcjonowaniu (Tabela 1).

Pierwsze miRNA odkryto u nicieni *Caenorhabditis elegans* [99]. Ich funkcyjność w nowym, odrębnym mechanizmie parę lat później [140][132].

Spośród wymienionych cząsteczek RNA w mechanizmie RNAi u człowieka cząsteczki miRNAs stały się głównym obiektem zainteresowań naukowców ze względu na ich niską specyficzność w rozpoznawaniu targetów i zagadkowy mechanizm regulacji fizjologicznej, który posługuje się takimi cząsteczkami. Ten zaskakujący zwrot ewolucji RNAi dotyczył przejścia od wewnątrzkomórkowej nieswoistej odpowiedzi immunologicznej wykorzystującej pełną komplementarność siRNA względem targetów na mechanizm kontroli poziomu ekspresji genów z częściową komplementarnością par miRNA/mRNA.

Tabela 1. Zestawienie funkcyjnych cząsteczek RNA lokujących się w kompleksie RISC

	miRNA	siRNA	piRNA
występowanie	rośliny i zwierzęta	rośliny i niższe zwierzęta, u ssaków – jeszcze nie rozstrzygnięto	komórki płciowe zwierząt
struktura	pojedyncza nić	podwójna nić	pojedyncza nić
długość	17-25 nt	21-22 nt	25-31 nt
stopień komplementarności	częściowa, dlatego pojedynczy miRNA posiada setki targetów	prawie 100% dopasowanie, blokowanie specyficznych mRNA, z minimalnym efektem off-target	wysoki [119]
biogeneza	własne geny, introny, regulują geny inne niż te z których pochodzą	regulacja tego samego genu z którego pochodzi [101], wykorzystane w sygnalizacji międzykomórkowej (wówczas egzogenne)	tylko niektóre loci, pseudogeny [163]
działanie	hamowanie translacji mRNA	przełamywanie mRNA	
funkcja	regulacja poziomu mRNA	wyciszanie genu u roślin i zwierząt pozbawionych przeciwciał lub komórkowo zależnej immunologii	wyciszanie retrotranspozonów

Biogeneza cząsteczek miRNAs odbywa się w jądrze i cytoplazmie. Jak każda endogenna cząsteczka RNA posiada swój gen. Geny miRNA są transkrybowane za pomocą polimerazy RNA typu 2 [100]. Polimeraza często wiąże się z promotorem ulokowanym blisko sekwencji, która po transkrypcji przyjmuje strukturę spinki, czyli prekursora miRNA (pre-miRNA). Kilkuset nukleotydowy transkrypt *primary* miRNA (pri-miRNA) tworzy jedną lub wiele struktur spinek, z których każda może stać się na pre-miRNA.

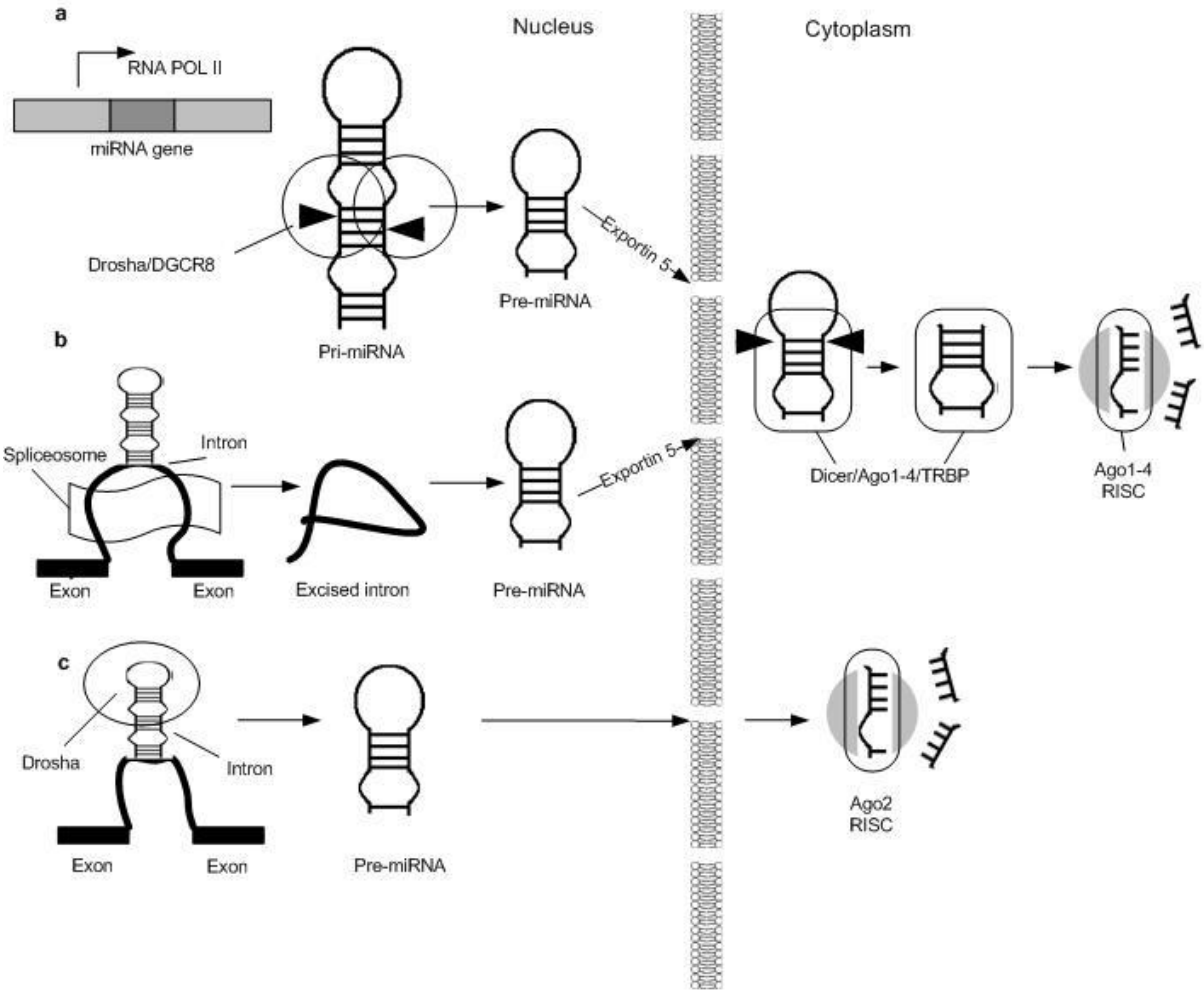
Etapy transkrypcji zależne są od kontekstu lokalizacji tego genu. Na podstawie bazy predykcji pri-miRNA [137] utworzonej w *Division of Molecular Pathophysiology Biocenter, Innsbruck Medical University* autor rozprawy przeprowadził zestawienie przynależności genów miRNAs (Tabela 2).

Dominującą grupę genów miRNA stanowią introny genów kodujących, następnie własne geny miRNA oraz geny, których produkt zostaje poddany degradacji w mechanizmie *nonsenese mediated decay*.

Regulacja genów miRNA związana jest z ich lokalizacją. Geny miRNA wewnątrz intronów podlegają „automatycznej" translacji podczas aktywacji genu gospodarza. Samodzielne geny albo bardzo często ich zgrupowanie powinny posiadać własne sekwencje promotorów oraz własne czynniki transkrypcyjne. Rozpoznanie tych genów stanowi większą trudność bioinformatyczną ze względu na brak charakterystycznej struktury w jaką są wyposażone geny kodujące.

Tabela 2. Zestawienie genów miRNA. Typ transkryptu, położenie miRNA w stukturze transkryptu: egzon, intron. Znaczenie typów transkryptów: antisense – transkrypt pochodzi z komplementarnej względem kodującej nici DNA, miRNA – własny gen, protein coding – pri-miRNA pochodzi z fragmentu transkryptu kodującego, nonsense mediated decay – transkrypt kodujący wywołujący proces kontroli przedwczesnych kodonów stopu, Processed transcript – transkrypt nie zawierający sekwencji kodującej ani też nie jest sekwencją intronu, Retained intron – transkrypt zawierający sekwencje kodującą zawierającą także zachowany intron, ambigous orf - Ambiguous Open Reading Frame - niekodujący transkrypt potencjalnie kodujący białko, z więcej niż jedną ramką odczytu, Processed pseudogene – przetworzony transkrypt pseudogenu, lincRNA - long intergenic non-coding RNA, sense intronic – niekodujący transkrypt wewnątrz intronu, snoRNA - małe jąderkowe RNA, unprocessed pseudogene – nieprzetworzone pseudogeny, non coding – niekodujące RNA, unitary pseudogene – rodzaj pseudogenu, ncrna host – klaster niekodujących RNA, sense overlapping – niekodujący transkrypt zawierający gen kodujący białko wewnątrz jego sekwencji intronicznej na tej samej nici, nie zachodzącej na sekwencje egzonu, Transcribed unprocessed pseudogene – rodzaj pseudogenu pochodzącego z nieprzetworzonego transkryptu, transcribed processed pseudogene – rodzaj pesudogenu z przetworzonego transkryptu, IG V gene – składowe geny immunoglobuliny.

typ transkryptu	l. premirna_id	l. pri-miRNA	exonic	intronic
antisense	25	43	2	24
miRNA	695	696	686	0
protein_coding	707	2666	58	659
nonsense_mediated_decay	191	319	6	180
processed_transcript	317	660	19	300
retained_intron	142	204	21	123
ambiguous_orf	1	1	0	1
processed_pseudogene	5	7	5	0
lincRNA	43	94	7	37
sense_intronic	5	5	1	4
snoRNA	7	7	1	0
unprocessed_pseudogene	5	5	0	5
non_coding	17	23	8	11
unitary_pseudogene	2	2	0	2
ncrna_host	5	5	5	0
sense_overlapping	1	11	0	1
transcribed_unprocessed_pseudogene	4	7	0	4
transcribed_processed_pseudogene	1	1	1	0
IG_V_gene	1	1	0	0
SUMA	2174	4757	820	1351



Rys. 2.5. Proces biogenezy cząsteczek miRNA dla trzech różnych typów genów. a) Własny gen transkrybowany przez polimerazę RNA II. pri-miRNA zostaje przetworzone przez kompleks Drosha(RNase III) i DGCR8 w pre-imRNA, który następnie zostaje transportowany do cytoplazmy przez Exportin 5. b) Gen introniczny. Zestaw białek spliceosom realizujący wycinanie intronów realizuje także wycięcie pre-miRNA. Po rozpoznaniu przez Exportin 5 zostaje przetworzone przez kompleks DICER. c) Wycinanie pre-miRNA za pomocą kompleksu Drosha bezpośredno pre-miRNA. (Rysunek wzorowany na ilustracji z pozycji [49]).

Uzyskany na podstawie genu transkrypt pri-miRNA poddany zostaje przetworzeniu do postaci pojedynczej spinki do włosów (*hairpin*), czyli pre-miRNA. W tej formie zostaje on przetransportowany do cytoplazmy, gdzie zostaje poddany dalszej obróbce przez kompleks białkowy RISC. Kompleks DICER rozpoznaje cząsteczka dominującą, która po wydzieleniu z dsRNA zostaje zamocowana w kompleksie RISC (Rys. 2.5).

Matryca genowa nie determinuje jednak ostatecznej struktury sekwencji miRNA. Szacuje się, że ok 6% ludzkiego miRNA podlega procesowi edycji RNA [124]. Konsekwencją tej edycji jest przyporządkowanie wielu różnych miRNA temu samemu prekursorowi.

2.4 Interakcja miRNA/mRNA

Różnica między roślinnym i zwierzęcym mechanizmem RNAi dostarcza istotnych faktów dotyczących mechanizmu interakcji miRNA z docelowym transkryptem. Do jej zrozumienia potrzebna jest znajomość struktury miRNA i powstania jej w ramach filogenezy. U roślin najczęściej następuje degradacja transkryptów docelowych – targetów z powodu niepełnej komplementarności pomiędzy miRNA a sekwencją jego targetu. Ewolucyjnie duplikacja regionu

antysensownej nici genu targetu stanowi podstawę dla powstania genu miRNA. Transkrypcja tego regionu i przetworzenie transkryptu przez RNAi na cząsteczkę siRNA w efekcie daje komplementarną sekwencję do odpowiedniego genu. Cząsteczki o takiej genezie dysponują jednym lub kilkoma targetami. Konsekwencją jednak takiej genezy genu miRNA jest konieczność koewolucji pary miRNA-mRNA.

U zwierząt ze względu na mały stopień komplementarności stwierdzono, że parowanie zasad występujących u funkcyjnych RNA dotyczy przede wszystkim pierwszych 7-8 nukleotydów od strony 5'. Ten fragment miRNA nazwano *seed*. Ponieważ specyficzność *seed* jest bardzo niska, łatwo przewidzieć, że potencjalnych targetów dla danego miRNA może być wiele. Oprócz tego w obrębie *seed* stwierdzono także pewne odstępstwa od pełnej komplementarności, które powodują wybrzuszenia zauważalne na poziomie strukturalnym. Ta częściowa komplementarność świadczy o konformacyjnym dopasowaniu, a nie tylko "sekwencyjnym" powstającego dupleksu [7]. Dalsze prace dotyczące rozpoznania miejsca lub miejsc wiązań w obrębie transkryptu dostarczyły szereg dodatkowych czynników, które współwystępują w miejscach wiązań się RISC'u z targetem. Są to:

1. klasyfikacja rodzajów *seeds* (8mer, 7mer-m8, 7mer-1A) i wynikająca z nich: *seed-pairing stability* (SPS), która jest funkcją koncentracji nukleotydów: adeniny i uracylu [62];
2. filogenetyczny konserwatyzm lub jego brak cząsteczek miRNA [59];
3. filogenetyczny konserwatyzm lub jego brak miejsc wiązania (*target sites*);
4. wkład regionu 3' miRNA w stabilność dupleksu [71];
5. wkład par AU (w otoczeniu dupleksu) na nici transkryptu [102];
6. ilość miejsc wiązania w obrębie transkryptu;
7. udział pozycji miejsca wiązania: odległość do najbliższego końca oznaczonego UTR targetu.

Obecnie nie wiadomo, jaka część interakcji miRNA/mRNA stosuje się do wymienionych reguł, szczególnie, jeśli miejsce wiązania wypada poza rejonem 3'UTR. Pewne prace [166][130][139] potwierdziły występowanie miejsc wiązania poza obszarem 3'UTR: w rejonie kodującym – CDS (które jest charakterystyczne dla roślin) oraz w rejonie 5'UTR. Narzędzia bioinformatyczne wykorzystując podane reguły wprowadzają ograniczenia na predykcję targetów i nie są pomocne przy odkrywaniu nowych targetów, które być może podlegają innym ograniczeniom czy wymogom.

Złożoność problemu interakcji miRNA/mRNA wynika z możliwości uczestnictwa RNAi w następujących procesach:

1. degradacja transkryptów mRNA zachodząca w sposób niepożądany, nadmiarowy, podobnie jak przy reakcji komórki na egzogenne (np. wirusowe) cząsteczki RNA;
2. nadmiarowość transkryptów mRNA wywołana przez zaburzenia ich regulacji transkrypcyjnej, czy też słabo poznane procesy ich degradacji;
3. możliwość interakcji kompetycyjnej (rywalizacji) między miRNA a egzogennymi RNA;
4. nieprawidłowości w mechanizmie regulacji wywołane przez niekorzystną mutację sekwencji albo niekorzystną konformację białka kompleksu Drosha lub RISC;

5. fizjologiczna regulacja potranskrypcyjna genu miRNA poprzez inhibitory translacji mRNA w różnych stadiach rozwoju i różnicowania organizmu.

2.5 Mikromacierze DNA

Wyznaczenie profilu ekspresji miRNAs w komórce jest często podejmowanym badaniem ze względu na współudział cząsteczek miRNAs w wielu istotnych procesach życiowych. Można wyróżnić trzy metody wykorzystywane w celu wyznaczenia profilu: ilościową reakcję PCR czasu rzeczywistego - *real-time quantitative PCR* (qRT-PCR) [27], metody bazujące na hybrydyzacji (mikromacierze DNA) [111][103] i metody wykorzystujące sekwencjonowanie nowej generacji (równoległe): *next-generation sequencing* (NGS) [127] lub RNA-seq. Spośród nich w przeprowadzonych dalej obliczeniach w tej rozprawie wykorzystano dane pochodzące z eksperymentu mikromacierzowego DNA. Mikromacierze DNA wykorzystywane są do identyfikacji genów powiązanych z chorobą, oceny zmienności ekspresji w grupach genów w odpowiedzi komórkowej na lek, infekcję czy egzogenny materiał genetyczny.

Technika mikromacierzy DNA pozwala na równoczesny (przy użyciu jednej mikromacierzy) pomiar ekspresji całego zbioru ludzkich (lub zwierzęcych) miRNAs, ze względu na relatywnie niską liczbę różnych miRNAs (ok 2500) w porównaniu do różnorodności mRNAs (ok 30000). Aktualnie oferowane są następujące platformy mikromacierzy miRNA: Affymetrix, Agilent, Exiqon, Life Technology i Illumina. Macierze te obejmują już większość albo wszystkie formalnie uznane miRNAs (na podstawie bazy miRBase) wielu biologicznych gatunków w tym oczywiście i ludzkie. Oprócz dojrzałych miRNAs oferowane są też pola (*spots*) na mikromacierzach wykrywające prekursory pre-miRNAs. Pionierską metodę udoskonalenia mikromacierzy miRNAs oferuje firma Exiqon. Zastosowana tam mikromacierz CURY LNA wykazuje wysoką czułość i swoistość nawet dla miRNA bogatych w adeniny i tyminy, jak również dużą powtarzalność na poziomie 99% korelacji między macierzami [51]. Wynika to z zastosowania modyfikowanych chemicznie nukleotydów (*locked nucleotides* - LNA), których hybrydyzacja charakteryzuje się większą stabilnością termodynamiczną.

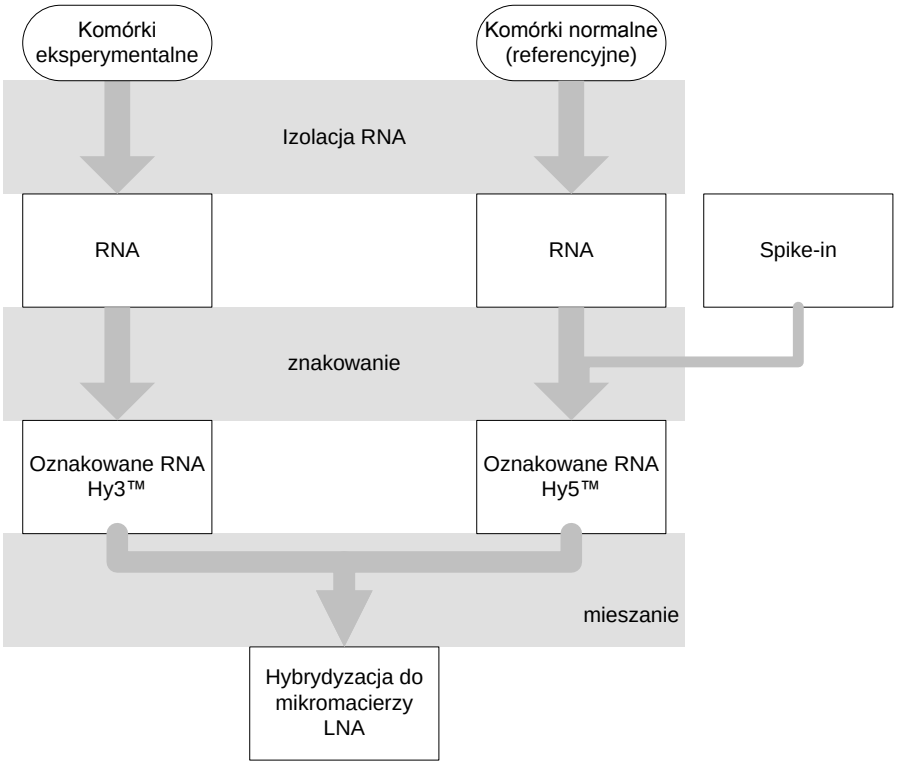
Do oceny ilościowej zawartości miRNAs w próbce metoda wykorzystuje hybrydyzację miRNA, poddanych uprzednio odwrotnej transkrypcji do ich deoksynukleotydowych odpowiedników, z selektywnie rozpoznanym fragmentem nukleotydowym, jednym z pośród wielu różnych znajdujących się na płytce mikromacierzy.

Eksperymenty z użyciem mikromacierzy posługują się płytkami, matrycami – regularną siatką zawierającą pola (*spots*). Każde takie pole zawiera wiele cząsteczek nukleotydowych o takiej samej sekwencji - sondy hybrydyzujące (*probe*) np. oligonukleotyd DNA, który jest starannie dobrany w ten sposób, aby hybrydyzował wyłącznie z konkretną cząsteczką RNA.

Mikromacierze DNA ze względu na metodę porównywania prób biologicznych dzielimy na dwukanałowe (dwukolorowe) i jednokanałowe (jednokolorowe). Ze względu na niską powtarzalność produkcji mikromacierzy (szczególnie mikromacierzy cDNA) nie jest możliwe porównywanie bezwzględnych wartości zmierzonego natężenia fluorescencji. Dlatego wprowadzono metodę dwukolorowego znakowania fluorescencyjnego. Wówczas każde pole macierzy hybrydyzuje z dwoma próbami: referencyjną oraz porównywaną. Uzyskane dwie wartości natężenia pozwalają wyznaczyć wartość względną poziomu ekspresji, jako iloraz

natężenia fluorescencji. Wartość ta nie powinna zależeć od wielkości pola mikromacierzy, czyli liczby związanych tam sond.

Przykładowy schemat przeprowadzenia eksperymentu dwukanałowego przedstawiono na Rys. 2.6. Metoda dwukanałowa pozwala od razu na jednej mikromacierzy DNA uzyskać względne wartości ekspresji porównywanych prób. Wykorzystuje ona dwa rodzaje barwników: czerwony Cyanina Cy3 lub jego odpowiednik Hy3 oraz zielony - Cyanina Cy5 lub Hy5 do znakowania fluoroscencyjnego prób. Wyekstrahowane z obu prób RNA jest przetwarzane osobno po to, aby wykorzystać barwniki niezależnie. Próba referencyjna jest znakowana podczas odwrotnej transkrypcji przy pomocy Cy3, a druga - eksperymentalna Cy5. Po wymieszaniu oznakowanych prób następuje ich hybrydyzacja do mikromacierzy. Po jej zakończeniu mierzy się poziom fluorescencji po uprzednim naświetleniu (pobudzeniu) laserem. Iloraz zmierzonych intensywności promieniowania powinien być niezależny od ilości sond mieszczących się na polu.



Rys. 2.6. Uproszczona procedura eksperymentu z mikromacierzą DNA oligonukleotydową, dwukanałową miRCURY LNA™ microRNA Array (7th Gen)

Populacja cząsteczek miRNA pochodzących z próby może zostać zakłócona przez hybrydyzacje miRNA z cząsteczkami cDNA lub nić sensowną powstałą podczas amplifikacji PCR przeprowadzonej na miRNAs. Powodem jest enzymatyczna własność odwrotnej transkrypcji lub amplifikacji [28]. Liniową zależność intensywności fluorescencji w funkcji koncentracji miRNA obserwuje się w przedziale dwóch rzędów wielkości ekspresji, co wystarcza do wyznaczania profili ekspresji miRNA, natomiast nie jest wystarczająca do porównywania poziomu ekspresji różnych miRNA w obrębie jednej próby. Służy to więc jedynie do porównywania ekspresji między próbami w różnych warunkach fizjologicznych. Jest to spowodowane efektywnością wiązania się miRNA z jego sondami DNA [108].

Trudność analizy miRNA na mikromacierzach związana jest z niską masą cząsteczkową w porównaniu do mRNA, ich małą długością i wysoką heterogenicznością - wysoka wariancja par GC. Ta heterogeniczność prowadzi do hybrydyzacji o różnym stopniu efektywności u różnych miRNA. Stanowi to utrudnienie przy równoczesnym pomiarze stężenia całej populacji miRNAs na jednej mikromacierzy.

Oligonukleotydowe mikromacierze wykorzystują też metody ułatwiające normalizację uzyskanych wyników swojej analizy. Specjalna pula krótkich cząsteczek RNA o znanych sekwencjach oraz znanej liczebności tzw. *RNA spike-ins* dodawana jest do wymieszanych prób. Zmierzony poziom hybrydyzacji tych cząsteczek z polami kontrolnymi pozwala na ustalenie rzeczywistego poziomu stężenia RNA w próbach.

Metoda jednokanałowa wymaga osobnych mikromacierzy w celu wyznaczenia względnych poziomów ekspresji. W tej technice wykorzystuje się tylko jeden znacznik. Zalety jakie się podaje dla tej metody to: brak wpływu fizycznego i chemicznego na etapie hybrydyzacji między porównywanymi próbami, prostsze porównywanie wyników z różnych eksperymentów, dla dużej liczby prób biologicznych – potrzeba mniejszej liczby mikromacierzy.

Podaje się następujące problemy techniczne stosowania mikromacierzy [52]: różny stopień intensywności fluorescencji zależny od użytego barwnika w metodzie dwukanałowej, różna efektywność oznaczania różnych miRNA zależna od jego sekwencji, ograniczenie stosowania mikromacierzy tylko do porównywania częściowo różniących się prób ze względu na utrudnioną procedurę normalizacji, nukleotydy hybrydyzujących cząsteczki łączą się także w inne oprócz Watsona-Cricka pary, duże odchylenie standardowe temperatury topnienia dwuniciowych DNA.

Podana problematyka oraz błędy techniczne powodują, że opracowanie danych mikromacierzowych wymaga wykorzystania metod statystycznych przy normalizacji, korekcji tła, itd., ale też odpowiedniego podejścia eksperymentalnego: wielokrotne próby tej samego materiału biologicznego, czy także stosowanie innych technik kontrolujących uzyskane wyniki.

2.6 Metody walidacji par miRNA/mRNA

Od czasu stwierdzenia zaledwie częściowej komplementarności sekwencyjnej cząsteczek dupleksu miRNA i targetu, predykcja par miRNA/mRNA za pomocą narzędzi bioinformatycznych powinna zostać potwierdzona eksperymentalnie, aby uniknąć fałszywie pozytywnych predykcji, które wynikają właśnie z tego braku jednoznaczności hybrydyzacji dwóch polinukleotydów.

Przed procedurą eksperymentalnej walidacji wykorzystuje się szereg programów predykcji miejsc wiązania w obrębie targetu – mRNA. Dopiero wówczas dupleksy miRNA/mRNA zostają potwierdzane eksperymentalnie [131][168]. Podstawowe stosowane metody to oznaczenie kodowanych białek metodami *reporter assays* [38] i *western blot* [78]. Metoda *reporter assays* opiera się na stwierdzeniu, że miRNA wiążąc się do miejsca wiązania specyficznego mRNA powoduje represję produkcji białka reporterowego, którego poziom może być łatwo porównany z próbą kontrolną. W tej metodzie przygotowuje się specjalny konstrukt przeznaczony do wprowadzenia do komórki (transfekcji), jako fuzję regionu targetu lub tylko jego miejsca wiązania z wektorem (reporter). Metoda pozwala na ocenę ilościową targetu przez pomiar aktywności reportera. Jako reportery najczęściej wybierane są: lucyferaza, GFP [26], YFP, lacZ/ β-galactosidase.

Podejście umożliwiające analizę funkcjonalną par miRNA/mRNA polega na wywołaniu przejściowej nadekspresji poprzez podstawienie "sobowtóra" (*mimic*) danego miRNA w komórce, która represjonuje produkt domniemanego targetu. Następnie zostaje to potwierdzone analizą immunoblotu - *western blot* [22], w której po rozdzieleniu w elektroforezie żelowej proteomu, rozpoznanie białka na membranie po jego przeniesieniu z żelu dokonuje się przy użyciu swoistego przeciwciała.

Eksperymenty takie jak *northern blot* [169], *quantitative real-time PCR* (qRT-PCR) oszacowują ilość całkowitego RNA wyizolowanego z próby i weryfikują koekspresję miRNA i mRNA. Eksperymenty skanujące cały transkryptom wykorzystują także mikromacierze DNA z nadeksperesją lub blokowaniem miRNA, znakowaniem izotopem aminokwasów w hodowli (SILAC) lub *Pulsed SILAC* (pSILAC) (*stable-isotope labeling by amino acids in cultured cells*) [72].

Metody eksperymentalne opierają się na analizie transkryptomu, analizie biochemicznej i analizie proteomu. Jedno z kryterium wyboru odpowiedniej metody związane jest ze spodziewaną liczebnością analizowanych targetów. Dla ich dużej liczby wybiera się metody transkryptomiczne lub proteomiczne, które rejestrują stan komórkowy po neutralizacji konkretnego miRNA lub po wywołaniu jego nadekspresji. Wadą takiego podejścia jest trudność rozróżnienia między bezpośrednim lub pośrednim wpływem danego miRNA na ekspresję danego targetu. Aby zlikwidować możliwość pośredniego wpływu miRNA w metodzie rejestrowania targetów stosuje się metodę immunoprecipacji AGO lub RISC w połączeniu z metodami mikromacierzowymi lub *deep sequence analysis* całkowitego RNA wyizolowanego z immunoprecipacji w obecności i podczas nieobecności danego miRNA.

Weryfikacja interakcji miRNA-target odbywa się przez eksperymentalną manipulację poziomem miRNA czy to przez jego nadekspresję, brak ekspresji czy podeksprejsję w liniach komórkowych lub fragmentach tkanek. Tego typu materiał biologiczny już na wstępie można przyjąć, że nie podlega kontroli całościowej fizjologicznej, sygnalizacyjnej międzykomórkowej, więc może wykazywać niefizjologiczną regulację także (albo przede wszystkim) na poziomie interferencji RNA (zmieniony profil miRNA). Prawdopodobnie też stąd bierze się łatwość uzasadnienia w metodologii biologicznej znaczenia aktywności niektórych miRNA w rozwoju organizmu.

Metody eksperymentalne weryfikacji targetów nie udzielają bezpośredniej odpowiedzi na pytanie, które z pośród potwierdzonych interakcji miRNA/mRNA występują w endogennych, fizjologicznych czy patologicznych warunkach całego organizmu. Podobnie jak w procesie regulacji genów przez czynniki transkrypcyjne, endogenne miRNA mogą wymagać odpowiednich komórkowych czynników do wiązania i regulacji swoich targetów. Tymczasem te czynniki nie muszą być zapewnione w warunkach eksperymentalnych weryfikacji [138]. Mając świadomość tej problematyki, ograniczoności stosowanej metodologii autorzy jednej z baz [181] wprowadzili kategorie różnicujące eksperymenty wg różnych kryteriów. Dokonując metaanalizy publikacji odnoszących się do eksperymentów walidacyjnych rozróżniono tzw. endogenne oraz egzogenne eksperymenty miRNA.

Wzrastająca liczba badań eksperymentalnych dostarcza informację o endogennych interakcjach miRNA-mRNA [41][4]. Wykorzystują one np. fuzję regionu 3'UTR domniemanego targetu z innym genem, który później w obecności ekspresji endogennego miRNA zostaje poddany (lub nie) represji.

Egzogenne miRNA eksperymenty mogą być ogólnie podzielone na dwie kategorie opierając się na metodach manipulacji poziomem miRNA: nadekspresja miRNA lub brak ekspresji, oraz eksperymentów podekspresji. Te pierwsze wykorzystują transfekcję dojrzałym miRNA [87], transfekcję prekursorem miRNA [61], pośrednio indukowaną nadekspresją miRNA(używając *DNA demethylating agent 5-Aza-Deoxycytidine*) [116], lub inhibitor deacetylazy histonowej – kwas fenylomasłowy - *phenylbutyrate* (PBA) [145]. Techniki wywołujące podekspresję miRNA obejmują zablokowanie miRNA (*knock-down*) przez siRNAs [128], zablokowanie miRNA przez antysensowny modyfikowany oligonekleotyd [179], *locked nucleic acids* (LNAs) [47], lub *2'-O-Me oligonucleotides* [97] oraz blokowanie genu miRNA [171].

Wprowadzona kategoria eksperymentów oznacza, że kiedy element wykonawczy oddziałuje bezpośrednio i celowo na poziom ekspresji miRNA mamy eksperyment egzogenny. Endogenny - gdy manipulujemy bezpośrednio targetami.

Inna przedstawiana w literaturze klasyfikacja przydziela eksperyment do jednej z trzech grup: poziom genowy (*target gene level*), poziom rejonowy (*target region level*)i poziom miejsca wiązania (*target site level*). Kiedy eksperyment wykazuje, że poziom pełnej długości produkt genu (mRNA lub białko) domniemanego targetu uległ redukcji w następstwie nad lub braku ekspresji miRNA, lub gdy pełnej długości produkt genu akumulował się po podekspresji miRNA, wtedy uznawany jest eksperyment poziomu genowego. Ta kategoria przede wszystkim obejmuje eksperymenty egzogenne, ponieważ tam poszukujemy ujemnej korelacji pomiędzy poziomem manipulowanego miRNA a pełnej długości produktami domniemanego targetu. Jest ona często uważana za pośredni dowód interakcji miRNA/mRNA, ponieważ poziom produktu genu może zmieniać się z innych powodów np. zmian w ekspresji innych białek, które są rzeczywistymi targetami danego miRNA.

Poziom regionu przypisujemy dla przypadków, kiedy region mRNA domniemanego targetu (krótszy niż pełnej długości transkrypt) jest odpowiedzialny za interakcję miRNA-target. Większość tego typu eksperymentów jest przeprowadzanych z wykorzystaniem fuzji regionu 3'UTR domniemanego targetu oraz odpowiedniego reportera, w następstwie czego obserwuje się, że ekspresja reportera zmniejsza się, lub zwiększa w odpowiedzi na nadekspresję miRNA, podekspresję miRNA lub jej brak.

Kategoria *target site level* dotyczy bardzo krótkich fragmentów mRNA (o porównywalnej długości z miRNA). Eksperymenty *reporter assays* wykorzystują fuzję konstruktów zrobionych z bardzo krótkich sekwencji miejsc wiązania oraz zmutowanych miejsc wiązania.

Kolejna kategoria rozróżnia eksperymenty ze względu na przedmiot pomiarów poziomu ekspresji domniemanych targetów: *reporter assays*, pomiar poziomu mRNA (PT-PCR, Northern blot, 5'RACE, mikromacierze DNA, *ribonuclease protection assay*, *Branched DNA probe assay* [29]), pomiar poziomu białka (*western blot*, ELISA, immunocytochemia) i inne metody analizy ekspresji targetów, np.: analiza fenotypowa, metoda TRAP[23] wykorzystująca dominującą negatywną mutację podjednostki RISC, następnie wychwytywanie par miRNA/mRNA i ograniczenie dalszego ich przetwarzania. Zapewnia ona, że nawet małe poziomy par miRNA i targetów zostaną wytrącone podczas immunoprecypacji RISC. Następnie zastosowanie metody NGS pozwala zidentyfikować nieznane targety, a metody qRT-PCR – potwierdzić te znane.

Metoda 3'LIFE (*Luminescence-based Identification of Functional Elements in 3'UTRs*) [178] metoda wysokoprzepustowa mapowania elementu regulatorowego w 3'UTR. Wykorzystuje *lucirerase*

reporter assays "przeskalowany" na metodę wielkoprzepustową. Jest ona rodzajem funkcjonalnej analizy, ponieważ detekcja i walidacja elementów zachodzi w tym samym czasie.

W obrębie miejsc wiązania wprowadza się także mutacje punktowe. Jeśli efektem tych mutacji jest zniesienie regulacji miRNA, miejsce wiązania jest potwierdzeniem weryfikacji targetu. Ta sama metoda znajduje zastosowanie w badaniu znaczenia generalnych cech wpływających na interakcje miRNA/target. Na przykład analizy znaczenia regionu 5' *seed* dla rozpoznawania targetu [43], czy jego dostępności [88].

3 Modele w analizie interakcji miRNA/mRNA

3.1 Wprowadzenie

Mechanizm regulacji genów w procesie RNAi przedstawiony w rozdziale "Biologiczne podstawy regulacji genów" jest analizowany i modelowany technikami eksperymentalnymi oraz metodami bioinformatycznymi. Proces regulacji genów charakteryzujący się niską specyficznością (swoistością) w rozpoznawaniu regulowanych transkryptów staje się inspiracją w modelowaniu procesu RNAi u zwierząt. Podstawowe dwie grupy istniejących rozwiązań bioinformatycznych dotyczą metod rozpoznawania genów miRNAs na sekwencjach chromosomowych oraz metod identyfikacji par interakcji miRNA/mRNA. Rozpoznanie genów miRNAs, ich kontekst i położenie w stosunku do genów kodujących oraz wzajemne ich położenia pozwala przewidywać ich aktywność i interakcje zachodzące między genami. Metody rozpoznawani genów miRNAs opierają się na badaniu homologii lub stosowaniu metod nauczania maszynowego. Metody homologiczne (wysoki stopień podobieństwa sekwencji miRNA z sekwencją zlokalizowaną w genomie) wykorzystują istniejący zbiór miRNAs odpowiedni dla danego organizmu. Nie pozwalają one jednak identyfikować genów niepoznanych jeszcze miRNAs. Także wtedy, kiedy chcemy wykorzystać zbiór miRNAs pokrewnego gatunku biologicznego, co jest spowodowane szybką ewolucją miRNAs [54], metody te okazują się nieprzydatne. Metoda nauczania maszynowego wykorzystuje przykładowe znane geny, które pochodzą z genomu badanego organizmu. Na ich podstawie przeprowadza się predykcję struktury II rzędowej, która jest podstawowym kryterium poprawnego przetworzenia genu miRNA na funkcyjne miRNA [11].

Funkcjonalność miRNA zależy od precyzyjnie rozpoznanych targetów. Narzędzia *in computo* służące do rozpoznania targetów opierają się na komplementarności, filogenetycznej konserwatywności zarówno miRNA jak i mRNA, oszacowaniu energii swobodnej dupleksu miRNA/mRNA oraz ocenie dostępności regionu transkryptu. Stosunkowo niedawno powstały narzędzia integrujące także ekspresję mRNA i miRNA w celu charakterystyki funkcyjnej *in situ* interferencji RNA [79][114].

Hybrydyzacja krótkiego 7-8 nukleotydowego fragmentu cząsteczki miRNA (*seed*) z targetem najczęściej zachodzi w rejonie 3'UTR. Ta krótka komplementarności stanowi znaczącą trudność prawidłowego rozpoznania targetu. Ponieważ specyficzność 7-8nt jest bardzo niska, można powiedzieć, że prawie każda para miRNA/mRNA przy założeniu takiej komplementarności może interferować. W obrębie *seed* stwierdzono i potem uwzględniono w predykcjach także pewne odstępstwa od pełnej komplementarności. Tzw. gapy w tym rejonie świadczą o konformacyjnym dopasowaniu, a nie tylko "sekwencyjnym" powstającego dupleksu.

Ten problem niejednoznaczności w określeniu par miRNA/mRNA może świadczyć o złożoności mechanizmu kryjącego się pod hasłem "interferencji RNA", ponieważ przypuszczać tu można występowanie innych czynników regulacyjnych, które *in situ* dokonują ostatecznej decyzji o hybrydyzacji i degradacji transkryptu. Może to być mechanizm kontroli jakości transkryptów, na co wskazuje ewolucyjne pochodzenie tego mechanizmu. Nieokreśloność tego mechanizmu sugeruje ostrożne wnioskowanie z przeprowadzanych eksperymentów weryfikujących wytypowane targety, które w swojej istocie dotyczą badania komórkowych mechanizmów regulacyjnych w stanie patologii indukowanej czy sztucznej. Należy przypuszczać, że obecne bazy informacji potwierdzonych eksperymentalnie interferujących ze sobą miRNA i mRNA dotyczą

samoregulacji będącej reakcją na skrajne, patologiczne zakłócenia funkcjonowania komórki, a nie wnoszą wiedzy o homeostatycznej kontroli rozwoju komórki w jednym z wielu mechanizmów regulacji, jakim podlega produkcja białek.

Zróżnicowanie sekwencyjne cząsteczek miRNAs pokrywa całą przestrzeń możliwych kombinacji sekwencyjnych. Dobór przez komórkę odpowiednich, aktywnych w danym momencie endogennych miRNA w cytoplazmie uwzględnia kontrolę całego transkryptomu, pozostawiając w działaniu tylko te transkrypty, które nie wykazują ani zbyt małej ani zbyt dużej komplementarności sekwencji par miRNA/mRNA. W takiej sytuacji wprowadzenie w eksperymentalnych technikach egzogennego miRNA do cytoplazmy stanowi zaburzenie całego misternego schematu interakcji wzajemnej puli miRNAs.

Trudności we wnioskowaniu rzeczywistych interakcji miRNA z mRNA znalazły przynajmniej częściowe rozwiązanie poprzez wyodrębnienie skorelowanych "sygnałów". Rozpoznano i scharakteryzowano szereg dodatkowych czynników, których uwzględnienie poprawia jakość predykcji interakcji, przynajmniej tych eksperymentalnie przeanalizowanych (patrz rozdział 2.4).

Uwzględnienie przedstawionych czynników we wnioskowaniu na temat interakcji miRNA/mRNA oraz zaproponowana przez autora tej pracy integracja informacji o ekspresji cząsteczek miRNA z pozostałymi parametrami może być istotnym elementem poprawy identyfikacji targetów, szczególnie przy ograniczonej ilości przeprowadzanych prób w ramach eksperymentów mikromacierzowych. Znaczenie integracji informacji powinno być odpowiednim kierunkiem rozwoju poznawania rzeczywistości w celu uzyskania obiektywnej informacji na temat mechanizmów interferencji miRNA.

3.2 Dane i zasoby informacji

Modele i metody analizy interakcji miRNA/mRNA wykorzystują różne dane wejściowe dla przeprowadzenia predykcji. Stosowane są techniki wysokoprzepustowe: ilościowa reakcja łańcuchowa polimerazy DNA (qRT-PCR) [27], technologia mikromacierzowa (*DNA microarrays*) [111], równoległe sekwencjonowanie DNA (*NGS*) [127], oraz technikę RNA-seq. Ta ostatnia technika wykorzystując NGS pozwala na wyznaczenie transkryptomu oraz jego sekwencje w wybranym momencie życia komórki [125] dostarczając informacji o ekspresji transkryptów.

Informację o znanych i scharakteryzowanych miRNAs: struktura cząsteczki, prekursory, ich geny, rodziny – pogrupowaniach wg różnych kluczy, przechowują różnego rodzaju bazy danych: miRBase, TargetScan, microRNA. Standardy odnośnie ustalania struktury nukleotydów zajmuje się *International Structural Genomics Organization* [81][167], jednak informacje z eksperymentów dotyczących badania struktury miRNAs są w małym stopniu rozwinięte [96]. Znajomość struktury niekodujących RNA jest istotna dla zrozumienia podstawowych mechanizmów regulacji potranskrypcyjnej w sytuacji niskiej komplementarności dupleksu miRNA/mRNA. Przykładowe bazy struktur RNA to: *Nucleic acid database* [14], *RNA base-pair structure* [182], *NCBI structure* [174], SCOR *database* [90], *RNA strand* [1].

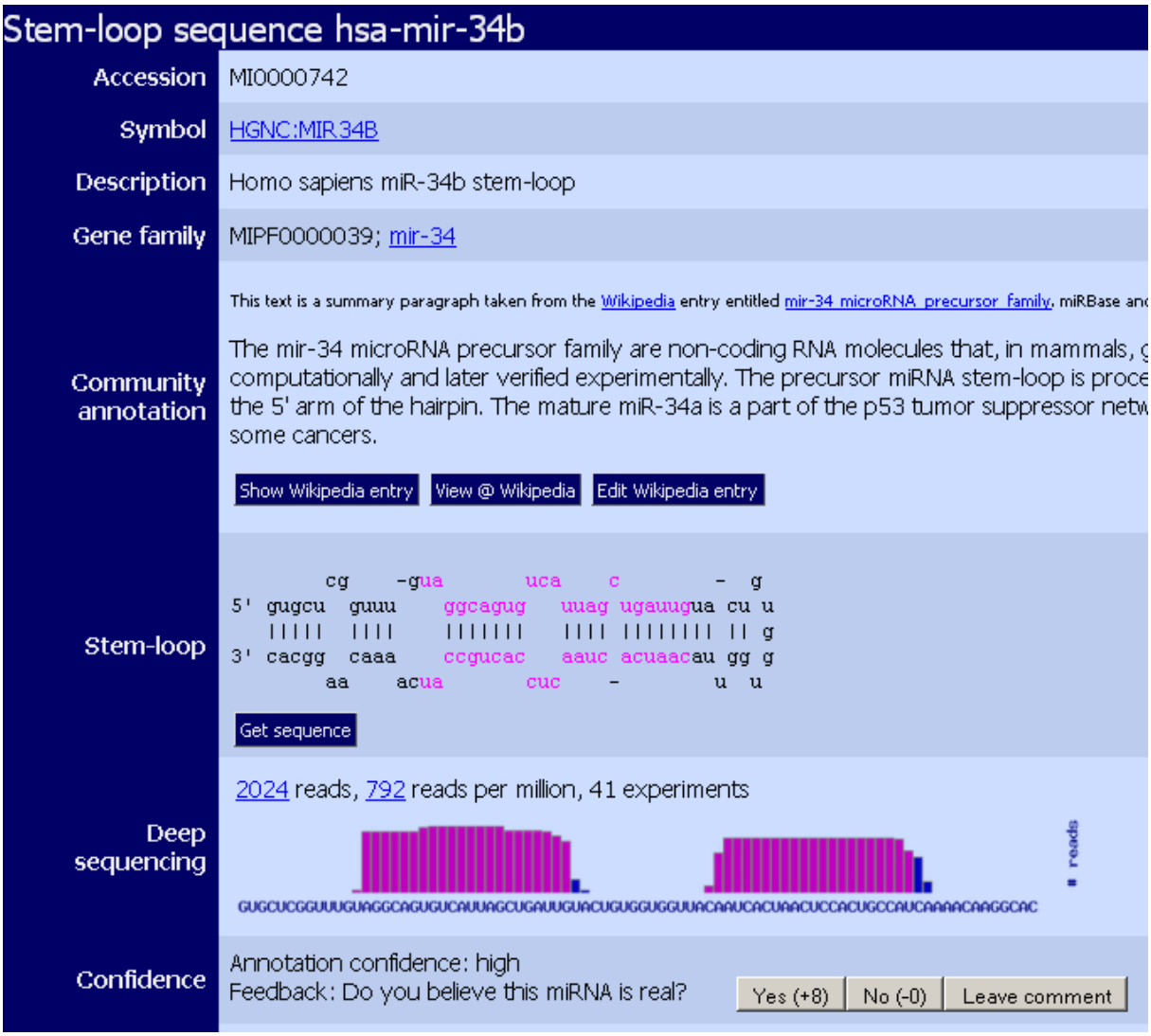
Informację o transkryptach (analizowanych później, jako docelowych targetów miRNAs) można uzyskać z NCBI. Wykorzystany w tej pracy zbiór transkryptów pochodzi z *Reference Sequence* (RefSeq) (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/). Stanowi on zbiór pełnych, sformatowanych, nieredundantnych sekwencji naturalnie występujących cząsteczek transkryptów [136]. Wykorzystywany często, jako zbiór referencyjny. Powstał on z sekwencji zgromadzonych w

ramach *International Nucleotide Sequence Database Collaboration* (INSDC). Informację zawarte w rekordzie każdego transkryptu dotyczą sekwencji w tym także różnych wariantów, adnotacji, a także fizycznej lokalizacji genomowej. Zasób zawiera ortologi, paralogi i alternatywne haplotypy (dla niektórych organizmów) a także transkrypty z alternatywnych splicingów kodujących te same białka lub jego odmienne formy izomeryczne. Są to: mRNA, mRNA hipotetycznych białek, niekodujące RNA, pseudogeny, miscRNA (małe RNA o różnorodnej funkcji), *small nucleolar* RNA, miRNA. Zbiór RefSeq dostępny jest w formacie FASTA lub pełne rekordy w formacie GenBank.

3.2.1 Baza miRBase

MicroRNAs registry [68] wprowadza nomenklaturę (nazewnictwo) miRNAs. Baza pozwala uzyskać prekursorowe i dojrzałe miRNAs. Jest podstawową, specjalistyczną bazą rejestrującą informację o cząsteczkach miRNAs. Zawiera sekwencje miRNAs występującą u różnych gatunków, w konkretnych tkankach lub liniach komórkowych. Oprócz sekwencji miRNA dostępne są sekwencje prekursorów miRNA (pre-miRNA). Baza podaje też "stopnień zaufania" danej cząsteczki, który potwierdza realność jej występowania. "Stopień zaufania" (*confindence*) oparty jest na krotności odczytywania danego nukleotydu na danej pozycji podczas sekwencjonowania (*deep sequencing*).

Baza podaje też informację o pochodzeniu danej cząsteczki: introniczne lub/i egzoniczne miRNA. Możliwość wyszukiwania miRNAs wg rodzin. Nazwa, identyfikator miRNA koduje informację o rodzinie – rodzina tutaj zdefiniowana oznacza zbiór cząsteczek o podobnej sekwencji, ale zasób nie podaje precyzyjnej definicji. Rodziny oznaczane są poprzez wspólny cyfrowy prefiks np.: ggo-mir34c, hme-mir-34-1, hsa-mir-34a. Natomiast pojęcie klastra odnosi się do genomowej lokalizacji. Geny miRNA pochodzące z tej samej lokalizacji na chromosomie najprawdopodobniej ulegają równoczesnej transkrypcji i przynależą do tego samego klastru (Rys. 3.1). Pierwotne transkrypty tzw. policistronowe, czyli obejmujące kilka genów, z reguły są długie na dziesiątki tysięcy nukleotydów (u ssaków) [8].



Rys. 3.1. Fragment wyniku wyszukiwania prekursora hsa-mir-34b (http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000742). Pokazana jest rodzina do której należy gen, predykcja struktury drugorzędowej oraz ocena wiarygodności cząsteczki (deep sequencing)

3.2.2 Interpretacja wyników mikromacierzy ekspresji miRNA

Uzyskane surowe dane z mikromacierzy dotyczą intensywności fluorescencji, które ze względu na ograniczenia techniczne metody muszą zostać odpowiednio przetworzone w ramach tzw. analizy niskiego poziomu [65]. Obejmuje ona korekcję tła, logarytmizację ilorazu intensywności zmierzonych poziomów fluorescencji i normalizację. Analiza wysokiego poziomu obejmuje wyznaczenie genów o istotnej ekspresji różnicowej przy pomocy np. testu t, modelu regresji, modelu mieszanego, empirycznej metody Bayesa albo SAM.

Błędy zawarte w danych surowych wynikają z procedury przygotowania próby mRNA, procesu amplifikacji materiału próby. Intensywność fluorescencji zależna jest także od stopnia hybrydyzacji dupleksów. Omówimy teraz kolejne kroki wykonywane w ramach analizy niskiego poziomu.

Korekcja tła – pierwszy krok w przetwarzaniu danych surowych, dzięki któremu można porównywać dane z różnych mikromacierzy. Ten etap koryguje iloraz intensywności i chroni go przed niedoszacowaniem. Najprostsza metoda korekcji polega na odjęciu wartości lokalnego tła (średnia lub mediana) otaczającego pole od sygnału samego pola. Konsekwencją tej procedury jest redukcja wysokich wartości na mniejsze. Negatywnym aspektem jest uzyskanie w wyniku tej

korekcji dla pewnych pól wartości ujemnych (gdy wartość oryginalna jest mniejsza od wartości średniej lub mediany), a więc takich, które dalej nie mogą zostać przetwarzane. Dlatego wprowadza się szereg innych metod wyznaczających lokalne tło w celu obejścia powyższego problemu. Wówczas wartość odejmowana pochodzi z matematycznych modeli, nieliniowych filtrów, jak również metod wyznaczania intensywności sygnału, które nie wykorzystują odejmowania, aby uzyskać korekcję tła (Kooperberg [91], Edwards [46], Normexp [142], VSN).

Metoda Kooperberga oparta jest na bayesowskim modelu wykorzystującym splot rozkładów normalnych do korekcji sygnałem tła każdego pola. Wartości średniej i odchylania standartowego wyznaczane są dla 3-4 sąsiednich pól wchodzących w skład kanału. Aby uzyskać wartość oczekiwaną rzeczywistego sygnału każdego pola wyliczana jest całka numeryczna. Metoda Edwards'a z kolei ustala próg i stosuje interpolację funkcją monotoniczną. Odejmowanie tła jest przeprowadzane tylko dla sygnałów pochodzących z pól, gdzie różnica między sygnałem i tłem przekracza wartość progową. Dla pozostałych sygnałów zostaje wykorzystana funkcja interpolująca. Normexp opiera się na algorytmie RMA [82]. Model ten zakłada, że sygnał tła podlega rozkładowi normalnemu, podczas gdy sygnał próby rozkładowi wykładniczemu. W odróżnieniu od RMA, algorytm Normexp dokonuje separacji pól i używa funkcji wiarygodności.

Transformacja logarytmiczna – poziom ekspresji przedstawiany jest jako iloraz intensywności fluorescencji dwóch znaczników obliczana dla każdego pola. Większe poziomy ekspresji uzyskują wartości od 1 do nieskończoności, mniejsze między 0 a 1. W celu zbalansowania tej różnicy wprowadza się przekształcenie logarytmiczne. Zastosowanie logarytmu o podstawie 2 powoduje, że dwukrotny przyrost poziomu ekspresji odpowiada poziomowi ekspresji $E_{level} = 1$, a dwukrotny spadek - $E_{level} = -1$. Główną zaletą tej transformacji jest uzyskanie porównywalnych poziomów ekspresji. Wartość E_{level} wyznaczana jest ze wzoru:

$$E_{level} = log_2 \frac{R_{fg} - R_{bg}}{G_{fg} - G_{bg}} \quad \textbf{(3-1)}$$

gdzie:

E_{level}– poziom ekspresji genu;

$R_{fg}, R_{bg}, G_{fg}, G_{bg}$ - intensywność promieniowania czerwonego i zielonego tła (*background*) i sygnału (*foreground*).

Normalizacja – jest procedurą obliczeniową korygującą błędy spowodowane różnymi parametrami barwników zastsowanych w tym samym eksperymencie, różnym stopniem efektywności znakowania, hybrydyzacji i skanowania, które występują podczas eksperymentu mikromacierzowego. Celem normalizacji jest uzyskanie danych porównywalnych z wynikami z innych mikromacierzy w taki sposób, aby jednocześnie nie utracić istotnych wartości biologicznych. Proces obejmuje normalizację sygnału intensywności ze wszystkich pól względem wspólnego czynnika. Czynnik ten może opierać się na statystycznych parametrach takich jak intensywność wszystkich sygnałów lub średni sygnał całego zbioru (globalna normalizacja) lub na kontroli sygnału tak zwanych *housekeeping genes*, dla których przyjmuje się stałą wartość między różnymi próbami (wewnętrzna normalizacja), lub dodatkowo – na kontroli sygnału z sond *spike-in* (normalizacja zewnętrzna).

Chociaż metody normalizacji zostały opracowane dla mikromacierzy z dużą liczbą pól i opierają się na podstawach statystycznych, ich zastosowanie dla mikromacierzy miRNA powinno zostać

zweryfikowane. Jest to spowodowane relatywnie małą liczbą pól na płytce mikromacierzy miRNA oraz faktem, że inaczej niż dla ekspresji mRNA, poziom miRNAs może zmieniać się znacząco pomiędzy próbami. Dodatkowo na razie nie stwierdzono w puli miRNAs odpowiedników *housekeeping genes*.

Dlatego uwagę przywiązuje się tutaj do metody *spike-ins*, czyli sztucznych RNA, dodanych do próby przed jej znakowaniem. Sygnał ze znakowanych *spike-in* jest uzyskiwany po ich hybrydyzacji do odpowiednich pól obecnych na macierzy. Dopiero detekcja prawidłowych sygnałów pochodzących z tych pól umożliwia wykorzystanie ich do normalizacji.

Normalizacja, która umożliwia porównywanie danych z różnych mikromacierzy stosowana jest po to, aby zlikwidować różnice techniczne, różnice znakowania, hybrydyzacji i skanowania. Przeprowadza się ją na podstawie parametrów, które przyjmuje się, jako stałe i niezależne od macierzy. W metodach normalizacji większość intensywności sygnału pozostaje nieruszona, więc statystyka ma na celu oddzielenie istotnego sygnału od większości niezmienionego sygnału. To podejście nie zdaje egzaminu, kiedy różnice miedzy próbami są duże i ilość niezmienionego sygnału do normalizacji jest zredukowana. Uważa się, że jeśli w eksperymentach na transkryptomach jest to rzadkie zjawisko, to jednak częste dla macierzy analizujących miRNAs.

Przykładowe metody normalizacji:

1. Skalowanie [113] – przeprowadzane na liniowej skali, przed logarytmowaniem, koryguje intensywności sygnału na podstawie wartości średniej lub mediany. Uzyskany faktor dla każdej macierzy jest używany wielokrotnie na każdym polu macierzy. Ponieważ macierze miRNA mają małą gęstość, skalowanie może nie być właściwą metodą.
2. Metoda kwantylowa [18] – opiera się na założeniu, że dwa zbiory z dwóch prób powiązanych ze sobą danych powinny uformować liniową (diagonalną) zależność, kiedy dane zostaną ustawione przeciwko sobie. Powoduje ona równy rozkład intensywności.
3. *Locally Weighted Scatterplot Smoothing* (Lowess) [12] – lokalna regresja do wygładzenia wykresu: ilorazu M/A (log ratio/log mean) w funkcji linearnego rozkładu. Zakłada się, że większość sygnałów między próbami się nie zmieniam Metoda forsuje równą średnią. Pozwala na korekcję systematycznego odchylenia na wykresie MA, korygując MA do lnie prostej.
4. *Variance stabilization and normalization* (VSN) [80] – łączy korekcje tła, addytywną i muliplikatywną korekcję. Stabilizacja wariancji jest używana, aby zredukować zależność wariancji i intensywności sygnału.

Analiza różnicowej ekspresji wchodzi w skład analizy wysokiego poziomu. Jednym z podstawowych zadań analizy statystycznej danych mikromacierzowych jest wytypowanie genów charakteryzujących się istotnie różniącą się ekspresją, jako konsekwencję porównywania ekspresji w różnych grupach pacjentów, tkanek, komórkach czy różnych biologicznych warunkach. Wykrycie genów o różnicowej ekspresji pomaga w zrozumieniu funkcji genów, mechanizmach regulacji i procesów komórkowych. Analiza ta poprzedza wieloczynnikowe analizy klasteryzacji, klasyfikacji czy analizę wzbogacania genów (*gene set enrichment*). Dane mikromacierzowe są wartościami ciągłymi, które przed analizą różnicową musi poprzedzić wstępna analiza. Uzyskane wartości istotności są korygowane procedurami jednoczesnego testowania wielu hipotez.

Metody parametryczne stosowane w wykrywaniu różnicowej ekspresji genów:

1. Analiza stopnia zmiany (*fold change*) – polega na wyliczenie parametru FC (*fold change*) wg. wzoru:

$$FC_i = log_2 \frac{\bar{x}_i}{\bar{y}_i} = log_2 \bar{x}_i - log_2 \bar{y}_i \quad \textbf{(3-2)}$$

   gdzie: $\bar{x}_i, \bar{y}_i$ wartości średnie poziomu ekspresji *i* – tego genu odpowiednio w grupie kontrolnej i badanej. Wyznaczenie tych wartości dla wszystkich genów pozwala na ustalenie wartości progowej rozdzielającej geny o niezmienionej i zmienionej ekspresji.
2. Test t – najczęściej stosowany do wyznaczenia genów o różnicowej ekspresji. Definiujemy statystykę dla dwóch prób:

$$t_i = \frac{\bar{x}_i - \bar{y}_i}{s_i} \quad \textbf{(3-3)}$$

   gdzie: s_i - błąd standardowy i-tego genu. Po wyznaczeniu statystyki t wartość p zostaje obliczona na podstawie rozkładu t. Dla przyjętej wartości progowej – alfa (np. α=0,05) można dokonać wydzielenie genów o istotnie zmienionej ekspresji.
3. *Significance Analysis of Microarrays* (SAM). Metoda ta została wprowadzona w celu poprawy oszacowania błędu wariancji w przypadku małej ilości powtórzonych prób danego eksperymentu. W tej metodzie zmodyfikowano test t poprzez dodanie stałej małej wartości w mianowniku statystyki. Dzięki tej modyfikacji geny z małą krotnością zmian ekspresji nie zostaną zaszeregowane do tych o zmiennej ekspresji.

$$t_i = \frac{\bar{x}_i - \bar{y}_i}{s_0 + s_i} \quad \textbf{(3-4)}$$

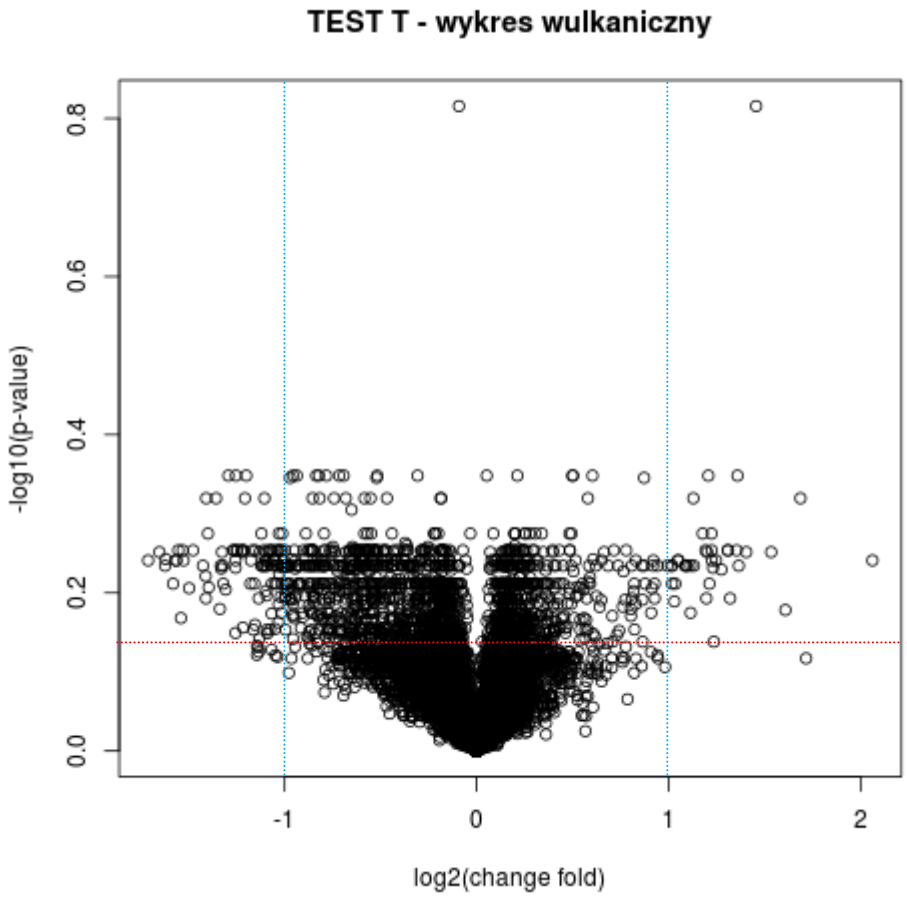
   Oprócz niewątpliwych korzyści zastosowania metody SAM (np. najlepiej radzi sobie z wytypowaniem najmniejszej liczby genów o różnicowej ekspresji) należy pamiętać, że wprowadza ona pewne założenia dotyczące rozkładu wartości mierzonych, które są trudne do weryfikacji szczególnie w przypadku dysponowania niewieloma macierzami (próbami). SAM zakłada symetryczny rozkład błędów przypadkowych wspólny dla wszystkich genów. Konsekwentnie wariancja dla małej ilości prób może być zaszumiona, geny statystycznie istotne, charakteryzujące się małą zmianą poziomu ekspresji mogą nie być istotnie biologicznie, itd.
4. Test Cyber-t. Modyfikacja testu t, która uwzględnia globalną uśrednioną wariancję.
   Limma – kolejna modyfikacja testu t podobna do SAM, ale wykorzystująca podejście Bayesa do wyznaczenia testu t. Każda wariancja genu jest obliczana, jako średnia ważona wariancji dla konkretnego genu i globalnej wariancji [155].

Niezależnie od rachunkowej analizy przeprowadza się także graficzną prezentację ekspresji genów. Do oceny różnicy ekspresji genów na podstawie sygnału fluoroscencyjnego zielonego i czerwonego tworzy się wykres MA [16], jako logarytm ilorazu promieniowania czerwonego i zielonego (M) w funkcji średniej geometrycznej (A).

$$M = log_2(\frac{R}{G}) \qquad A = log_2\sqrt{RG} \tag{3-5}$$

Wykres MA pozwala także na ocenę błędów stałych i systematycznych.

Zmienność ekspresji między testowym i referencyjnym zbiorem transkryptów ocenia się na podstawie tzw. wykresu wulkanicznego (*volcano plot*). Wykres zestawia istotność zmienności ekspresji dla każdego genu (p-value) w funkcji krotności zmian poziomu ekspresji (log2(change fold)). Przykładowy wykres wulkaniczny przedstawiono na Rys. 3.2 [173]. Na wykresie wprowadzono linie pomocnicze: wszystkie punkty (geny) powyżej czerwonej linii poziomej posiadają wartości istotności mniejsze od 0,05, punkty na lewo i na prawo od odpowiednich niebieskiej lewej i prawej linii pionowej charakteryzuje większa niż dwukrotna zmiana poziomu ekspresji.



Rys. 3.2. Wykres typu wulkanicznego: NHDF PK15 (*porcine endogenous retroviruses*) porównane z NHDF (*human dermal fibroblasts*)

Głównym problemem analizy profilu miRNA jest to, że ekspresja genu może podlegać czynnikom zakłócającym wynikłym z przestrzennej niejednorodności lub nasycenia sygnału [39]. Czynniki te mają wpływ na cały badany profil w eksperymencie. Prowadzi to do sytuacji, kiedy uzyskuje się nadmiarową liczbę genów klasyfikowanych jako o różnicowej ekspresji. Są to wyniki fałszywie dodatnie. Stosuje się wówczas następujące rozwiązania [135]:

1. Standaryzowany test t.
2. Test *statistitics null distribution*.
3. Metodę bootstrapu.

3.2.3 Baza - TargetScan punktacja kontekstowa i konserwatywność

Zasób TargetScan udostępnia dane dotyczące efektywności represji transkryptów CS (*Context Score*) oraz konserwatywności targetu P_{CT} (*Probability of Conserved Targeting*). W niniejszej pracy wykorzystano zbiory znajdujące się pod adresem http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61. Zbiory te w formacie tekstowym CSV (*Comma-Separated Values*) zawierają tabelaryczne dane i pozwalają uzyskać wartości CS i P_{CT} dla konkretnego genu, transkryptu i miRNA. Zawartość zbiorów jest następująca (Tabela 3):

Tabela 3. Wybrane zasoby bazy TargetScan

"miR_Family_Info.txt"	zawiera sekwencję miRNA i przynależną rodzinę(miR_family, seed_m8, species_ID, miRBase_ID, mature_sequence, family_Conservation, miRBase_Accession);
"Conserved_Site_Context_Scores.txt" "Nonconserved_Site_Context_Scores_Human.txt"	wartości CS oraz powiązane parametry (gene_ID, gene_Symbol, transcript_ID, gene_Tax_ID, miRNA, site_Type, uTR_start, uTR_end, prime3_pairing, local_AU, position, tA, sPS, context_score, context_score_percentile)
"Summary_Counts.txt"	podsumowanie dla każdej pary gen/rodzina miRNA (transcript_ID, gene_Symbol, miRNA_family, species_ID, total_num_conserved_sites, number_of_conserved_8mer_sites, number_of_conserved_7mer_m8_sites, number_of_conserved_7mer_1a_sites, total_num_nonconserved_sites, number_of_nonconserved_8mer_sites, number_of_nonconserved_7mer_m8_sites, number_of_nonconserved_7mer_1a_sites, representative_miRNA, total_context_score, aggregate_PCT)

Zasób TargetScan wprowadza własne pojęcia *seed* oraz rodzin miRNAs, które wiążą się bezpośrednio z pierwszymi 7 nukleotydami od strony 5':

- "seed" – fragment sekwencji miRNA zawierający od 2-7 nukleotydu od strony 5';
- "seed+m8" – *seed* z kolejnym (ósmym) nukleotydem.

Rodzina cząsteczek miRNAs- definiowana jest, jako podzbiór miRNAs posiadających taki sam region "seed+8m".

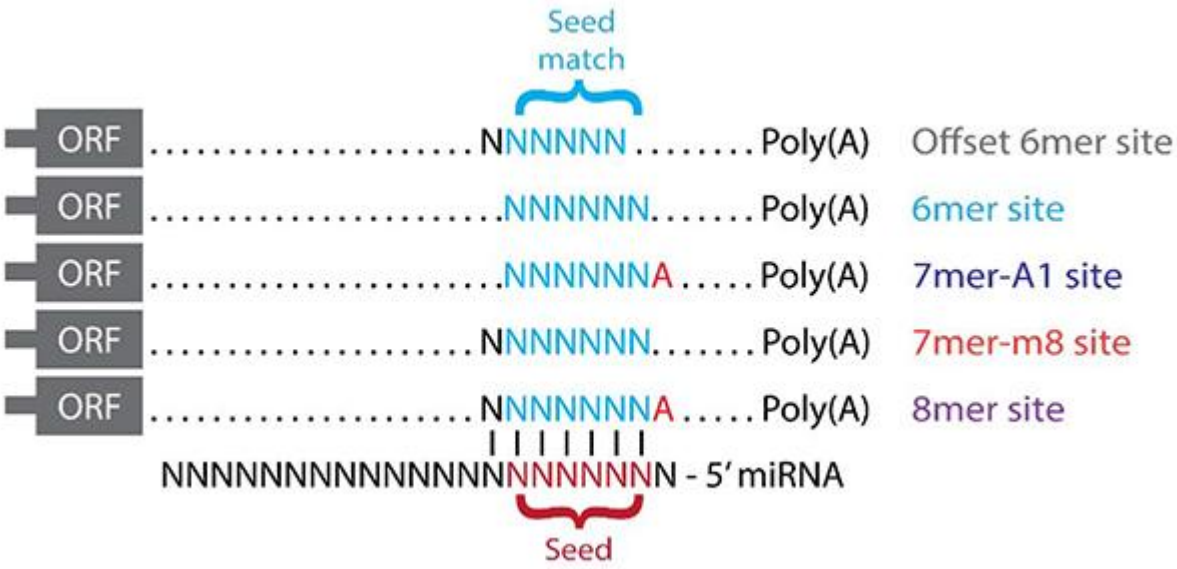
Wartość CS powstała, jako integracja wiedzy kontekstowej, czyli specyficzności charakterystyki sekwencji targetu powyżej i poniżej regionu wiązania targetu. Stanowi ona sumę sześciu punktacji (Tabela 4):

Tabela 4. Parametry punktacji kontekstu sekwencji transkryptów. Rodzaje miejsc wiązania uszeregowane wg ich efektywności

	parametr	opis	wersja	
1	rodzaj *site*	7mer-1a (1), 7mer-m8(2), 8mer (3), 6mer (4)	context score 2007	context+ score 2011
2	3'pairing contribution	wartość oceniająca stopień komplementarności pary miRNA/target w regionie poza *seed*.		
3	local AU contribution	wartość odpowiadająca koncentracji adenin i uracyli powyżej i poniżej przewidywanego miejsca wiązania.		
4	position contribution	odległość miejsca wiązania od najbliższego końca UTR targetu.		
5	TA (TargetSite AbundanceContribution)	wartość odpowiadająca liczebności miejsc wiązania dla całej rodziny miRNA w obrębie wydzielonych regionów 3'UTR transkryptów.		
6	SPS (Seed-Pairing Stability Contribution)	wartość oceniająca stabilność dupleksu miRNA/target jako funkcji koncentracji par A-U w regionie *seed*.		

Wprowadzone przez TargetScan rodzaje miejsc wiązania, które wpływają na stabilność dupleksu (Rys. 3.3.):

- "7mer-1a" -*site* które wiąże się z dojrzałym miRNA na jego pozycjach (5') 2-7 występujących po adyninie.
- "7mer-m8" - *site* który wiąże się z dojrzałym miRNA na jego pozycjach (5') 2-8.
- "8mer" - *site* który wiąże się z dojrzałym miRNA na jego pozycjach (5') 2-8, występujących po adyninie.
- "6mer" - *site* który wiąże się z dojrzałym miRNA na jego pozycjach (5') 2-7.



Rys. 3.3. Kanoniczne rodzaje miejsc wiązania.

Wprowadzonym rodzajom miejsc wiązania przypisuje się wartości konserwatywności, która zależna jest od długości gałęzi drzewa filogenetycznego. Dla każdego rodzaju miejsca wiązania wprowadza się różne progi konserwatywności:

- 8mer >= 0.8,
- 7mer-m8 >= 1.3,
- 7mer-1A >= 1.6,

- 6mer - brak konserwatywności.

Dla genów z wielokrotnymi miejscami wiązania dla jednej rodziny miRNA, całkowita wartość CS jest wartością wyznaczoną dla "najlepszego" miRNA w rodzinie (tzw. reprezentatywny miRNA).

Wartość P_{CT} (*probability of preferentially conserved targeting*) jest wynikiem analizy komparatywnej ortologicznych genów (lub mRNA). Termin **ortologiczne geny** oznacza rodzinę zduplikowanych genów, które pojawiły się u różnych gatunków w wyniku specjacji, czyli rozdzielenia się gatunków. Jeśli spokrewnione geny występują w obrębie tego samego organizmu wówczas nazywamy je paralogami. Stwierdzenie ortologiczności genów albo ich fragmentów obywa się na podstawie oszacowania podobieństw genów u różnych obecnie żyjących organizmów "poukładanych" na drzewie filogenetycznym. Odnalezione wspólne, podobne sekwencje w różnych genomach poddaje się badaniu konserwatywności. Konserwatywność jest miarą nie tylko podobieństwa lub identyczności (homologii) sekwencji nukleotydowych występujących u różnych gatunków (ortologiczne sekwencje), ale przede wszystkim miarą zachowania ich funkcyjności w ewolucji. Szczególnie widoczne to doprecyzowanie jest w przypadku badania konserwatywności krótkich sekwencji, których konserwatywność – rozumiana, jako homologiczność sekwencji, może być przypadkowa a nie wynikiem doboru naturalnego. Homologiczność długich sekwencji u różnych gatunków oznacza ich konserwatywność z racji małego prawdopodobieństwa przypadkowości zachowania ich homologii.

Konserwatywność sekwencji podkreśla ich znaczenie funkcyjne. Mutacje w regionach chromosomowych o wysokiej konserwatywności prowadzą z reguły do powstania formy niezdolnej do życia lub takiej, która zostaje wyeliminowana przez naturalną selekcję. W mechanizmie RNAi hybrydyzacja miRNA i mRNA dotyczy przede wszystkim rejonu *seed*. W tym przypadku analizę komparatywną przeprowadza się dla regionów genów 3'UTR ze szczególnym akcentem stawianym na regiony miejsc wiązania, które odnajdujemy w UTRs różnych gatunków.

Oznaczenie konserwatywności krótkich sekwencji 6-8merów (*sites*), które stanowią wynik dopasowania *seeds* miRNAs [19] jest wykorzystywane w predykcji targetów. Uważa się, że uwzględnienie konserwatywności regionów miejsc wiązania powinno poprawić jakość predykcji par miRNA/mRNA i zmniejszyć ilość wyników fałszywie pozytywnych, ze względu na "podpowiedź" jaką udziela natura.

Ocena konserwatywności sekwencji jest bardzo złożoną, wieloetapową procedurą obliczeniową. Wykorzystuje się w niej szereg metod:

- wielosekwencyjne uliniowienie (*multialignment*) służące do uzyskania ortologów,
- do tworzenia drzew filogenetycznych,
- do wyznaczania długości gałęzi drzew filogenetycznych,
- do wyznaczania lokalnej konserwatywności,
- obliczania statystycznej istotności uzyskanych rezultatów.

Na użytek zasobów TargetScan wprowadzono parametr P_{CT}. Procedura obliczeniowa tego parametru analizuje konserwatywność 6-8 merów traktowanych jako potencjalne miejsca wiązania. W wyniku tej analizy uzyskano ponad 45 000 konserwatywnych site'ów wewnątrz ludzkich regionów 3'UTRs oraz przeszło 60% ludzkich genów kodujących, których regiony 3'UTR podlegają selekcji naturalnej łączenia się w pary z miRNAs [58].

Badanie konserwatywności 6-8merowych miejsc wiązania przeprowadzono dla określonego podzbioru miRNAs [58]. Miejsca wiązania powstają wówczas, jako wynik dopasowania *seeds* miRNAs do wstępnie wyselekcjonowanych, ortologicznych regionów 3'UTR. Zbiór miRNA podzielony został na trzy klasy:

1. szeroko konserwatywne (2) – konserwatywność u większości kręgowców, zwykle po danio pręgowany (*zebrafish*);
2. konserwatywne (1) – konserwatywność u większości ssaków, ale zwykle nie dalej niż po ssaki łożyskowe;
3. wąsko konserwatywne (0) – wszystkie pozostałe.

W Tabela 5 przedstawiono zrealizowane podsumowanie zawartości zbioru miRNAs

Tabela 5. Zestawienie liczebności tych klas w zbiorze miRNAs (biosql_test.mirna)

	NULL	0	1	2	suma
miRNA	861	1255	91	186	2393

Ponieważ mutacje, konwersja genów, *crossover* są różne dla różnych regionów genomu, konsekwentnie różne UTRs mają różny podstawowy poziom konserwatywności. Oprócz podstawowego poziomu konserwatywności, który należy uwzględnić w obliczeniach, sekwencje UTRs poza zaangażowaniem w mechanizm RNAi wykazują także inną funkcyjność, która we własnym zakresie "dba" o konserwatywność. Dlatego miejsca wiązania, które znajdują się wewnątrz UTRs z wysokim poziomem konserwacji są mniej prawdopodobne, by były konserwatywnymi w mechanizmie RNAi, niż te ulokowane w szybko ewolucyjnie zmieniających się UTRs.

Do wyznaczenia konserwatywności ortologiczne geny uzyskano przez wielosekwencyjne uliniowienia (*multialignment*) znanych lokalizacji 3'UTR dla 28 genomów kręgowców. Na podstawie uśrednionych odległości między dopasowaniami utworzono drzewo filogenetyczne, dla którego w następnym kroku wyznaczono długości gałęzi w sposób uwzględniający indywidualny poziom podstawowy konserwatywności. Aby w dalszych obliczeniach uwzględniać indywidualny stopień konserwatywności danej grupy UTRs podzielono cały zbiór UTRs na 10 podzbiorów od najmniej po najbardziej konserwatywne. Dla każdego podzbioru wyznaczono drzewo filogenetyczne, które mają identyczny układ gałęzi, mimo że różnią się ich długościami. Konserwatywność danej sekwencji – jednego z miejsc wiązania danego miRNA jest wtedy oceniana, jako suma długości wszystkich gałęzi na drzewie łączących podzbiór gatunków posiadających perfekcyjnie dopasowaną tą sekwencję. Długości do obliczeń wybiera się z odpowiedniego drzewa, które reprezentuje dany podzbiór zawierający dany UTR o znanym współczynniku konserwatywności tła. Miejsca wiązania, które znajdują się w podzbiorach bardziej zróżnicowanych UTRs charakteryzujących się mniejszą konserwatywnością, wystarczy, że będą konserwatywne w mniejszej liczbie ortologów, aby osiągnąć taką samą sumaryczną długość. Wynika to z tego, że długości gałęzi w takim drzewie reprezentują bardziej zróżnicowane UTRs i są one dłuższe niż te w podzbiorze i drzewie będącym bardziej konserwatywnym.

Na podstawie wielosekwencyjnych uliniowień miejsca wiązania uznanawane są za konserwatywne, jeśli pozostają zachowane we wszystkich genomach w ortologicznych lokalizacjach. Natomiast niekonserwatywne lub słabo konserwatywne, gdy nie występują lub mają zmiany w jednym z genomów. W przypadku, gdy w analizie wykorzystano aż 28 genomów

kręgowców powyższy warunek byłby zbyt surowy. Powodowałby on odrzucenie także konserwatywnych miejsc wiązań, które występują u większości, ale nie we wszystkich genomach. Może to być spowodowane specyficznością, odmiennością genetyczną, lub nawet błędami sekwencjonowania lub przeprowadzanych uliniowień. Aby zatem uwzględnić także te konserwatywne miejsca wiązania opracowano specjalną metodę ilościową. Polega ona na zsumowaniu wszystkich gałęzi drzewa filogenetycznego z danym konserwatywnym miejscem. Drzewo filogenetyczne musi oczywiście obejmować analizowane gatunki. Wyznaczona wartość powinna dopowiadać ewolucyjnemu czasowi, jaki dzieli miejsca wiązania u uwzględnionych gatunków. Wartość progowa może być ustalona w przedziale od zera do sumy wszystkich długości gałęzi, stanowiąc kompromis między wysoką czułością a swoistością.

Następnym krokiem po identyfikacji konserwatywnych miejsc wiązań w ortologicznych lokalizacjach genów jest rozróżnienie tych miejsc, które zostały zachowane w wyniku doboru naturalnego od tych, które przypadkowo zachowały się niezmienione. Tutaj mają zastosowanie statystyki:

- Z-score,
- metody bootstrapu – generacja grupy reprezentatywnych sekwencji podobnych do miRNAs (kohorty).

Metody te jednak nie uwzględniają: lokalnych skłonności mutacyjnych, współczynnika konserwatywności dinukleotydów [60] i lokalnego współczynnika konserwatywności.

Autorzy parametru P_{CT} zaproponowali trzy innowacje w celu uniezależnienia się od przypadkowej konserwatywności lub spowodowania jej inną funkcyjnością. Dopasowanie k-merów uwzględnia ich zawartość GC i oczekiwaną konserwację wynikającą z jego struktury dinukleotydowej. Zróżnicowanie rodzajów miejsc wiązań i odjęcie poziomu sygnału i tła dłuższych miejsc wiązań od tych krótszych, które mogą być zawarte wewnątrz nich.

Dla każdego miejsca wiązania z dopasowania *seed* został wyliczony sygnał konserwatywności (sygnał) oraz konserwatywność podstawowa (szum). Wartość P_{CT} obliczana jest na podstawie wzoru:

$$P_{CT} = \begin{cases} \frac{S/_B - 1}{S/_B} \\ \approx 0 \; dla \; S/_B < 1 \end{cases} \quad \textbf{(3-6)}$$

gdzie:

S– sygnał, czyli współczynnik konserwatywności powstaje przez sumowanie wszystkich długości gałęzi, które zawierają konserwatywne miejsce wiązania;

B– poziom podstawowy (szum) średnia wartość współczynnika konserwacji danego drzewa.

Konserwatywność miejsca wiązania: filogenetyczne długości gałęzi wszystkich gatunków zawierających to miejsce.

Parametr P_{CT} został wyznaczony dla wszystkich wysoce konserwatywnych miRNAs.

$$P_{CT} = 1 - (\; (1 - P_{CT})_{site1} \; x \; (1 - P_{CT})_{site2} \; x \; (1 - P_{CT})_{site3} \ldots \;) \quad \textbf{(3-7)}$$

3.2.4 Bazy potwierdzonych targetów

Do przeprowadzenia walidacji testowanego w ramach opracowania własnego modelu predykcji targetów niezbędne jest uzyskanie danych wzorcowych, które pozwolą na jego weryfikację. Pośród zasobów otwartych, które można użyć w tym celu, można wymienić następujące:

1. miRTarBase
   http://mirtarbase.mbc.nctu.edu.tw

Baza mirTarBase zawiera publicznie dostępne dane zweryfikowane eksperymentalnie. Zawiera ona informacje o interakcjach miRNA/mRNA. Powstała na podstawie manualnej eksploracji danych literaturowych dot. studiów nad funkcjonowaniem miRNA po uprzednio przeprowadzonej procedurze data mining. Informacje zawarte w udostępnianych zasobach to: identyfikator miRNA, gatunek biologiczny: miRNA i genu, symbol genu, rodzaj eksperymentu. Metody walidacji: *western blot, Luciferase assay, pSILAC, Microarray*, NGS.

2. TarBase (Papadopoulos GL et al., Nucleic Acids Res. 2009)
   http://diana.cslab.ece.ntua.gr/tarbase/

Aktualnie TarBase v7.0 określa się, jako największa manualnie przetwarzana baza targetów. Informacje te uzyskane są ze specyficznych analiz, wielkoprzepustowych typu mikromacierzowego i proteomicznych. Szczególną uwagę zwrócono na targety pochodzące z eksperymentu sekwencjonowania takich jak HITS-CLIP i PAR-CLIP. Informacje zawarte w udostępnianych zasobach: identyfikator miRNA, symbol genu, rodzaj eksperymentu (*reporter gene*, *nothern blot, western blot*, qPCR, *proteomic*, *microarray, sequencing, degradome seq*, inne).

3. miRecords (Xiao F et al., Nucleic Acids Res. 2009)
   http://c1.accurascience.com/miRecords

miRecords zawiera eksperymentalnie potwierdzone targety pochodzących ze skrupulatnej analizy literatury. Cechą charakterystyczną tego zestawienia jest różnorodność dostarczanej informacji związanej z metodą eksperymentalną walidacji. Jako jedyna podaje ona miejsce przyłączania się pierwszego nukleotydu miRNA (target_site_position). W jej ramach przeprowadzono także metanalizę eksperymentalnych metod walidacji. Niestety zasób zawiera liczne błędy wyłapane przy analizie danych: błędy literowe w oznaczeniach, błędy w adresacji miejsc wiązań, szereg lokalizacji wskazuje na położenie *target site position* leżące poza obszarem transkryptu. Stwierdzono to na podstawie przeprowadzonej wyrywkowej weryfikacji danych.

4. miR2Disease (Jiang Q. et al, Nucleic Acids Res. 2009)
   http://www.mir2disease.org/

miR2Disease jest manualnie uzyskiwanymi informacjami o interakcjach miRNA/mRNA, szczególnie regulacji RNAi zachodzącej w stanach patologicznych. Dostarcza ona informacji o korelacjach miRNAs z niektórymi jednostkami chorobowymi.

3.3 Realizacje bioinformatyczne

Implementacje algorytmów predykcji targetów stanowią istotne narzędzie w pracy badacza zajmującego się genetyką. Bioinformatyczne narzędzia stworzone do tego celu zostały uznane za niezbędne na etapie poprzedzającym eksperymentalne potwierdzenie uzyskanych wyników. Dzięki tym narzędziom uzyskuje się zbiór par miRNA/target uporządkowanych wg odpowiednich

rang, wokół których jeśli w ogóle, koncentruje się weryfikacja par na poziomie eksperymentalnym. Algorytmy te znacząco redukują liczbę potencjalnych dupleksów i odrzucają te najmniej prawdopodobne. Korzyścią ich zastosowania jest więc znacząco mniejsza pula dupleksów do dalszej weryfikacji.

Najistotniejsza zdaniem autora niniejszej pracy kategoryzacja narzędzi predykcji targetów dokonuje się w domenie funkcyjności. Narzędzia możemy wówczas podzielić na dwie kategorie: predykcja targetów w ogóle i predykcja funkcyjnych targetów. Narzędzia z pierwszej kategorii na podstawie danej grupy miRNAs oraz puli transkryptów typują wszystkie prawdopodobne targety. Druga kategoria obejmuje narzędzia, które w konkretnym eksperymencie wskazują najbardziej prawdopodobne targety na podstawie informacji precyzującej dane doświadczenie. Uzyskane w niej predykcje powinny zatem stanowić podzbiór predykcji uzyskanych przy użyciu narzędzi zaliczanych do kategorii pierwszej. Informacje precyzujące dane doświadczenie to przede wszystkim poziom ekspresji genów, specyficzność tkankowa, wpływ czynników użytych w doświadczeniu na czynniki transkrypcyjne, rodzaj wprowadzonego czynnika do komórek, itd. W celu jego realizacji wykorzystuje się dane pochodzące z technik wysokoprzepustowych (mikromacierze DNA, NGS) oraz rezultaty narzędzi pierwszej kategorii.

Etap eksperymentalny analizy polega na weryfikacji uzyskanych targetów i pozwala uniknąć ograniczeń, czy błędów, jakie wprowadzają metody wysokoprzepustowe. Do tych ograniczeń należą: brak uwzględnienia regulacji potranslacyjnej genów oraz inne regulacje poza interferencją RNA, koregulacje genów, zależności występujące w sieci powiązań wzajemnych miRNAs.

W predykcji targetów w ogóle algorytmy analizują komplementarność całego miRNA z odpowiednim rejonem transkryptu. Ich celem jest predykcja i wytypowanie miejsc wiązań, które potencjalnie mogą stanowić region funkcyjny. Najczęściej brane są pod uwagę regiony 3'UTR transkryptu. W drugim etapie w celu poprawy predykcji w obliczeniach uwzględnia się jeden lub więcej czynników: siłę wiązania się dupleksu, liczbę par Watson-Cricka w obrębie *seed*, stopień komplementarności fragmentu 3' miRNA, stopień konserwatywności miejsc wiązań i/lub oszacowanie energii swobodnej powstałego dupleksu.

Przykładowe realizacje narzędzi tego typu zestawia Tabela 6. Każde tam występujące rozwiązanie proponuje odmienne algorytmy działania oraz odmienne metody klasyfikacji transkryptów. Definiują one własny współczynnik fałszywie pozytywnych i fałszywie negatywnych predykcji. Przedstawione metody dostarczają różniące się miedzy sobą predykcje. Stopień pokrywania się ich zbiorów wynikowych jest zmienny i czasami bywa mały lub nawet żaden [151].

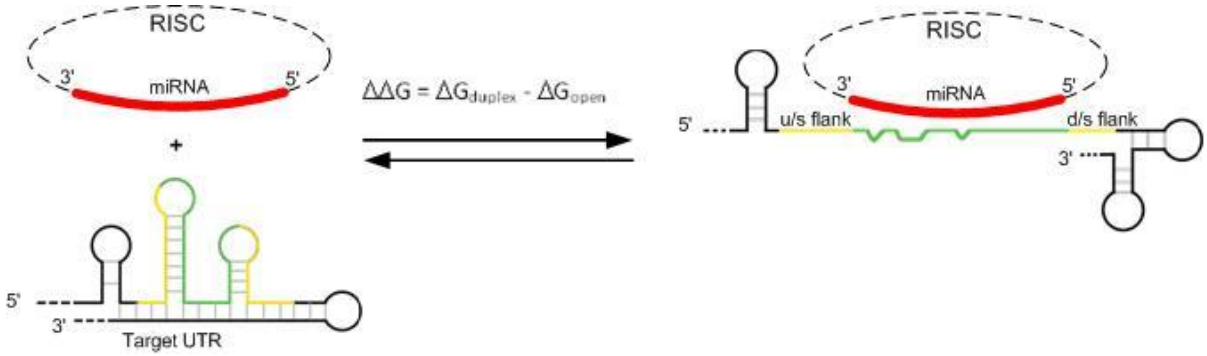
Tabela 6. Wybrane, charakterystyczne algorytmy predykcji targetów

Metoda	Rodzaj metody	Ref	Dostępność	Adres usługi
DIANA microT	kontekstowość konserwatyzm	(Kirakidou et al., 2004)	online	http://diana.cslab.ece.ntua.gr/
miRanda	komplementarność termodynamika	(John et al., 2004)	lokalnie	http://www.microrna.org
PITA	termodynamika	Kertesz et al., 2007		http://genie.weizmann.ac.il/pubs/mir07/mir07_dyn_data.html
MiRtarget2	Support Vector Machine (SVM)	(Wang and El Naqa, 2008)	online	http://mirdb.org
miRWalk 2.0	algorytm szukający *seeds*	Dweep, H et al. 2015	online	http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/index.html
PicTar	termodynamika	(Krek et al., 2005)		http://pictar.mdc-berlin.de/
RNAHybrid	termodynamika i model statystyczny	(Rehmsmeier et al., 2004)	lokalnie	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid
Target Scan	komplementarność *seeds*	(Lewis et al., 2005)	online	http://www.targetscan.org
mirSVR	model regresji	Betel D. et all 2010		http://www.microRNA.org
ComiR	SVM	C. Coronnello et all 2012	online	http://www.benoslab.pitt.edu/comir/
RNA22	*pattern recognition*	Miranda, KC et al 2006)	online	https://cm.jefferson.edu/rna22/Interactive/

Krótka charakterystyka wybranych narzędzi predykcji targetów z pierwszej kategorii:

1. Program **miRanda** realizuje wyszukiwanie w dwóch etapach: uliniowienia sekwencji cząsteczek miRNA i mRNA metodą programowania dynamicznego. Stopień dopasowania sekwencji oceniany jest na podstawie stopnia komplementarności lokalnego dopasowania. Punktacja uwzględnia różne rodzaje komplementarności par nukleotydów oraz rozbudowany system oceny dopasowania: kara za początek gapu, jego wydłużenie, punktacja za dopasowanie, stopień znaczenia dopasowania w rejonie *seed*. Oprócz par Watson-Cricka uwzględnia pary niekomplementarne (*wobble*) G:U.Drugi etap działania programu polega na oszacowaniu stabilności termodynamicznej lokalnego dopasowania. Wyznaczenie energii swobodnej odbywa się na podstawie wygenerowanej fikcyjnej jednoniciowej sekwencji i obliczeniu struktury w pakiecie ViennaRNA [112].
2. Algorytm **PITA** koncentruje się na strukturze przyjmowanej przez parujące się cząsteczki i wynikającą z nich fizyczną dostępność do regionu miejsca wiązania. Najpierw więc

określa strukturę drugorzędową cząsteczki mRNA a szczególnie jej końca 3'UTR. W metodzie wprowadzono model *parameter-free model*, który dotyczy interakcji miRNA/mRNA, i który oblicza różnicę pomiędzy energią swobodną uzyskiwaną przez dupleks i kosztami energetycznymi "rozprostowania" regionu miejsca wiązania właśnie w celu jego dostępności (Rys. 3.4).



Rys. 3.4. Ilustracja interakcji miRNA/mRNA. Punktacja ΔΔG obliczana jako energia swobodna uzyskana przy przejściu ze stanu w którym miRNA i mRNA są zwinięte (lewa strona rysunku) i stanem w którym miRNA jest związany z jego targetem (prawa strona). (Rysunek wzorowany na ilustracji z bibliografii [31]).

Program PITA przeprowadza obliczenia dla wskazanego mRNA i sekwencji odpowiadającego mu regionu 3'UTR oraz zbioru miRNAs. Najpierw region 3'UTR zostaje przeskanowany w poszukiwaniu potencjalnych miejsc wiązań korzystając z parametrów: długości *seed*, stopnia niedopasowania w rejonie *seed*, kontekstu *seed*. Następnie dla tych regionów przeliczana jest punktacja podstawowym algorytmem PITA [88].

3. **TargetScan** predykcję opiera na poszukiwaniu różnych rodzajów miejsc wiązań w obrębie 3'UTR transkryptów, które wiążą region *seeds* miRNA. Konserwatywność tych miejsc wiązań uprzednio została wyznaczona (patrz rozdział 3.2.3). Przerwy (*gaps*) w dopasowaniu *seed* mogą być skompensowane przez konserwatywność pozostałych fragmentów miRNA: 3' oraz środkowej części. Ranking uzyskanych wyników uwzględnia informację kontekstową (metoda *context++ scores*) oraz, jeśli taka opcja zostanie wybrana, także konserwatywność targetów.
4. Metoda **miRWalk** stanowi właściwie bazę interakcji miRNA/mRNA opracowaną na podstawie własnego algorytmu miRWalk do przewidywania miejsc wiązań na kompletnej sekwencji genów (także mitochondrialnych) porównując je z wynikami innych 12 programów. Dodatkowo zawiera informację o 449 ścieżkach biologicznych i 2356 zaburzeniach zdrowotnych z bazy OMIM - *Online Mendelian Inheritance in Men*). Następnie informacje o potwierdzonych interakcjach miRNA/mRNA w połączeniu do genów, ścieżek, chorób, organów, zaburzeń OMIM, linii komórkowych. Algorytm podąża wzdłuż sekwencji transkryptu wyłapując heptamery perfekcyjnie komplementarne do regionu *seed*. Po znalezieniu dopasowania poszerza długość dopasowania aż do uzyskania pierwszego braku dopasowania. W efekcie tego spaceru algorytm zwraca wszystkie możliwie dopasowania o długości 7 lub więcej nukleotydów. Następnie uzyskane dopasowania są rozdzielane wg regionu położenia: promotor, 5'UTR, CDS, 3'UTR, mitochondrialny. Rozkład prawdopodobieństwa losowych dopasowań subsekwencji w analizowanej sekwencji jest obliczana na podstawie rozkładu Poissona. Należy oczekiwać, że dłuższa perfekcyjna komplementarność *seed* jest powiązana z

niższym prawdopodobieństwem. Normalizacja punktacji za dopasowanie względem długości targetu i miRNA.

5. Algorytm **RNA22** oparty jest na opracowanym wzorcu miejsc wiązań, który służy do wyszukiwania targetów w sekwencji, a dopiero w dalszej kolejności dopasowania do danego targetu miRNA. Na podstawie analizy sekwencji znanych miRNAs algorytmem Teiresias, uzyskano wzorzec, który następnie przetworzono na odwrotnie komplementarny, umożliwiający zastosowanie go na sekwencjach transkryptów. Dzięki takiemu podejściu w tej metodzie możliwe jest uzyskanie miejsc wiązań, które "należą" do jeszcze nieodkrytych cząsteczek miRNAs. Pozwala ustawić próg czułości i swoistości, rodzaj miejsca wiązania, liczbę sparowanych zasad, energię wiązania, pary G:U w obrębie miejsca wiązania.
6. **ComiR** (*Combinatorial miRNA targeting*) sprawdza, jakich miRNAs targetem jest dany mRNA. W tym celu wykorzystuje informację o ekspresji miRNA s oraz rezultaty predykcji czterech algorytmów: miRanda, PITA, TargetScan, mirSVR. Uzyskane punktacje z tych czterech algorytmów są wykorzystane, jako składowe do klasyfikacji metodą *support vector machine* (SVM) w celu określenia targetów [30].
7. Narzędzie **RNAhybrid** jest dostępny *on line* http://bibiserv.techfak.uni-bielefeld.de/rnahybrid. Pozwala na ustawianie różnych użytecznych opcji np. odrzucania par G:U w regionie *seed* lub opcję forsowania długości *seed*. Podstawowy algorytm jest pewną wariacją algorytmów predykcji II rzędowej struktury. Jednak przeciwieństwie do innych rozwiązań określa on najbardziej preferowaną hybrydyzację obu parowanych cząsteczek.

Druga kategoria narzędzi poprawia jakość predykcji oraz dokonuje oceny funkcyjności miRNAs właśnie przez doprecyzowanie warunków eksperymentalnych. Metody z tej kategorii opierają się na hipotezie, że regulacyjna aktywność miRNAs może mieć wyraz w zmianach ekspresji transkryptów będących ich targetami. Weryfikacja tej hipotezy opiera się na pomiarze zmian ekspresji genów w komórkach po transfekcji lub inhibicji konkretnych miRNAs [109][5].

Informacja o poziomie ekspresji transkryptów uzyskana z różnych wielkoskalowych technik wymaga wstępnego przetworzenia. Dlatego pełną realizację rozpoznawania transkryptów przez te narzędzia można przeprowadzać poprzez platformy webowe oferujące przejście poprzez kolejne etapy aż po bezpośrednią metodę integracji danych o ekspresji. Można tu wskazać takie platformy jak:

- Babelomics (http://www.babelomics.org/),
- GeneSpring GX (http://www.silicongenetics.com/),
- Platforma Integromicznych Analiz Danych z Mikromacierzy DNA https://lifescience.plgrid.pl/pl/users/sign_in.

Przykłady narzędzi, które "same" integrują dane o ekspresji, czyli umożliwiają funkcjonalną interpretację ekspresji mRNA i miRNA zawiera

Tabela 7. Wykorzystanie profilów ekspresji miRNA i transkryptów pozwala na predykcję par miRNA/mRNA przez identyfikację par o odwrotnej korelacji ich ekspresji.

Tabela 7. Metody integracji danych o ekspresji w predykcji targetów

Metoda	Rodzaj metody	Ref	Dostępność	Adres usługi
TopKCEMC	*Cross entropy Monte Carlo*	Lin and Ding 2009	lokalnie	http://www.stat.osu.edu/~statgen/SOFTWARE/TopKCEMC/
SigTerms	asocjacja klas genów	Creighton et al. 2008	lokalnie (MS Excel)	http://sigterms.sourceforge.net/
miRGen++	wnioskowanie Bayesa	Huang et al., 2007b	lokalnie (Matlab)	http://www.psi.toronto.edu/genmir
TargetScore	wnioskowanie Bayesa	Li Y. et all 2014	lokalnie	http://www.bioconductor.org/packages/release/bioc/html/TargetScore.html
Roleswitch	algorytm iteracyjny	Li Y. et all 2014	lokalnie	http://www.bioconductor.org/packages/release/bioc/html/Roleswitch.html

Krótka charakterystyka wybranych narzędzi predykcji targetów w drugiej kategorii:

1. Metoda **TopKCEMC** (http://www.stat.osu.edu/~statgen/SOFTWARE/TopKCEMC/) integruje rezultaty pochodzące z różnych analiz tych samych danych. Każdy reprezentowany jest przez listę rankingową. Algorytm globalnej optymalizacji (Crosss Entropy Monte Carlo) znajduje jedną optymalną listę łączącą wszystkie pozostałe [110]. Realizuje on iteratywne przeszukiwanie, aż do uzyskania optymalnej listy, która minimalizuje sumę ważonych odległości między proponowaną listą, a każdą z wejściowych list rankingowych. Odległość między dwoma listami mierzona jest zmodyfikowaną miarą *Kendall's tau* oraz *Spearman's footrule* [53]. Znajduje zastosowanie do analizy wyników predykcji targetów oraz list genów o różnicowej ekspresji.
2. Metoda **GenMIR** (http://www.psi.toronto.edu/genmir/) wykorzystuje sieć Bayesa i nauczanie maszynowe. Algorytm bierze pod uwagę wzorce ekspresji genu używając ekspresyjne dane miRNA i zbiór kandydatów na targety. W efekcie zbiór funkcjonalnych targetów uzyskujemy z danych przeliczonych algorytmem Bayesa. Za pomocą tego modelu ekspresja transkryptów będących targetami może być wyjaśniona przez regulacyjną aktywność wielu miRNAs. GenMiR++ pozwala na akuratną identyfikację targetów miRNA z sekwencji i danych ekspresyjnych i pozwala uzyskać istotną liczbę eksperymentalnie zweryfikowanych targetów.

3.4 Model probabilistyczny – wnioskowanie bayesowskie

Model statystyczny stanowi rozszerzenie modelu matematycznego na przypadki, kiedy zmienne i parametry modelu podlegają fluktuacjom, których wartości są znaczące w porównaniu z samymi zmiennymi. W takich sytuacjach rolę zmiennych przejmują zmienne losowe, które przyporządkowują zmiennym określone wartości prawdopodobieństwa. Zmienna losowa w wyniku doświadczenia lub predykcji przyjmuje różne wartości z określonym prawdopodobieństwem. Do takich zmiennych zaliczają się wykorzystywane także w niniejszej

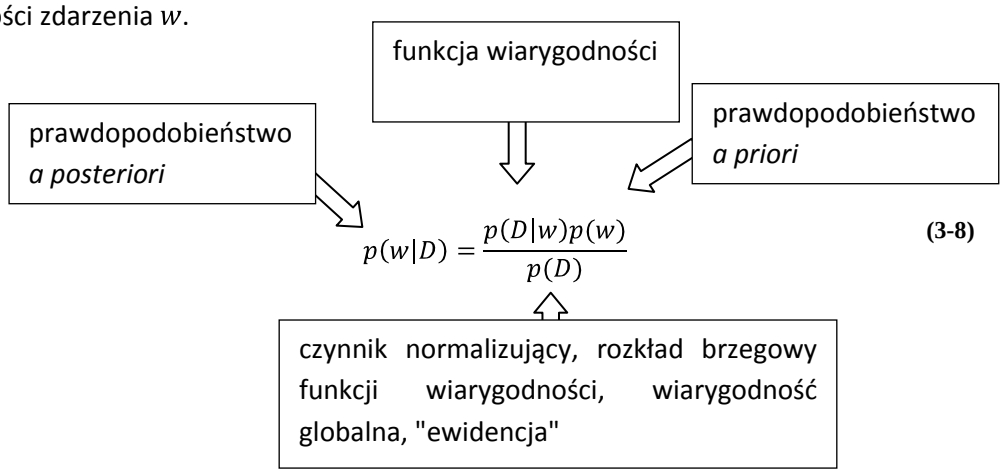
pracy dane mikromacierzowe, ze względu na mnogość zakłóceń i artefaktów pojawiających się w wieloetapowej procedurze eksperymentalno-obliczeniowej.

3.4.1 Uogólnienie twierdzenia Bayesa

Historycznie wyróżnia się dwa podejścia określające pojęcie prawdopodobieństwa. Podejście klasyczne - częstotliwościowe (Abrahama de Moivre'a [37]) oznacza prawdopodobieństwo wystąpienia zdarzenia losowego, jako częstość występowania tego zdarzenia w odpowiedniej liczbie przeprowadzonych identycznych prób. Zakłada ono, że możemy dowolnie zwielokrotniać nasze doświadczenie w celu określenia prawdopodobieństwa, więc znajduje zastosowanie tylko dla powtarzalnych doświadczeń. Definicja częstotliwościowa nie precyzuje, jaka liczba doświadczeń jest wystarczająca, aby uzyskany iloraz częstości reprezentował prawdopodobieństwo.

W podejściu klasycznym zakładamy identyczność prób i jednocześnie możliwość ich dalszego przeprowadzania. W tym podejściu nie określa się ilości prób odpowiedniej dla danego doświadczenia. Ze względu na niemożliwość rozróżnienia próby w serii doświadczeń uzyskane prawdopodobieństwo stanowi miarę niepewności (lub wiarygodności). Wprowadzenie do prawdopodobieństwa i prawdopodobieństwo zmiennej losowej ciągłej zostały wyjaśnione w Dodatku A i B.

Thomas Bayes propagował podejście oceny niepewności, które pozwala na badanie zdarzenia w kontekście innych czynników wywierających wpływ na dane zdarzenie. W świetle przeprowadzonych badań uwzględnienie innych czynników zdarzenie ocenione prawdopodobieństwem *a priori* zostaje przekształcone w tzw. prawdopodobieństwo *a posteriori*. Stąd uogólnienie twierdzenie Bayesa dotyczy oszacowania niepewności dla przypadków, kiedy nie możemy zwielokrotniać doświadczenia. Podejście Bayesa umożliwia: badanie zdarzeń powtarzalnych i niepowtarzalnych, predykcję zdarzenia. Opisuje niepewność zdarzenia bez znajomości czynników, przyczyn tej niepewności. Przypadkowość jest traktowana, jako ograniczoność informacji, jaką posiadamy. Podejście to uwzględnia ilość prób doświadczenia, odnosi się do faktycznych danych [9]. Pozwala ono szacować niepewność zdarzenia w postaci prawdopodobieństwa *a posteriori* $p(w|D)$ po tym jak zaobserwowaliśmy dane D. Prawdopodobieństwo *a priori* zostaje skorygowane przez uwzględnienie danych D przy założeniu słuszności zdarzenia w.



Wartość $p(D|w)$ po prawej stronie wzoru Bayesa jest szacowana dla obserwowanych danych D i może być rozpatrywana, jako funkcja parametrów wektora w. Wyznacza ona

prawdopodobieństwo uzyskania danych D przy założeniu prawdziwości wektora w. W takiej sytuacji człon $p(D|w)$ nazywa się funkcją wiarygodności (*likelihood function*). Ta funkcja wyraża jak prawdopodobne są obserwowane dane dla różnych wartości parametrów wektora w. W celu estymacji funkcji wiarygodności używa się metody maksymalizacji prawdopodobieństwa. Wybierane są wówczas takie wartości wektora w, aby uzyskać maksymalną wartość tej funkcji. $p(w)$ - jest prawdopodobieństwem *a priori*, ustalanej arbitralnie przed uzyskaniem danych D. $p(w|D)$ jest prawdopodobieństwem *a posteriori*, czyli prawdopodobieństwem $p(w)$ skorygowanym funkcją wiarygnodności, czyli stan naszej wiedzy po uwzględnieniu danych D. $p(D)$ stanowi czynnik normalizujący, niezależny do wartości wektora w.

3.4.2 Rozkład mieszany Gaussa

Kombinacja liniowa podstawowych rozkładów takich jak rozkład Gaussa (patrz Dodatek C) nazywana jest rozkładem mieszanym. Dla K rozkładów Gaussa wzór na wypadkową gęstość prawdopodobieństwa:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu_k}, \Sigma_{\mathbf{k}}) \tag{3-9}$$

gdzie:

π_k - współczynnik mieszania.

Każdy składnik - rozkładu Gaussa $N(\boldsymbol{x}|\boldsymbol{\mu_k}, \Sigma_{\mathbf{k}})$ zawarty we wzorze (3-9) zwany jest komponentem modelu i charakteryzuje się własną średnią $\boldsymbol{\mu_k}$ oraz kowariancją $\Sigma_{\mathbf{k}}$. Jeśli scałkujemy obustronnie to równanie względem $\boldsymbol{x}$ przy założeniu, że $p(\boldsymbol{x})$ oraz każdy komponent Gaussa jest znormalizowany uzyskamy:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{3-10}$$

Wymaganie $p(\boldsymbol{x}) \geq 0$ oraz $N(\boldsymbol{x}|\boldsymbol{\mu_k}, \Sigma_{\mathbf{k}}) \geq 0$ oznacza, że $\pi_k \geq 0$ dla wszystkich k. Uwzględniając te warunki oraz (3-10) uzyskujemy:

$$0 \leq \pi_k \leq 1 \tag{3-11}$$

W ten sposób uzasadniliśmy, że współczynnik mieszania spełnia warunki prawdopodobieństwa.

Rozkład brzegowy uzyskany z reguły sum oraz reguły iloczynów jest dany wzorem:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(k)p(\boldsymbol{x}|k) \tag{3-12}$$

Porównując go z (3-9) uzyskujemy $\pi_k = p(k)$ jako prawdopodobieństwo *a priori* wylosowania k-tego komponentu oraz gęstość $N(\boldsymbol{x}|\boldsymbol{\mu_k}, \Sigma_{\mathbf{k}}) = p(\boldsymbol{x}|k)$ jako prawdopodobieństwo $\boldsymbol{x}$ pod warunkiem k. W takiej sytuacji prawdopodobieństwo *a posteriori* $p(k|\boldsymbol{x})$ możemy wyznaczyć z twierdzenia Bayesa:

$$p(k|\boldsymbol{x}) = \frac{p(k)p(\boldsymbol{x}|k)}{\sum_l p(l)p(\boldsymbol{x}|l)}$$
$$= \frac{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu_k}, \Sigma_\mathbf{k})}{\sum_l \pi_l N(\boldsymbol{x}|\boldsymbol{\mu}_l, \Sigma_\mathbf{l})} \quad \textbf{(3-13)}$$

3.4.3 Algorytm maksymalizacji wartości oczekiwanej dla mieszanego modelu Gaussa

Algorytm maksymalizacji wartości oczekiwanej (*expectation-maximization*- EM) [40] oszacowuje parametry modelu w sposób iteratywny, zaczynając od pewnych przyjętych (estymowanych) wartości początkowych. Każda iteracja zawiera krok *expectation* (E), który odnajduje rozkład ukrytych zmiennych, mając znane wartości dla zmiennych obserwowalnych i bieżące estymowane wartości parametrów. Krok maksymalizacji (M), który reestymuje parametry tak, aby odpowiadały maksymalnemu prawdopodobieństwu, przy założeniu, że rozkład znaleziony w kroku E jest poprawny. Można wykazać, że każda taka iteracja poprawia wyznaczoną wartość prawdopodobieństwa lub pozostawia ją bez zmian, jeśli lokalne maksimum zostało osiągnięte.

Algorytm EM jest zatem metodą heurystyczną pozwalającą rozwiązać wyrażenia matematyczne, które ze względu na zbyt dużą liczbę niewiadomych nie można rozwiązać metodami analitycznymi. Wymaga on puli danych obserwowalnych, podania wyrażenia, funkcji pewnych nieznanych parametrów oraz ukrytych zmiennych, która podlega minimalizacji lub maksymalizacji. Funkcję tą może być pewna miara odległości lub funkcja wiarygodności. Nieznane parametry dotyczą tworzonych klastrów lub stanowią parametry rozkładów. Zmienne ukryte stanowią binarny wskaźnik przynależności do danego klastru. Znajduje on zastosowanie przy klasteryzacji, dopasowywaniu krzywych, rozkładów, itd.

Wykorzystanie algorytmu EM w mieszanym modelu Gaussa zmierza do wyznaczenia optymalnych parametrów mieszanego rozkładu Gaussa (3-9). Zakładamy, że dane eksperymentalne pochodzą z pewnej mieszaniny rozkładów o znanych rodzajach, ale nieznanych parametrach tych rozkładów. Jego wykorzystanie wymaga *a priori* podania liczby zmiennych losowych, czyli w tym przypadku liczby rozkładów składowych. Celem algorytmu jest maksymalizacja funkcji wiarygodności względem trzech parametrów: średnich, kowariancji i współczynników mieszania. Przebieg algorytmu [15]:

Przyjęcie wartości średnich $\boldsymbol{\mu}_k$, kowariancji Σ_k i wartości współczynnika mieszania π_k. Wyliczenie wartości początkowej logarytmu funkcji wiarygodności.

1. Krok E. Wyznaczenie odpowiedniości γ wykorzystując bieżące parametry:

$$\gamma(z_{nk}) = \frac{\pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}_j|\boldsymbol{\mu}_j, \Sigma_j)} \quad \textbf{(3-14)}$$

2. Krok M. Ponowne przeliczenie parametrów używając bieżących wartości γ:

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n \quad \textbf{(3-15)}$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})^T \quad \textbf{(3-16)}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{3-17}$$

3. Wyznaczenie wartości logarytmu funkcji wiarygodności.

$$\ln p(\boldsymbol{X}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) = \sum_{n=1}^{N} ln\left\{\sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)\right\} \tag{3-18}$$

4. Sprawdzenie współczynnika konwergencji parametrów lub logarytmu prawdopodobieństwa. Jeśli kryterium konwergencji nie jest spełnione następuje powrót do p. 2.

Wśród ograniczeń stosowania tego algorytmu wymienia się: konieczność założenia liczby składowych rozkładów normalnych modelu, wrażliwość algorytmu na wartości początkowe parametrów, które mogą prowadzić do odnalezienia lokalnego maksimum.

Algorytm EM może zostać uogólniony na potrzeby wnioskowania wariacyjnego.

3.4.4 Metoda wariacyjna we wnioskowaniu Bayesa

Metoda wariacyjna zastosowana do wnioskowania Bayesa (*Variational Baeysian Expectation Maximization* (VB-EM)) jest techniką aproksymacji skomplikowanych całek pojawiających się we wzorze Bayesa (w mianowniku) gdzie model statystyczny zawiera obserwowane zmienne – dane, nieznane parametry i ukryte zmienne. Stosuje się ją m.in. w aproksymacji rozkładu *a posteriori* zmiennych ukrytych we wnioskowaniu statystycznym. Wynikiem zastosowania tej metody jest analityczne, lokalno-optymalne dopasowanie aproksymowanego rozkładu. Uważa się ją za rozszerzenie zastosowań algorytmu EM, który estymuje pojedyncze parametry rozkładu *a posteriori* do pełnej, bayesowskiej estymacji, która oblicza, aproksymuje cały rozkład *a posteriori* parametrów i ukrytych zmiennych. Podobnie jak algorytm EM, metoda ta znajduje zbiór optymalnych parametrów modelu. Zaletą metody wariacyjnej jest uniknięcie przedopasowania (*overfitting*).

Operację, która każdej funkcji pewnej przestrzeni funkcyjnej R przypisuje pewną liczbę nazywamy funkcją funkcji lub funkcjonałem.

$$F: f \rightarrow F(f) \tag{3-19}$$

Przez analogię do operacji na funkcjach możemy określić pochodną funkcjonału:

$$\frac{dF}{df} \tag{3-20}$$

którą można wykorzystać do wyszukiwania ekstremów funkcjonału.

Najczęstszą operacją funkcjonału jest całka oznaczona funkcji. Rachunek wariacyjny zajmuje się szukaniem funkcji, dla których dany funkcjonał przyjmuje wartości ekstremalne. Najczęściej funkcjonał dany jest całką oznaczoną funkcji.

Koncepcja optymalizacji wariacyjnej w zastosowaniu do wnioskowania Bayesa zakłada, że mamy pełny bayesowski model, w którym wszystkie parametry podane są poprzez rozkład *a priori*. Model może mieć ukryte zmienne, jak również nieznane parametry. Zbiór wszystkich nieznanych wartości oznaczmy przez $\boldsymbol{Z}$. Podobnie wszystkie obserwowane zmienne przez $\boldsymbol{X}$. Na przykład

mamy zbiór N zmiennych niezależnych i parametrów o identycznych rozkładach, dla których: $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ oraz $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$**.** Łączny rozkład wynosi $p(\boldsymbol{Z}, \boldsymbol{X})$.

Wnioskowanie Bayesa realizowane jest na podstawie poniższego równania.

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z})}{\int p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z}} \qquad \textbf{(3-21)}$$

$$= \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{p(\boldsymbol{X})}$$

gdzie przez $\boldsymbol{Z}$ oznaczono zmienne ukryte, a przez $\boldsymbol{X}$ dane weryfikujące prawdopodobieństwo *a priori*.

Przekształcenie realizuje konwersję rozkładu *a priori* przez funkcję wiarygodności na postać *a posteriori*. Praktycznie w ocenie rozkładu *a posteriori* największą trudność sprawia wyznaczenie $p(\boldsymbol{X})$ - czynnika normalizującego lub inaczej rozkładu brzegowego funkcji wiarygodności, który stanowi mianownik we wzorze Bayesa. Aby rozwiązać ten problem stosuje się dwa podejścia: stochastyczne i analityczne (strukturalne) wykorzystujące metodę wariacyjną. W metodzie wariacyjnej aproksymuje się rozkład, który jest maksymalnie podobny do rozkładu *a posteriori*.

Naszym celem jest zatem znalezienie aproksymacji dla rozkładu *a posteriori* $p(\boldsymbol{Z}|\boldsymbol{X})$ poprzez wyznaczenie rozkładu brzegowego $p(\boldsymbol{X})$. Rozkład *a posteriori* ukrytych zmiennych $p(\boldsymbol{Z})$ mając dane $\boldsymbol{X}$ jest aproksymowany przez rozkład wariacyjny $q(\boldsymbol{Z})$:

$$p(\boldsymbol{Z}|\boldsymbol{X}) \approx q(\boldsymbol{Z}) \qquad \textbf{(3-22)}$$

Rozkład $q(\boldsymbol{Z})$ musi należeć do rodziny rozkładów prostszych niż $p(\boldsymbol{Z}|\boldsymbol{X})$, wybrany tak, aby można go było dopasować do rozkładu *a posteriori*. Stopień dopasowania rozkładów $q(\boldsymbol{Z})$ i $p(\boldsymbol{Z}|\boldsymbol{X})$ jest oceniany miarą – funkcją odległości między tymi dwoma rozkładami. Dlatego dobór tego rozkładu powinien minimalizować użytą miarę. Jako miarę najczęściej wykorzystuje się dywergencję Kullback-Leiblera. Dywergencja Kullbacka-Leiblera zwana jest też relatywną entropią i określa rozbieżność między dwoma rozkładami prawdopodobieństwa. W teorii informacji określa ona średnią ilość dodatkowej informacji przy transmisji wartości x podlegającej rozkładowi $p(x)$ o nieznanym rozkładzie do odbiornika za pomocą kodowania o teoretycznym rozkładzie $q(x)$ (więcej na ten temat w rozdziale 5.2.3).

Przekształcając wzór (3-21) uzyskujemy:

$$\ln p(X) = ln\frac{p(X,Z)}{p(Z|X)} \tag{3-23}$$

$$= \int q(Z)\, ln\frac{p(X,Z)}{p(Z|X)}\, dZ$$
$$= \int q(Z)\, ln\frac{p(X,Z)}{p(Z|X)}\frac{q(Z)}{q(Z)} dZ$$
$$= \int q(Z)\left(ln\frac{q(Z)}{p(Z|X)} + ln\frac{p(X,Z)}{q(Z)}\right) dZ$$
$$= \int q(Z)\ ln\frac{q(Z)}{p(Z|X)} dZ + \int q(Z) ln\frac{p(X,Z)}{q(Z)} dZ$$

$KL[q||p]$ — odległość między $q(Z)$ a $p(Z|X)$

$\mathcal{L}(q)$ — kres dolny

Logarytm wiarygodności globalnej $p(X)$ można zatem przedstawić w postaci:

$$ln\, p(X) = KL[q||p] + \mathcal{L}(q) \tag{3-24}$$

dywergencja ≥ 0 + kres dolny (łatwy do oszacowania dla danego q)

Powszechną praktyką przy aproksymacji rozkładu *a posteriori* jest jego fragmentacja (faktoryzacja) na niezależne partycje względem zmiennych ukrytych $Z = \{Z_1, \dots, Z_m\}$:

$$q(Z) = \prod_i^m q_i(Z_i) \tag{3-25}$$

gdzie: $q_i(Z_i)$ jest aproksymacją rozkładu *a posteriori* dla i-tego podzbioru parametrów.

Wówczas:

$$\mathcal{L}(q) = \int q(Z) ln\frac{p(X,Z)}{q(Z)} dZ \tag{3-26}$$
$$= \int \prod_i q_i \times \left(\ln p(X,Z) - \sum_i \ln q_i\right) dZ$$

Można wykazać używając rachunku wariacyjnego, że najlepszy rozkład q_j dla każdego fragmentu (partycji) uzyskujemy stosując wzór [15]:

$$q_j(Z_j) = \frac{e^{E_{i\neq j}[\ln\ p(Z,X)]}}{\int e^{E_{i\neq j}[\ln p(X,Z)]} dZ_j} \tag{3-27}$$

gdzie:

$E_{i\neq j}$ jest wartością oczekiwaną logarytmu rozkładu łącznego danych i zmiennych ukrytych dla wszystkich zmiennych poza analizowanym fragmentem.

Postać logarytmiczna powyższego wzoru:

$$\ln q_j(Z_j) = E_{i\neq j}[\ln p(Z, X)] + constant \quad \textbf{(3-28)}$$

Stała w powyższym wzorze jest wartością normalizującą (mianownik wzoru (3-27)). Wykorzystując własność wartości oczekiwanej, wyrażenie $E_{i\neq j}[\ln p(Z, X)]$ może zostać uproszczone do postaci funkcji określonej przez hiperparametry rozkładu *a priori* i zastrzeżeniu, ze ukryte zmienne nie są zawarte w obecnie analizowanym fragmencie. Takie podejście tworzy wzajemną zależność pomiędzy parametrami rozkładu zmiennych w jednym fragmencie, a zmiennymi w pozostałych partycjach. Sugeruje to zastosowanie algorytmu iteratywnego podobnego do tego stosowanego w EM, ale zmodyfikowanego, czy uogólnionego w obecnej metodzie. Algorytm ten gwarantuje zbieżność.

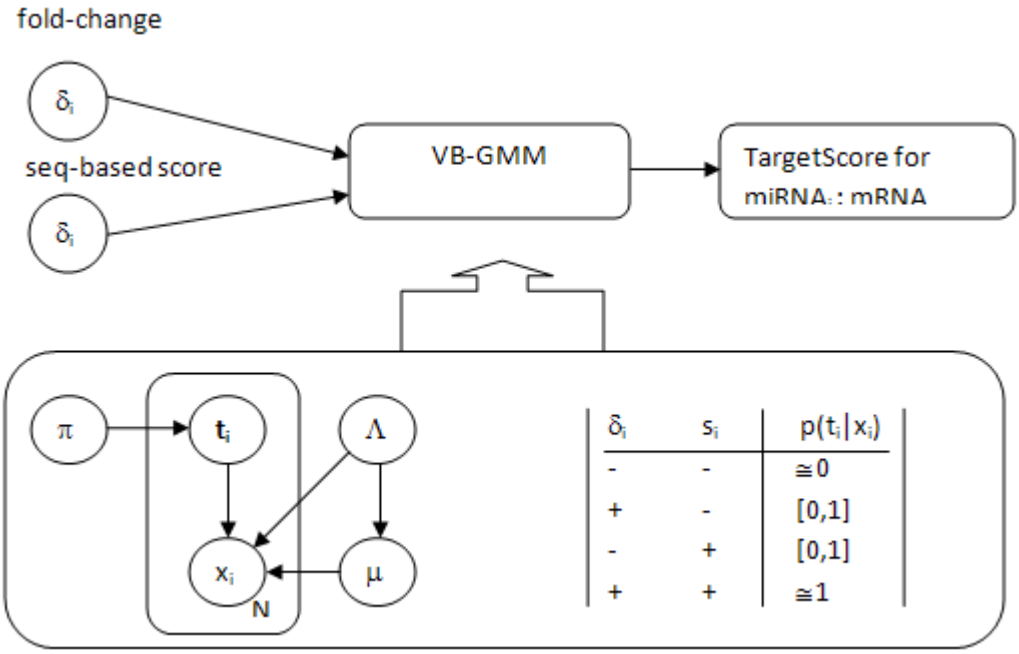
3.4.5 Model TargetScore

Przykładem wykorzystania algorytmu VB-EM przy analizie danych mikromacierzowych jest model TargetScore [104]. Biblioteka TargetScore oblicza prawdopodobieństwa interakcji jednego, konkretnego miRNA z całą pulą transkryptów. Dane wejściowe modelu stanowią:

1. wektor zmienności ekspresji każdego transkryptu o długości N wyrażony jako logarytm ilorazu poziomu ekspresji *fold change* (logFC) uzyskanych czy to z eksperymentu mikromacierzowego czy RNA-seq;
2. wektor od długości N punktacji kontekstowej *context score* (Cs) odpowiedni dla danego miRNA oraz transkryptów;
3. wektor prawdopodobieństwa konserwatywności regionu wiązania miRNA *probability of conserved targeting* (P_{CT}) również o długości N.

Pojęciem "punktacja cech sekwencji" (*sequence feature scores*) określa się obydwie punktacje Cs i P_{CT}. Na podstawie tych trzech wektorów, z czego można powiedzieć dwa stanowią parametry modelu, obliczane są prawdopodobieństwa transkryptów będących targetem danego miRNA.

Biblioteka TargetScore opiera się na modelu *Variational Bayesian –Gaussian Mixture Model* (VB-GMM), czyli zastosowaniu algorytmu VB-EM dla rozkładów Gaussa. Wnioskowanie wariacyjne oraz bayesowski model rozwiązują problemy, jakie wynikają z podejścia maksymalizacji dopasowania rozkładu do danych wejściowych (*maximum likelihood approach*). *Gaussian Mixture Model* stanowi liniową kombinację rozkładów gaussowskich. Metoda znalezienia rozwiązania maksymalizacji dopasowania rozkładu dla modelu ze zmiennymi ukrytymi to algorytm EM, a dokładnie jego generalizacja oparta na wnioskowaniu wariacyjnym. TargetScore integrując informację o zmienności ekspresji genów i punktacji cech sekwencji odpowiada na pytanie: jakie transkrypty w konkretnym eksperymencie podlegają regulacji konkretnym miRNA? Nie jest to więc pytanie ogólne o **wszystkie** cele danego miRNA, tylko konkretne pytanie szczegółowe.



Rys. 3.5. Schemat funkcjonowania modelu TargetScore. Znaczenia symboli: x_i – dane eksperymentalne, wartości zniezależne, π - współczynnik mieszania składowych rozkładów Gaussa, t_i–binarna zmienna K –wymiarowa, przyjmująca wartość $t_k = 1$, a dla pozostałych wymiarów wartości zero ($t_k \in \{0, 1\}$, $\sum_k t_k = 1$, Λ - macierz precyzji (odwrotność macierzy kowariancji) dla każdego rozkładu Gaussa, μ – wektor wartości średnich dla każdego rozkładu Gaussa.

Każda zmienna wejściowa (wektory): logFC, Cs, P_{CT} stanowi niezależną zmienną losową modelu VB-GMM. Model VB-GMM jest stosowany niezależnie dla każdej z nich. Wyboru metody wariacyjnej maksymalizacji dopasowywania rozkładu dokonano ze względu na brak efektu przedopasowania (*overfitting*) przy zastosowaniu tej metody. Założono dwa rodzaje podejścia zależne od typu zmiennej wejściowej. Trzy-komponentowy model VB-GMM zastosowano dla wektora logFC. Pozwala on wyróżnić transkrypty – targety charakteryzujące się ujemnymi wartościami logFC oraz targety o małej wartości dodatniej. Małe wartości dodatnie odnoszą się do transkryptów podlegających efektowi *off-target*. Drugie podejście dotyczy pozostałych dwóch zmiennych wejściowych, dla których zastosowano model dwu-komponentowy VB-GMM. Na Rys. 3.5 przedstawiono model w postaci acyklicznego grafu skierowanego. Graficzne sposoby przedstawienia sieci Bayesa omówiono w Dodatku D. Zależność między macierzą precyzji Λ a wektorem $\boldsymbol{\mu}$ wynika z faktu, że wariancja rozkładu Gaussa jest także funkcją macierzy precyzji. Optymalizację parametrów modelu dopasowywania uzyskano metodą EM. Komponenty z największą bezwzględną średnią wartością obserwowane dla ujemnych wartości logFC lub cech sekwencji (Cs i P_{CT}) są powiązane z targetami miRNA. Dlatego nazwano je *target component*. Pozostałe komponenty nazwano - *background component*. Wnioskowanie interakcji miRNA/mRNA przeprowadzono jako ekwiwalent rozkładu *a posteriori target component* dla obserwowanych zmiennych wejściowych. Rezultat, czyli prawdopodobieństwo liczone jest, jako średnia ważona przekształcenia sigmoidalnego logFC prawdopodobieństwa *a posteriori target component* po wszystkich składowych wejściowych [106].

Zakładamy, że mamy *N* genów, którym przyporządkowujemy wektory $\boldsymbol{x} = (x_1, \dots, x_N)^T$. Wektory te należą do przedziału $\boldsymbol{x} \in \{\boldsymbol{x_f}, \boldsymbol{x_1} \dots, \boldsymbol{x_L}\}$,

gdzie $\boldsymbol{x_f}$ - wartość logarytmu krotności zmiany poziomu ekspresji (logFC);

$\boldsymbol{x_1}, \ldots, \boldsymbol{x_L}$ – punktacja sekwencji, $\boldsymbol{L}$ – liczebność grup punktacji. W analizowanym przypadku *L = 2.*

Oznacza to, że każdy gen ma przyporządkowane *L+1* różne wektory wartości. W celu uproszczenia obliczeń przyjmuje się, że $\boldsymbol{x}$ reprezentuje jeden wektor ze zmiennych niezależnych.

W celu wnioskowania o genach będących targetami danego miRNA mając wartości wektora $\boldsymbol{x}$ należy uzyskać rozkład *a posteriori* $p(\boldsymbol{z}|\boldsymbol{x})$ zmiennej ukrytej $\boldsymbol{z} \in \{z_1, \ldots, z_K\}$, gdzie *K=3* gdy wartości $\boldsymbol{x}$ są zarówno dodatnie i ujemne – czyli dla logFC, oraz *K=2* gdy wartości $\boldsymbol{x}$ przyjmują tylko wartości dodatnie (punktacja sekwencji). Przez D oznaczmy wymiar modelu, czyli liczbę zmiennych, danych, $D = L + 1$.

Model TargetScore zawiera implementacje wielowymiarowego ($D > 1$) modelu GMM. I dla takiego przypadku został przedstawiony poniższy model. Niemniej w przeprowadzanych praktycznie obliczeniach dla każdego wektora danych $\boldsymbol{x}$ zastosowany jest jednowymiarowy model GMM ($D = 1$).

Zmienna ukryta $\boldsymbol{z}$ jest próbkowana z częstością $\boldsymbol{\pi}$ - współczynnik mieszania i przyjmuje się dla niej a *priori* rozkład Dirichleta $Dir(\boldsymbol{\pi}|\boldsymbol{\alpha_0})$ z parametrami $\boldsymbol{\alpha_0} = (\alpha_{0,1}, \ldots, \alpha_{,0,K})$. Aby obliczyć relatywną częstość targetów i nie-targetów w zbiorze genów przyjęto $\alpha_{0,1} = aN$ dla komponentu związanego z targetami, a dla pozostałych komponentów $\alpha_{0,k} = (1 - a) \times N(K - 1)$, gdzie $a = 0{,}01$.

Zakłada się dla $\boldsymbol{x}$ aprioryczny rozkład gaussowski $N\left(\boldsymbol{x}\middle|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right)$, gdzie $\boldsymbol{\Lambda}$ - macierz precyzji, która jest odwrotnością macierzy kowariancji. Wektor wartości oczekiwanych $\boldsymbol{\mu}$ oraz macierz precyzji $\boldsymbol{\Lambda}$ o łącznym rozkładzie $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ podlegają rozkładowi Gaussa-Wisharta:

$$\prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (\beta_0\Lambda)^{-1})\, W(\Lambda_k|\boldsymbol{W}_0, \upsilon_0) \tag{3-29}$$

gdzie jego parametry: $\{\boldsymbol{m}_0, \beta_0, \boldsymbol{W}_0, \upsilon_0\} = \{\widehat{\boldsymbol{\mu}}, , 1, \boldsymbol{I}_{DxD}, D + 1\}$. We wzorze tym $\widehat{\boldsymbol{\mu}}$ - wartość oszacowana; υ_0- stopnie swobody; $\boldsymbol{W}_0$ - macierz skali. Niech $\boldsymbol{\theta} = \{\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$. Można dokonać dekompozycji logarytmu rozkładu brzegowego funkcji wiarygodności na dwa składniki:

$$\ln p(\boldsymbol{x}) = \mathcal{L}(q) + KL(q||p) \tag{3-30}$$

$$\ln p(\boldsymbol{x}) = \int q(\theta)\, ln\frac{p(\boldsymbol{x}, \theta)}{q(\theta)} + \int q(\theta)\, ln\frac{q(\theta)}{p(\theta|\boldsymbol{x})} \tag{3-31}$$

gdzie:

$KL(\boldsymbol{q}||\boldsymbol{p})$ - dywergencja Kullbacka-Liblera;

$\mathcal{L}(\boldsymbol{q})$ - kres dolny;

$q(\theta)$ – proponowany rozkład $p(\boldsymbol{\theta}|\boldsymbol{x})$.

Ponieważ $\ln p(\boldsymbol{x})$ jest wartością stałą (normalizującą), maksymalizacja $\mathcal{L}(q)$ oznacza minimalizację $KL(q||p)$. Generalne optymalne rozwiązanie $\ln q_j^*(\theta_j)$ jest wartością oczekiwaną j z uwzględnieniem innych zmiennych, $\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{x}, \boldsymbol{\theta})]$.

W szczególności definiujemy:

$$q(\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = q(\boldsymbol{z})q(\boldsymbol{\pi})\boldsymbol{q}(\boldsymbol{\mu}, \Lambda) \quad \textbf{(3-32)}$$

Wartości oczekiwane tych trzech członów w skali logarytmicznej: $\ln q^*(\boldsymbol{z}), \ln q^*(\boldsymbol{\pi}), \ln q^*(\boldsymbol{\mu})$, mają taką samo postać jak pierwotny rozkład (tzw. rozkłady sprzężone). Należy dokonać oszacowania parametrów $\{\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$, które zależą od wartości oczekiwanej dla $\boldsymbol{z}$ lub rozkładu *a posteriori*:

$$p(z_{nk}|\boldsymbol{x}_n, \theta) \equiv \mathbb{E}[x_{nk}] = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}} \quad \textbf{(3-33)}$$

gdzie:

$$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln|\Lambda_{\boldsymbol{k}}|] - \frac{D}{2} ln\,(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_{k,\Lambda_k}}[(\boldsymbol{x}_n - \mu_k)^T \Lambda_{\boldsymbol{k}}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)] \quad \textbf{(3-34)}$$

Wewnętrzna zależność między wartościami oczekiwanymi i parametrami modelu wyznaczona jest poprzez algorytm VB-EM. Polega on na przypisaniu wartości początkowych wynikających z apriorycznych rozkładów i losowych danych o średniej $\boldsymbol{\mu}$. W i-tej iteracji oszacowaniu podlega równanie (3-33) używając parametrów modelu (krok VB-E), a następnie aktualizacja parametrów modelu używając (krok VB – M). Iteracja zostaje zatrzymana, kiedy $\mathcal{L}(q)$ przyrasta w danym kroku o wartość mniejszą niż ta przyjęta (default 10^{-20}).

Wartość prawdopodobieństwa, że dany gen jest targetem danego miRNA definiowane jest wzorem integrującym uzyskane prawdopodobieństwa dla poszczególnych zbiorów wejściowych (3-33) oraz bezpośrednich wartości logFC konwertowanych funkcją sigmoidalną:

$$targetScore = \sigma(-logFC)(\frac{1}{1+K}\sum_{x\in\{x_f, x_1, \ldots, x_L\}} p(t|\boldsymbol{x})) \quad \textbf{(3-35)}$$

gdzie:

$\sigma(-logFC) = \frac{1}{1+\exp(logFC)}$;

$p(t/\boldsymbol{x})$ – prawdopodobieństwo *a posteriori.*

Integracja danych wejściowych, które uwzględniają konkretny profil transkryptowy tkanki i hodowli komórkowej wyrażony skrótem logFC oraz zebrane i oszacowane informacje kontekstowe miejsca wiązania targetu Cs, informacje filogenetyczne o konserwatywności regionu oraz miRNA czyli P_{CT}, stanowi główną zaletę opisywanego modelu. Wg autora biblioteki wykorzystanie tych parametrów pozwala na redukcję liczby przeprowadzanych prób w eksperymencie. Uwzględnienie całego zbioru logFC wyklucza problem ustalania progu statystycznie istotnej zmienności ekspresji.

Biblioteka TargetScore posiada zastosowanie przede wszystkich w eksperymentach wykorzystujących transfekcję. Transfekcja jest procedurą wprowadzania obcej cząsteczki polinukleotydu tut. miRNA do komórki eukariotycznej. Tego rodzaju eksperyment naśladuje rodzaj sygnalizacji międzykomórkowej. Rejestrowanie techniką mikromacierzy transkryptomu – dokładniej jego zmienności, może stanowić weryfikację odpowiedzi komórki na podany bodziec.

Inaczej to wyrażając, można przyjąć, że jest to forma sygnalizacji jednokierunkowej między naukowcem, a odseparowanym fragmentem tkanki biologicznej. Pojęcie sygnalizacji międzykomórkowej ogranicza się do cząsteczkowego nośnika.

Rejestrując poziom ekspresji transkryptów przed i po transfekcji można wnioskować na temat targetów danego miRNA. Z punktu widzenia technicznego wydaje się być to rozwiązanie najkorzystniejsze przy poszukiwaniu targetów danego miRNA. Tym bardziej, że funkcjonowanie TargetScore'a uwzględnia w pewnym stopniu także efekt *off-target*, który dotyczy nieswoistego czy niezamierzonego blokowanie ekspresji innych genów.

4 Definiowanie modelu biocybernetycznego

Zastosowanie rozwiązań techniczno-matematycznych jako modeli układów żywych znacząco poszerza naszą wiedzę o świecie. Modele biocybernetyczne nie tylko umożliwiają weryfikację wiedzy biologicznej, ale także wprowadzają nową jakość w rozumieniu biologicznych procesów. Pozwalają (między innymi) oceniać i porównywać efektywność rozwiązań technicznych z analogicznymi modelami procesów biologicznych. Biocybernetyka jako dziedzina przynależna do nauk przyrodniczych, biologicznych i technicznych stosuje metody jednej nauki do przedmiotu zawartego w drugiej grupie. Zakłada więc ona wspólność metod i modeli w dziedzinach wiedzy opisujących materię ożywioną i nieożywioną.

Pojęcie modelu biocybernetycznego odnosi się do opisu matematycznego procesu biologicznego wyizolowanego z całego organizmu. Charakterystyczną cechą tego rodzaju modelowania jest skupienie się na mechanizmach regulacji, kontroli i komunikacji. Mechanizm interferencji RNA stanowi w biologii teoretycznej modelowy przykład kontroli i regulacji poziomu transkryptów w cytoplazmie komórki biologicznej. W niniejszej pracy skupiono się w związku z tym na ilościowym opisie relacji między danymi eksperymentalnymi dotyczącymi poziomu ekspresji transkryptów i miRNAs w mechanizmie interferencji RNA.

Odkrywanie wiedzy na podstawie danych biologicznych wykorzystujących techniki eksploracji danych ulega poprawie dzięki zastosowaniu zasad modelowania biocybernetycznego [160][162]. Biocybernetyka wprowadza kontrolę i sterowanie w tworzonym modelu w taki sposób, aby doprowadzić do jak największej jego zgodności z modelowaną rzeczywistością biologiczną [161][176]. Dla rozważanego w tej pracy modelu dane wejściowe zostają w taki sposób przekształcone, aby uzyskać tą zgodność.

Centralny dogmat biologii, opisywany wyżej (patrz Rys. 2.1) paradoksalnie dotyczy przepływu informacji, która do tej pory była domeną nauk technicznych i matematycznych. Informacja zawarta w chromosomalnych sekwencjach, poprzez odpowiedni system kodowania, zostaje wykorzystana przy syntezie białka (syntezie na matrycy RNA sekwencji aminokwasowej – proces translacji). W technice, połączenie hardware'u z programem (software'em) [172] dokonał twórca pierwszej maszyny liczącej - komputera John von Neuman przy współpracy Johna W. Mauchly'ego oraz Johna Prespera Eckerta .

Odkrycie kodu genetycznego (zbiór reguł pozwalających dowolnej kombinacji nukleotydowej przyporządkować odpowiednią sekwencję aminokwasową, trójki nukleotydów stanowią tzw. kodon) było wielkim zaskoczeniem w środowisku genetyków i informatyków. Wprowadziło ono na grunt biologii idee abstrakcji programu i danych informatycznych. Zasady hierarchiczności (poziomy), niezależności (przenośności, uniwersalności) obowiązujące w inżynierii oprogramowania odkryto na poziomie organizacji fundamentów przyrody ożywionej. Informacyjna własność została uzyskana poprzez wykorzystanie bardzo podobnych struktur par nukleotydowych.

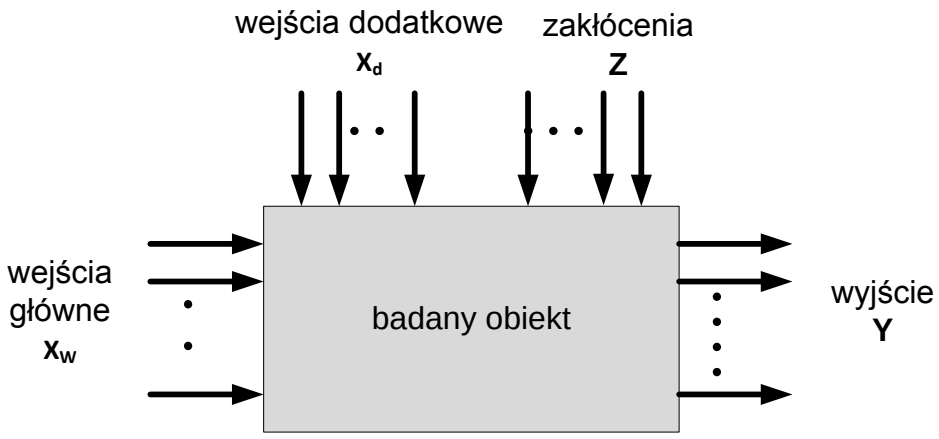
W ten sposób, poprzez centralny dogmat biologii, a dokładniej w biologii molekularnej utworzono furtkę, przez którą wprowadzono do tej dyscypliny wiedzy teorię informacji Shannona, teorię regulacji oraz teorię maszyn cyfrowych. Te trzy teorie razem połączone stanowią podstawę utworzonej w 1946 roku przez Norberta Wienera nowej interdyscyplinarnej dyscypliny naukowej – cybernetyki, ponieważ informacja występuje zawsze z regulacją.

Teoria informacji przedstawiona została w pracy *"A Mathematical Theory of Communication"* autorstwa Claude Shannon i Warren Weaver. Zajmuje się ilościowym opisem informacji, wprowadza pojęcia bitu informacji, entropii zdefiniowanych analogicznie do pojęć termodynamiki (traktujących o energii obiektów). Z teorii informacji wywodzą się metody telekomunikacyjne, teorie kodowania. Już na pierwszy rzut oka widoczna jest analogia dogmatu biologicznego do schematu transmisji sygnału cyfrowego z elementami nadawcy – kanału – odbiorcy. Rzecz jasna, że przypisanie funkcji nadawcy, kanału, czy odbiorcy jest sprawą czysto umowną w rzeczywistych układach biologicznych, które charakteryzuje duża złożoność wyrażona w hierarchiczności, wielopoziomowości i licznych pętlach sprzężenia zwrotnego modelowaniu choćby tylko na samym poziomie molekularnym. Obiektywnie i najprościej można przypisać funkcję nadawcy podmiotowi, który nazwano ewolucją, a która zakodowała informację w odpowiednich sekwencjach chromosomowych. Kanał transmisji, który charakteryzuje podatność na zakłócenia stanowi szlak przetwarzania tej zakodowanej informacji na postać zdekodowaną w formie docelowej sekwencji aminokwasowej. Zakłócenia stanowią „literówki", czyli mutacje wywołane czynnikami mutagennymi, chociaż mogą oczywiście wystąpić także dłuższe zakłócenia, na przykład insercje czy delecje wywołane poślizgiem w procesie duplikacji nici DNA, czy *crossing over* na etapie rozdzielania chromatyny przy podziale komórkowym.

Metody cybernetyczne polegają na [63]:

1. możliwie precyzyjnym opisie jakościowym i ilościowym mierzalnych wielkości występujących w procesie;
2. ustaleniu bądź założeniu w postaci hipotez związku między odpowiednimi wielkościami;
3. wykorzystaniu metod matematycznych do opisu badanego zjawiska;
4. wykorzystaniu teorii sterowania, informacji, komunikacji do analizy złożonych układów z pętlami sprzężeń zwrotnych.

Metoda "czarnej skrzynki" stosowana często w pierwszym etapie wtedy, kiedy nie jest znana wewnętrzna struktura badanego obiektu, polega na traktowaniu obiektu jako względnie odosobnionego [66]. Obiekt tego typu pozwala poprzez wyróżnione punkty brzegowe na wpływ sygnału zewnętrznych na procesy wewnętrzne. Oprócz nich wyróżnione są punkty, przez które następuje oddziaływanie obiektu na otoczenie. Są to wejścia i wyjścia układu (Rys. 4.1.). Z reguły tego typu wejścia uwzględniają zmienne mierzalne. Wejścia główne $\mathbf{X_w}$ mają zasadniczy wpływ na przebieg zjawiska. Wejścia pomocnicze $\mathbf{X_d}$ – sygnały, które należy uwzględnić w obliczeniach. Zakłócenia **Z –** są niemierzalne, ale można czasami ustalić ich parametry statystyczne. Zaletą stosowanie tej metody jest brak założeń, co do sposobu działania badanego mechanizmu.



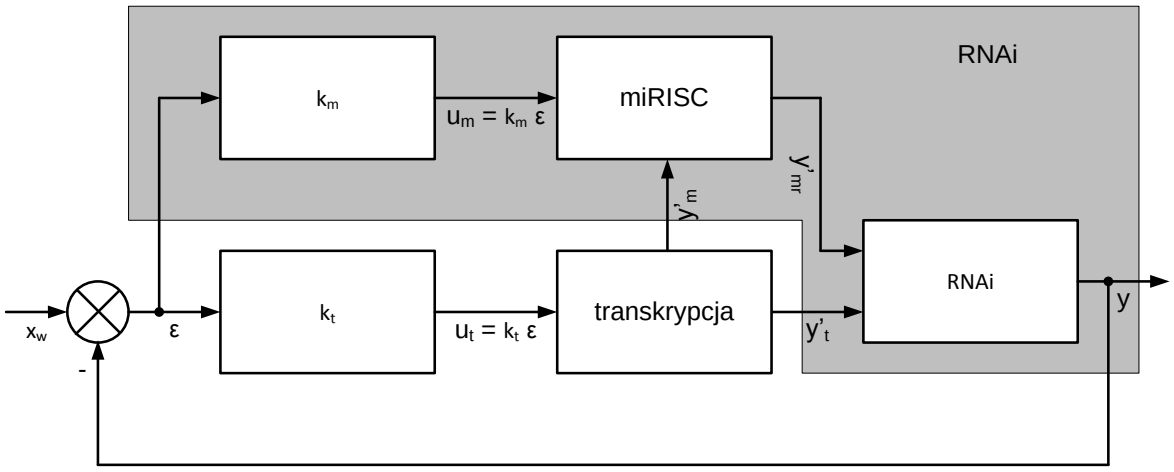
Rys. 4.1. Metoda "czarnej skrzynki".

Zasadniczy wkład ujęcia cybernetycznego polega na ustaleniu związków między sygnałami wejściowymi, a wyjściowymi. Ustalenie wpływu sygnałów dodatkowych czy też zakłóceń. Poszukiwane zależności można przedstawić w ogólnej postaci z wykorzystaniem nieznanej funkcji F:

$$\boldsymbol{Y} = \boldsymbol{F}(\boldsymbol{X}_w, \boldsymbol{X}_d, \boldsymbol{Z}, \boldsymbol{s}) \qquad \textbf{(4-1)}$$

gdzie ***s*** **–** zespół parametrów charakteryzujących stan obiektu. Celem badań jest zatem ustalenie postaci funkcji **F**. Pełne, cybernetyczne podejście do modelowania wymaga dalszych kroków takich jak ustalenie rodzaju sprzężenia zwrotnego, ocena korzyści zastosowania układu regulacyjnego (np. redukcja wpływu zakłóceń sygnałów analogowych, realizacja układu śledzącego wartość zadaną), badanie stabilności modelu.

Mechanizm interferencji RNA traktowany jako regulator poziomu transkryptów w cytoplazmie można przedstawić jako prosty model regulacyjny z pętlą ujemnego sprzężenia zwrotnego (Rys. 4.2). Przyjmując, że mamy informację o wartości pożądanej ekspresji N transkryptów oznaczoną wektorem X_w, , poziom ekspresji N transkryptów w cytoplazmie uzyskujemy w wyniku regulacji podstawowej sygnałem transkrypcji oraz dodatkowo przez kompleksy miRISC w mechanizmie RNAi. Schemat nie rozstrzyga, jakie czynniki powodują wybór dodatkowego mechanizmu regulacji w postaci RNAi dla wybranych genów. Stąd wektor sygnałów błędu ε jest podany na oba człony dopasowujące k_t i k_m.



Rys. 4.2 Model regulacji genów z wyróżnionym mechanizmem interferencji RNA (część szara)

Znaczenie symboli z Rys. 4.2:

X_w – wektor wartości zadanych poziomów ekspresji transkryptów;

ε – wektor sygnałów będących różnicą wektorów wartości zadanej X_wi wartości poziomów ekspresji uzyskanej y;

k_m,k_t – wektor współczynników proporcjonalności modyfikacji sygnału błędu ε na sygnał wykonawczy odpowiednio U_t i U_m;

U_t, U_m – wektor sygnałów transkrypcji i procesu wytwarzania miRISC;

y'_t,y'_m,y'_{mr},,y – wektor wartości poziomów ekspresji odpowiednio: transkryptów, pri-miRNA, miRISC, transkryptów po regulacji.

Sygnały przedstawione na schemacie są wartościami zmiennymi, zależnymi od wielu czynników "topograficznych", ontologicznych, środowiskowych.

W przedstawionym modelu możemy wyróżnić dwa wektory sygnałów dotyczące poziomu ekspresji, które to wartości w postaci zmiennej losowej są dostępne, jako rezultat eksperymentu mikromacierzowego przeprowadzonego w ustalonych warunkach. Są to wartości poziomu ekspresji transkryptów i miRNAs w cytoplazmie, już po regulacji oznaczone na schemacie przez y. Jak łatwo wywnioskować wartości te są mniejsze w stosunku do pierwotnych poziomów transkryptów oznaczonych na schemacie y'_t,y'_m,y'_{mr}. W niniejszej pracy postanowiono skupić się na matematycznym opisie relacji ilościowych między poziomami ekspresji transkryptów i cząsteczek miRNAs.

4.1 Założenia modelu rozważanego w tej pracy

Ze względu na opracowania różnych rozwiązań bioinformatycznych problem wyszukiwania targetów, czyli transkryptów, do których odpowiednim regionie przyłącza się konkretna cząsteczka miRNA, można rozwiązywać poprzez poprawę i udoskonalenia obecnych narzędzi bioinformatycznych. Ze względu na przedstawioną wcześniej problematykę, poznania procesu interferencji RNA w pracy skupiono się na rozwiązaniach, które uwzględniają najszerszy z możliwych zasobów informacji dotyczących tego procesu. Zaawansowany matematycznie model VB-GMM zaimplementowany w pakiecie TargetScore umożliwia stosunkowo proste go

rozwinięcie i umożliwia uwzględnienie dodatkowej informacji, która powinna poprawiać jakość analizy przy określaniu targetów. Ta dodatkowa informacja w stosunku do zestawu danych i parametrów wejściowych modelu TargetScore [106] polega na uwzględnieniu także rzeczywistego profilu ekspresji miRNAs w analizowanych próbach.

Założenia modelu:

1. model służy do wyszukiwania targetów dla cząstek miRNAs na podstawie danych mikromacierzowych,
2. zakłada się uwzględnienie w danych wejściowych profilu ekspresji transkryptów oraz miRNAs,
3. uwzględnia się wykorzystanie informacji kontekstowej i filogenetycznej w przeprowadzonej analizie,
4. zakłada się, że zmienność poziomu transkryptów między porównywanymi próbami biologicznymi wywołana jest przede wszystkim potranskrypcyjną regulacją genów. Zakładamy zatem, że rejestrowane zmiany poziomu ekspresji transkryptów są odwrotnie skorelowane ze zmianami poziomu cząsteczek miRNA.

Wykorzystanie informacji o ekspresji pozwala na wyznaczenie funkcjonalnych targetów miRNAs.

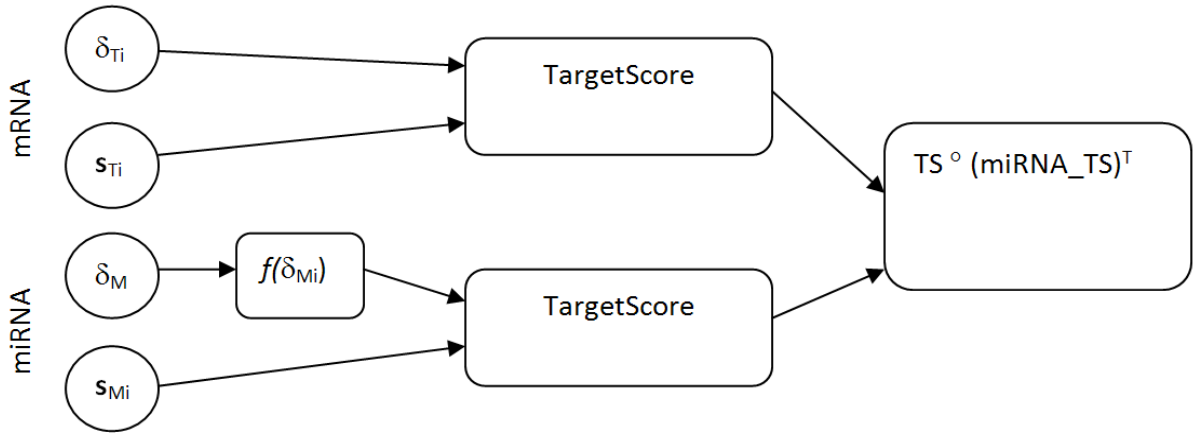
4.2 Model biTargetScore - rozwinięcie modelu TargetScore

Wykorzystanie przedstawionej wcześniej biblioteki TargetScore (patrz rozdział 3.4.5) wydaje się być możliwe w zastosowaniu do innych sytuacji eksperymentalnych, które jako rezultat podają zmianę poziomu ekspresji genów w odpowiedzi na niepatologiczne (normalne) czynniki, jakim poddana zostaje hodowla komórkowa. Mamy wówczas do czynienia ze zmianą zamiast jednego miRNA z eksperymentu transfekcji ze zmianą w całym profilu miRNA. Pojawia się pytanie, w jaki sposób wykorzystać TargetScore dla takiego przypadku? Rejestrowanie informacji o sparowanych dwóch zbiorach: poziomu ekspresji miRNA oraz mRNA wydaje się też być podejściem bardziej obiektywnym w stosunku do założeń niezmienności endogennego profilu miRNA w transfekcji.

Autor niniejszej pracy proponuje uogólnienie modelu TargetScore na przypadek uwzględnienia wielu miRNA w miejsce jednego oraz uwzględnienia w obliczeniach także poziomu zmienności zbioru miRNA. Udoskonalony model nazwano roboczo biTargetScore. W aspekcie biologicznym TargetScore odpowiada na pytanie, jakie jest prawdopodobieństwo interakcji każdego z transkryptów z danym miRNA. W biTargetScore pytanie zadawane jest w obydwóch kierunkach: tym samym, co powyżej oraz drugim - jakie jest prawdopodobieństwo interakcji każdego miRNA ze zbioru z danym targetem. Analiza przebiega dwutorowo. Najpierw przeliczany jest pierwszy, podstawowy zbiór danych wejściowych, czyli: logFC, Sc, P_{CT}. W drugim kroku w miejsce transkryptów podawany jest logFC ale dotyczący zbioru miRNA i odpowiednie Sc i P_{CT}. Schemat proponowanego rozwiązania przedstawiono na Rys. 4.3.

Proponowany model zakłada odmienne dopasowanie wartości logFC miRNA. Stąd schemat uwzględnia człon funkcji dopasowującej. Dla celów testowych modelu założono, że spadkowi poziomu ekspresji transkryptów będących celem miRNA towarzyszy wzrost poziomu ekspresji miRNA. Czyli najprostszy przypadek – funkcja inwersji. Funkcja dopasowania może zostać w toku walidacji modelu zmieniona. Najczęściej zakłada się ujemną korelację poziomu ekspresji miRNA i mRNA np. obliczaną metodą korelacji Pearsona. Wydaje się, że na doświadczalny sposób wyznaczenia tej funkcji pozwalają dobre rezultaty deklarowane przez autora TargetScore.

Każda komórka organizmu występuje w pewnym warunkowo zależnym stanie równowagi. Część transkryptów podlega translacji, a jakaś część podlega degradacji na wielu różnych etapach procesu produkcji białka. Oznacza to, że wyznaczony w eksperymencie poziom ekspresji transkryptów jest mniejszy niż rzeczywista wartość ekspresji. Analogicznie transkrypty niekodujące, które np. stanowią protoplastę pri-miRNA również podlegają procesom regulacji. Zmienność ekspresji genów podlega wreszcie modulacji czynnikami transkrypcyjnymi. W takiej sytuacji trudno jest jednoznacznie udzielić odpowiedzi, jaka jest relacja między ekspresją miRNA a transkryptami kodującymi i regulowanymi przez miRNA, szczególnie na poziomie ogólnym?



Rys. 4.3. Model biTargetScore. δ_{Ti} – wektor logFC transkryptów od długości N, S_{Ti} – macierz Sc i P_{CT} o wymiarach N x M, δ_{Mi} – wektor logFC miRNAs, S_{Mi} – macierz SC i P_{CT} o wymiarach M x N. $f(\delta_{Mi})$ – funkcja dopasowująca, TargetScore – model podstawowy, TS – macierz prawdopodobieństw dla transkryptów N x M, miRNA_TS – macierz prawdopodobieństw dla miRNA M x N.

W wyniku dwutorowych obliczeń w modelu biTargetScore uzyskujemy dwie macierze: TS jako wynik iteracji (dla każdego miRNA obliczamy prawdopodobieństwo interakcji z każdym mRNA ze zbioru) oraz macierz miRNA_TS, jako wyniki iteracji każdego mRNA z każdym miRNA ze zbioru. Kolumny tych macierzy posiadają prawdopodobieństwa zintegrowane tzn. są wynikiem jednokrotnego zastosowanie modelu VB-GMM. Iloczyn Hadamarda (iloczyn komórek o tych samych współrzędnych) macierzy TS oraz transpozycji macierzy miRNA_TS daje ostateczny rezultat.

Nowatorski charakter modelu biTargetScore zaproponowanego przez autora niniejszej rozprawy polega na wykorzystaniu biblioteki TargetScore do integracji informacji o zmienności ekspresji mRNAs jak i miRNAs w celu wyznaczenia targetów.

4.3 Ograniczenia modelu

Przedstawiony model biTargetScore skupia się na ustaleniu relacji między poziomami różnicowej ekspresji transkryptów i miRNAs z uwzględnieniem parametrów pomocniczych. Zakłada on odwrotną korelację wartości tych poziomów będącą następstwem przyjętego modelu regulacji RNAi. Model matematyczny VB-GMM w celu przeprowadzenia obliczeń zakłada wzajemną niezależność poziomów transkryptów i miRNAs. To podstawowe założenie staje się zarazem ograniczeniem jego stosowania, ponieważ nie uwzględnia ona następujących stanów/sytuacji:

1. Brak uwzględnienia rywalizacji wielu miRNA do tego samego targetu. Przyłączenie jednego kompleksu RISC jednym miRNA może powodować brak dostępu dla innych cząsteczek miRNAs.
2. Ograniczenia populacji kompleksów RISC np. będące efektem regulacji RNAi genów białek kompleksu.
3. Pomiar ekspresji dokonywany jest zatem dwupunktowo przed wprowadzeniem czynnika zmieniającego ekspresję oraz po jego wprowadzeniu. Brak zatem badania dynamiki zmian ekspresji. Czynnik wywołujący zmianę może mieć różny charakter: podany lek, symulacja infekcji, zmiana warunków hodowli, itd. Może też nim być wprowadzenie w procesie transdukcji cząsteczki miRNA. Wstępna analiza jakościowa wpływu tych czynników na procesy wewnątrzkomórkowe wydaje się być bardzo istotna, szczególnie gdy w jej toku stwierdzimy zmianę czynników transkrypcyjnych.
4. Wzajemny wpływ miRNAs na siebie, wpływ miRNA poprzez interferencje RNA na geny innych lub własne.
5. Efekt wysycenia (tłumaczenie własne, oryginalne *dilution-effects*) polegający na tym, że miRNA posiadający wiele targetów z mniejszą skutecznością obniża poziom ekspresji docelowych transkryptów w porównaniu do tych miRNA, które mają niewiele targetów [3].
6. Wyszukiwanie targetu tylko w obrębie regionu 3'UTR. Rozdział 2 cytował przypadki występowania miejsc wiązań także w obrębie 5'UTR lub CDS.

4.4 Przewidywane rezultaty

Uwzględnienie informacji o ekspresji miRNAs oraz mRNA w predykcji targetów jest metodą identyfikacji funkcyjnych cząsteczek miRNA. Można założyć, że metoda ta w sposób automatyczny poprawi jakość predykcji targetów. Czynniki wywołujące zmianę ekspresji stanowią jednak na tyle złożoną sieć relacji, że każdą predykcję zawsze należy weryfikować eksperymentalnie.

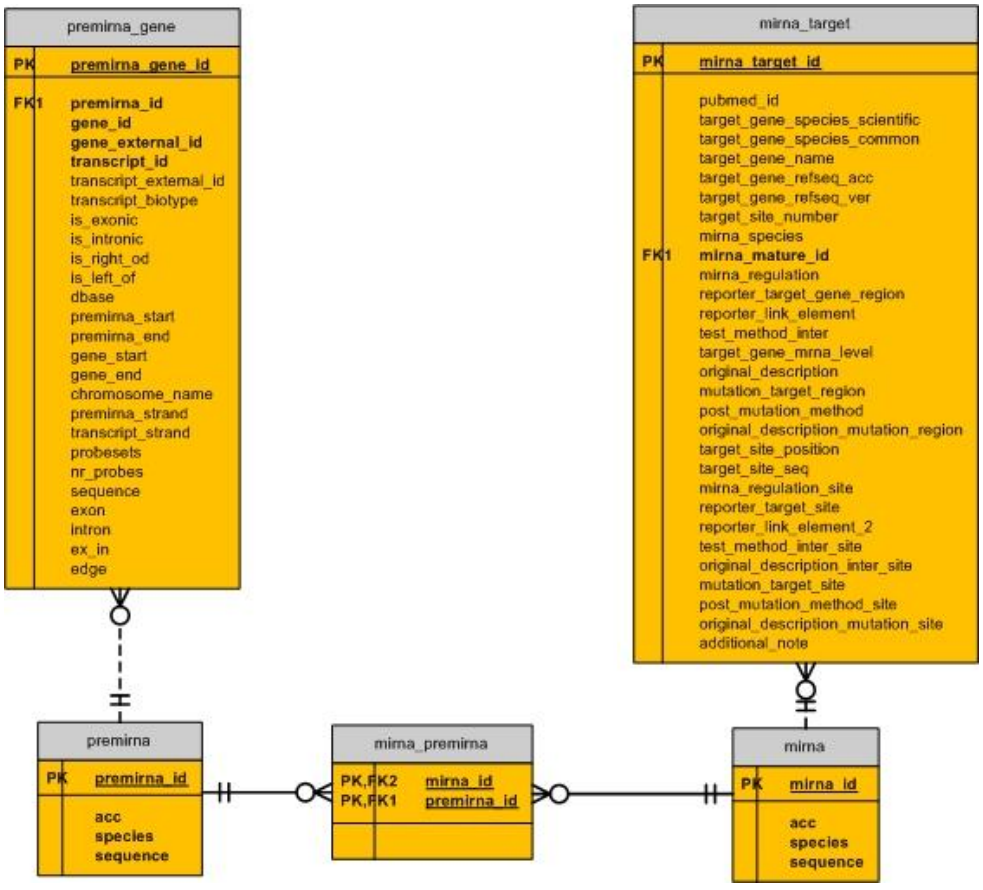
5 Opis implementacji

5.1 Lokalna baza danych bioinformatycznych w oparciu o model BioSQL

Model obiektowy BioSQL (http://www.biosql.org/) zawiera uniwersalny model relacyjny sekwencji, cech sekwencji, notacji, taksonomii i ontologii, PhyloDB (drzewa filogenetyczne), oraz interfejs obiektowo-relacyjny związany z wykorzystywanym językiem programowania. Pierwotnie powstał na użytek lokalnych zasobów GenBank oraz jako wspólna platforma dla Bio* projektów: BioPerl, BioPython, BioJava, BioRuby, które wykorzystują mapowanie relacyjno-obiektowe.

W modelu zastosowano wzorzec słabej typizacji, czyli systemu typów, w którym typ wyrażenia może być automatycznie zmieniony, jeśli kontekst tego wymaga. Słaba typizacja nie wykrywa pewnych rodzajów błędów, natomiast podkreśla się, że jest wygodniejsza w praktyce, w szczególności dla danych biologicznych, nie zawsze jednoznacznie zdefiniowanych. Obiekt jest określony poprzez grupę predefiniowanych encji, utworzonych dla określania różnych typów obiektów. W ramach encji specjalistyczne atrybuty określają typ obiektu. Zbiór typów pochodzi ze słownika ontologicznego, który przypisuje znaczenie obiektowi (wierszu oraz atrybutom).

Lokalna implementacja BioSQL zawiera rozpakowane zasoby informacji o transkryptach ludzkich (ftp://ncbi/refseq/H_sapiens/mRNA_Prot/human.rna.gbff) w formacie GenBank. Struktura relacyjna BioSQL została na potrzeby pracy rozbudowana o tabele dedykowane do przechowywania pełnej informacji o miRNAs. Pełny schemat struktury BioSQL zamieszczono w Dodatku E.



Rys. 5.1. Fragment lokalnej struktury relacyjnej bazy evolBioSQL dotyczący miRNAs.

Struktura BioSQL została poszerzona o dodatkowe tabele (Rys. 5.1). Informacje zapisane w tych tabelach uzyskano na podstawie różnych zasobów bioinformatycznych. Podstawową tabelę "mirna" wypełniają dane z referencyjnej bazy miRBase. Tabela "mirna_target" pochodzi z zasobów bazy miRecords. Pozostałe tabele: "premirna_gene", "premirna", "mirna_premirna" zostały uzupełnione na podstawie predykcji pri-miRNA zrealizowanej w Biocentrum Patofizjologi (Division of Molecular Pathophysiology Biocenter, Innsbruck Medical University) [137].

Struktura powiązania tabel (Rys. 5.1) odpowiada formalnemu wyobrażeniu przepływu informacji w biologii: od genu do funkcyjnej cząsteczki miRNA. Zawarte na schemacie relacje między tabelami uwzględniają znane relacje ilościowe miedzy poszczególnymi etapami wytwarzania dojrzałego miRNA. Tabela "mirna_premirna" uwzględnia relacje typu n:m, czyli ten sam miRNA może pochodzić z różnych prekursorów i odwrotnie ten sam prekursor może być macierzysty względem różnych miRNAs.

Lokalna implementacja bazy BioSQL została wykorzystana do uzyskanie odpowiednich sekwencji transkryptów, informacji o położeniu regionów CDS i 3'UTR i innych celów.

5.2 Analiza zbioru sekwencji miRNA

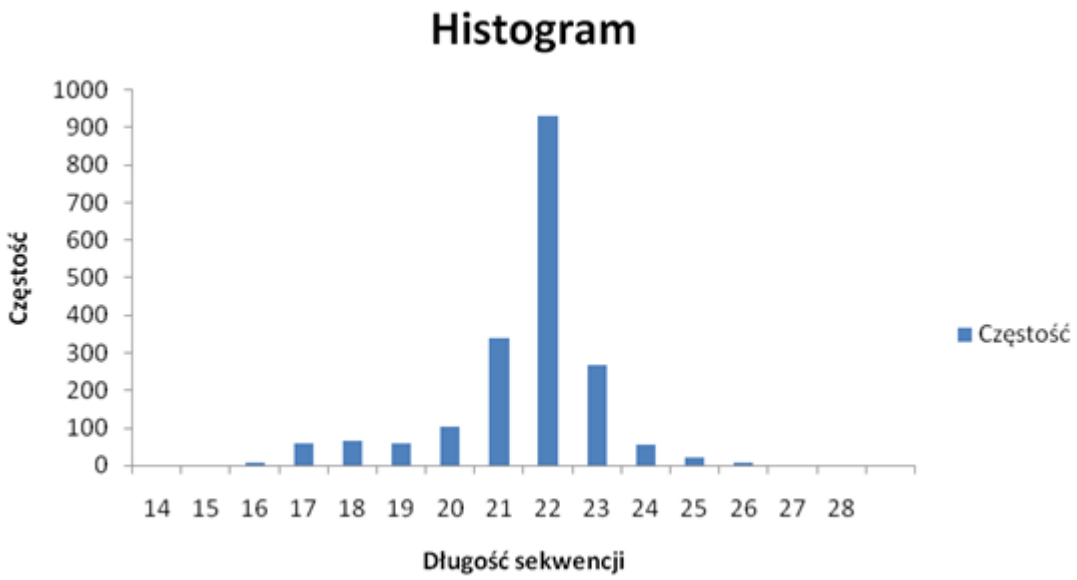
Charakterystyka aktualnego zbioru cząsteczek miRNAs pozwala na jego ocenę i poznanie właściwości. W ramach analizy zbioru przeprowadzono analizę długości sekwencji miRNAs, analizę częstości występowania homologów miRNAs w transkryptach, duplikację homologów w obrębie tego samego transkryptu oraz oceniono stopień losowości sekwencji wyznaczając entropie blokową.

Szacuje się, że ok 60% wszystkich ludzkich genów podlega regulacji miRNAs [122]. Średnio podaje się, że jedna rodzina miRNA posiada kilkaset targetów [59]. Liczba rodzin zdefiniowana na podstawie miRBase wynosi 1543. Przypisanie miRNA do rodziny odbywa się na podstawie dojrzałego miRNA, sekwencji oraz/lub struktury pre-miRNA [92]. Geny miRNA w rodzinie mogą prezentować pełną konserwatywność dojrzałego miRNA, częściową – sekwencję *seed*. Warto dodać, że geny z tej samej rodziny miRNAs nie są losowo zlokalizowane, ale celowo zajmują pozycję wokół genów zaangażowanych w infekcję, system immunologiczny, system sensoryczny i neurodegeneracyjne choroby, rozwój oraz nowotworzenie [120]. Przeprowadzono badania nad wkomponowaniem mechanizmu RNAi w sieć regulacyjną genów (*gene regulatorynetwork*) [36]. Sieć ta kontroluje, które geny są aktywowane w odpowiedzi na sygnał biologiczny.

W wyniku przeprowadzonej przez autora analizy długości sekwencji miRNAs uzyskano następujące parametry rozkładu:

ilość sekwencji:	1921
wartość maksymalna długości:	27 nt
wartość minimalna długości:	15 nt
średnia długość:	21,56 nt

oraz histogram (Rys. 5.2.):



Rys. 5.2. Rozkład długości sekwencji miRNAs

5.2.1 Badanie częstości homologów miRNAs w transkryptach

Z punktu widzenia biochemicznego niestabilność wiązania się cząsteczki miRNA z docelowym transkryptem z staje się pretekstem do porównania komplementarności sekwencji miRNAs i zbioru transkryptów. Relatywnie niski stopień komplementarności dupleksów miRNA/mRNA, potwierdzony eksperymentalnie, podsuwa pytanie: czy w transkryptomie znajdziemy lokalne dopasowania sekwencji miRNA lepiej do nich pasujące? W celu odpowiedzi na to pytanie przeprowadzono eksperyment obliczeniowy wyszukiwania dopasowań sekwencji miRNAs i sekwencji transkryptów.

W pierwszym kroku oceniono komplementarność całej sekwencji miRNA i zbioru transkryptów. Interesująca staje się ocena „popularności" - częstości komplementarnej sekwencji miRNA w zbiorze transkryptów. Analiza dotyczy, zatem komplementarnych homologów sekwencji miRNA. W celu jej przeprowadzenia wykorzystano gotową implementację algorytmu Needlemana-Wunscha. Algorytm ten stosowany jest przy globalnych uliniowieniach z tzw. liniową karą za przerwy. Przyjęto trzy stopnie jakości dopasowania: ACCEPTABLE , GOOD, PERFECT. Sposób obliczenia progowych wartości punktacji:

PERFECT:	(= length * 5)
GOOD:	(>= (length - 1) * 5 - 4) LUB (= (length - 3) * 5 - 4)
ACCEPTABLE:	>= (19 * 5)

gdzie, *length* – długość sekwencji *query*, tut. sekwencji miRNA.

Wyjaśnienie użytych progów:

- PERFECT - odnalezienie w sekwencji transkryptu dopasowania tożsamego,

- GOOD – odnaleziona sekwencja dopasowania może różnić się jednym albo trzema nukleotydami w jednym ciągu (w pewnym stopniu gwarantuje to jedna kara za przerwę „-4”),
- ACCEPTABLE - arbitralnie ustalona minimalna długość dopasowania (przyjęto 19nt).

Do analizy wykorzystano program *glsearch* z pakietu FASTA uruchamiany z ustalonymi parametrami:

glsearch36 -n -Q -E "5000 1.0" -m 10 *query library*

query to kolejne sekwencje miRNA (ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gz)

library – sekwencje mRNA z pliku human.rna.fna (baza NCBI).

Deklaracja parametru estymacji /–E "5000 1.0"/ powoduje, że w pliku wynikowym pojawiają się także dopasowanie suboptymalne typu HSP (*high-scoring alignments*)

http://fasta.bioch.virginia.edu/fasta_www2/fasta_guide.pdf.

Uzyskane rezultaty analizy częstości homologów miRNAs w transkryptach przedstawia Tabela 8.

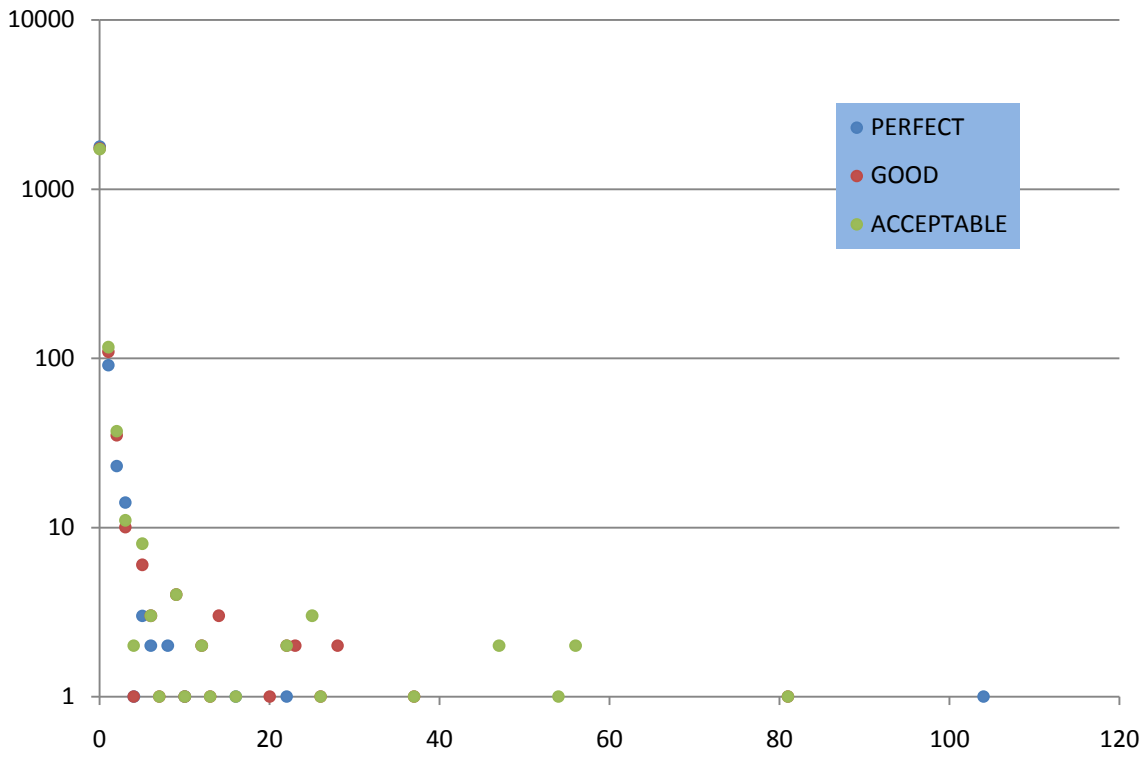
Tabela 8. Rezultaty badania częstości homologów.

Legenda:

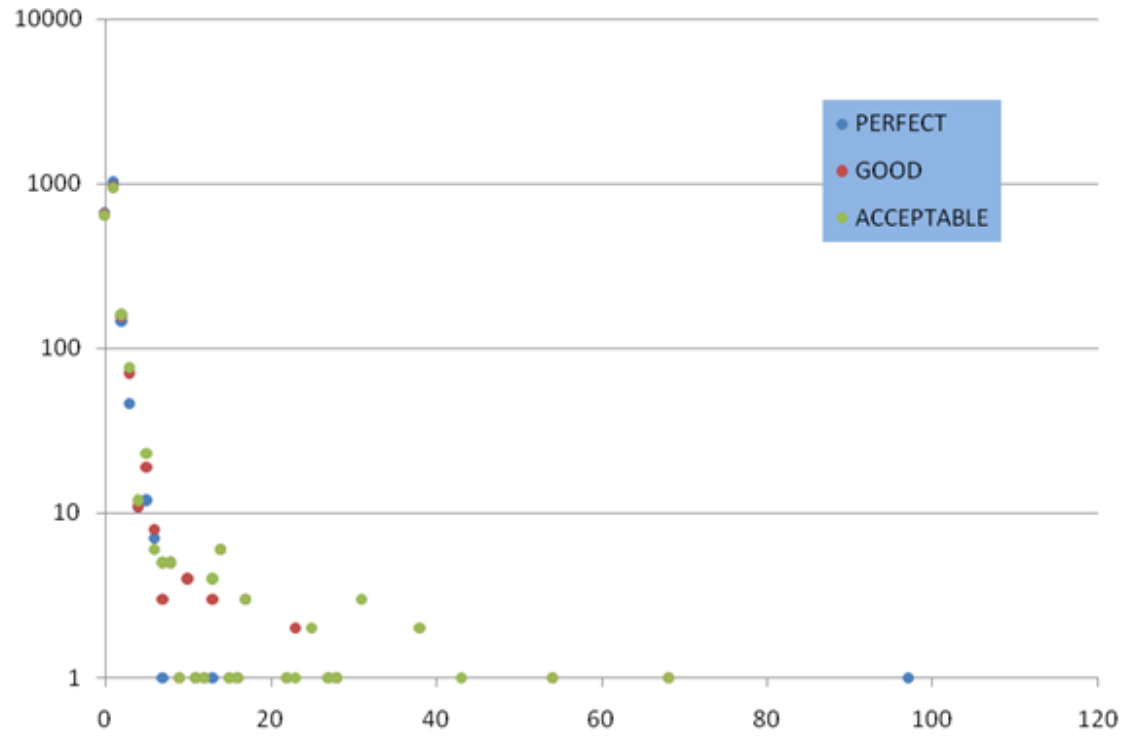
Wszystkie trafienia – suma wszystkich znalezionych sekwencji podobnych poza HSP
Wszystkie HSP – suma wszystkich alternatywnych dopasowań (HSP) względem podstawowych trafień
Wszystkie trafienia + HSP - suma wszystkich trafień łącznie z alternatywnymi dopasowaniami
ACCEPTABLE , GOOD, PERFECT – trzy progi podobieństw omówione wcześniej
Wszystkie zaakceptowane - suma kontrolna dla progów dopasowań
Wszystkie zaakceptowane HSP - suma kontrolna dla progów dopasowań uwzględniająca tylko HSP
„Osierocone” HSP – te HSP, których główne hity zostały zaakceptowane, a które były za słabe, aby same zostać zaakceptowane.

	ALL	FORWARD	REVERSE
Wszystkie trafienia	14582675		
Wszystkie HSP	11560499		
Wszystkie trafienia + HSP	26143174		
ACCEPTABLE	2176	1088	1088
GOOD	3710	1894	1816
PERFECT	2138	1760	378
Wszystkie zaakceptowane	8024	4742	3282
Wszystkie zaakceptowane HSP	303	140	163
„Osierocone” HSP	3821	1956	1865

Dwa kolejne wykresy (Rys. 5.3, Rys. 5.4) przedstawiają zależność liczby cząsteczek miRNAs w funkcji liczby znalezionych lokalnych dopasowań na nici transkryptów z uwzględnieniem trzech progów podobieństwa. Tabelaryczne zestawienie wyników zawarte jest w Tabela 9.



Rys. 5.3. Zależność liczby miRNAs od liczby odpowiadających homologów w transkryptach dla nici komplementarnej (REVERSE).



Rys. 5.4. Zależność liczby miRNAs od liczby odpowiadających homologów w transkryptach dla nici dominującej (FORWARD).

Tabela 9 Tabelaryczne zestawienie liczby lokalnych kopi (homologów) znalezionych na niciach komplementarnych transkryptów.

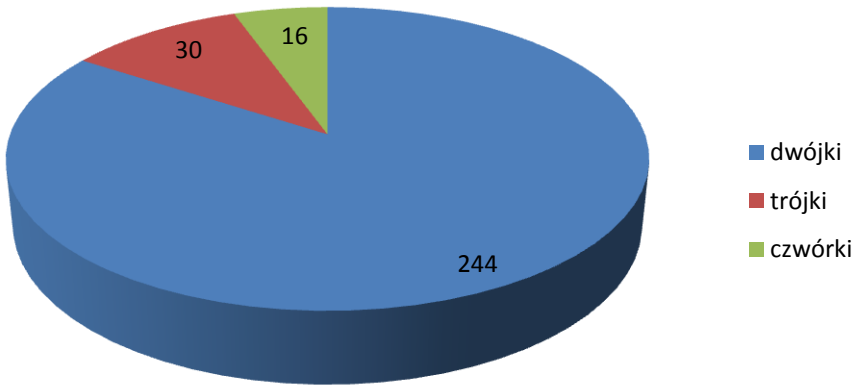
l. kopii	l. miRNA	l. dopasowań
0	1716	0
1	116	116
2	37	74
3	11	33
4	2	8
5	8	40
6	3	18
7	1	7
9	4	36
10	1	10
12	2	24
13	1	13
16	1	16
22	2	44
25	3	75
26	1	26
37	1	37
47	2	94
54	1	54
56	2	112
81	1	81
215	1	215
282	1	282
403	1	403
591	1	591
873	1	873
	1921	3282

5.2.2 Duplikacje homologów w obrębie transkryptów

Znalezione dopasowania sekwencji miRNAs i sekwencji zbiorów transkryptów wykazują także duplikację w obrębie tego samego transkryptu. W Tabela 10 zebrano liczby miRNAs, które wykazują największą liczbę duplikatów w obrębie transkryptów. Na pytanie jak wiele duplikatów znajduje się w obrębie pewnych transkryptów odpowiada Rys. 5.5.

Tabela 10. Zestawienie duplikatów dla nici REVERSE

	Nazwa miRNA	liczba transkryptów z duplikatami
1	hsa-miR-5585-3p MIMAT0022286	2
2	hsa-miR-1273g-3p MIMAT0022742	4
3	hsa-miR-5096 MIMAT0020603	37
4	hsa-miR-1273d MIMAT0015090	5
5	hsa-miR-574-5p MIMAT0004795	88



Rys. 5.5. Porównanie liczby duplikatów "dwójek", "trójek", "czwórek" w obrębie tego samego transkryptu dla REVERSE

5.2.3 Entropia blokowa

Teoria informacji znalazła także zastosowanie w ocenie informacji genetycznej. Ilościowe pojęcie informacji np. entropia Shannona *H* [153] i dywergencja Kullbacka-Leiblera zwana też entropia relatywną D_{KL} [95] są definiowane następująco:

$$H = -\sum_i p_i \log(p_i) \quad \textbf{(5-1)}$$

$$D_{KL}(P||Q) = \sum_i p_i \log(p_i/q_i) \quad \textbf{(5-2)}$$

Gdzie P i Q są gęstościami prawdopodobieństwa, $P = \{p_i\}$ i $Q = \{q_i\}$. Entropia *H* jest miarą niepewności pewnego zdarzenia o prawdopodobieństwie p_i, a relatywna entropia D_{KL} jest miarą podobieństwa pomiędzy dwoma rozkładami zmiennej (nie jest ona prawdziwą metryką np. $D_{KL}(P||Q)$ nie jest równe $D_{KL}(Q||P)$ z wyjątkiem przypadku, gdy $P = Q$).

Entropia blokowa [148] stanowi pewną modyfikację entropii Shannona. Jest ona obliczana nie dla każdego nukleotydu osobno, ale dla wydzielonych bloków nukleotydowych, czyli subsekwencji. Rozważmy jednokierunkową, nieskończoną sekwencję $s_1, s_2, \ldots$, gdzie $s_t \in \{0,1,\ldots,d-1\}$. Dla

sekwencji nukleotydowych możemy przyjąć $d = 4$. Załóżmy stochastyczny proces s_1, s_2, ... z prawdopodobieństwami:

$$p_t(s_1, \dots., s_n) = prob\{s_{t+1}, \dots, s_{t+n} = s_n\} \quad \textbf{(5-3)}$$

Zakładając, że rozkłady są stacjonarne możemy usunąć indeks t w $p_t(s_1, \dots., s_n)$ i zdefiniować entropię blokową jako:

$$H_n = -\sum_{s_1,\dots,s_n} p(s_i, \dots, s_n) \log p(s_i, \dots, s_n) \quad \textbf{(5-4)}$$

która mierzy średnią ilość informacji zawartą w słowie o długości n.

Przedstawiono obliczenie tzw. blokowej entropii Shannona przeprowadzone dla całego znanego zbioru ludzkich dojrzałych (*mature*) miRNAs. Obliczenia przeprowadzono wg różnych schematów:

1. dla całego zbioru miRNA, wydzielonych podgrup ze względu na pochodzenie: intronicznych i egzonicznych miRNA, ze względu na ich stopień konserwatywności;
2. obliczenia dla okna: przesuwającego się po 1nt lub aktualnej wielkości bloku;
3. dla subsekwencji miRNA: pierwszych 8nt, środkowych 8nt.

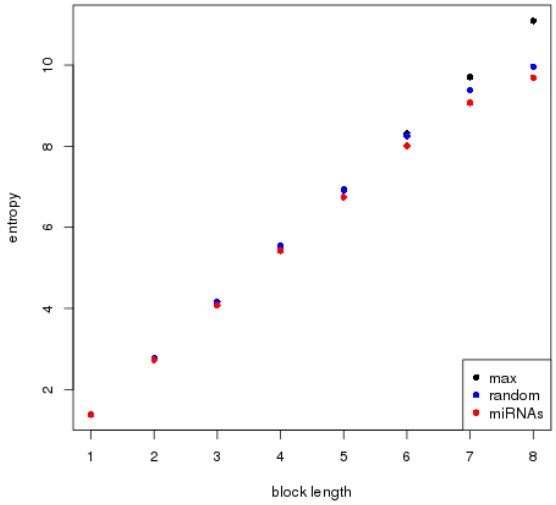
Zbiór miRNA stanowi 1921 sekwencji o średniej długości 21nt. Całkowita długość wszystkich sekwencji to ok 40 000nt. Wielkość wydzielanych bloków: 1-8nt. Wyniki obliczeń przestawiono na wykresach łącznie z obliczoną maksymalną entropią:

$$H_{max} = -ln\frac{1}{4^x} \quad \textbf{(5-5)}$$

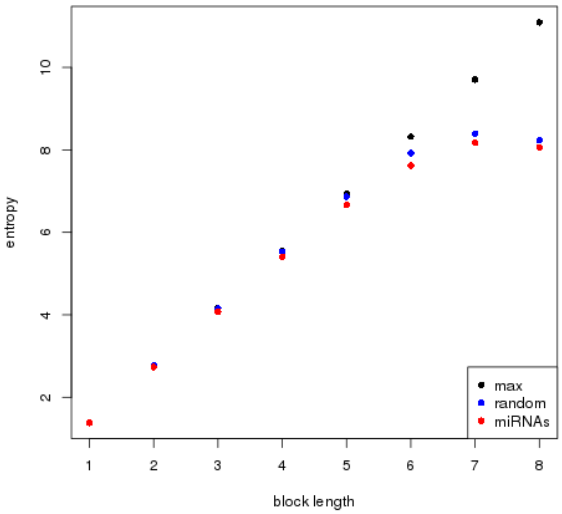
gdzie:

x – długość bloku,

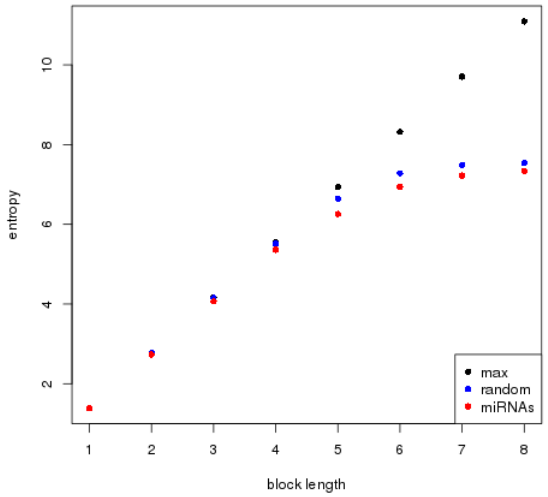
oraz entropią wyliczoną dla sekwencji losowej o równomiernym rozkładzie oraz identycznej długości z sekwencją obliczaną. Wyniki przedstawiono na wykresach na zbiorczym rysunku (Rys. 5.6, Rys. 5.7, Rys. 5.8, Rys. 5.9, Rys. 5.10, Rys. 5.11).



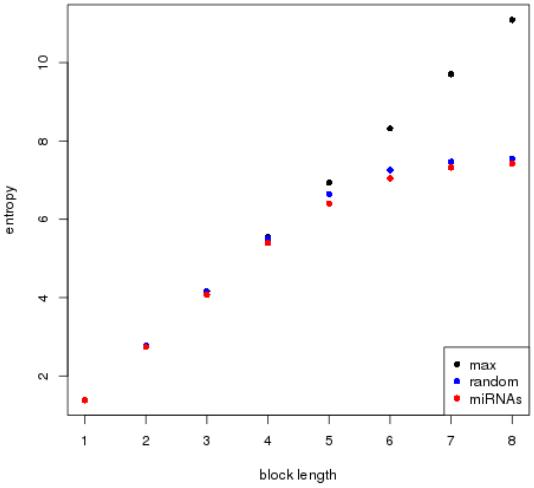
Rys. 5.6. Blokowa entropia zbioru miRNAs dla okna co 1nt.



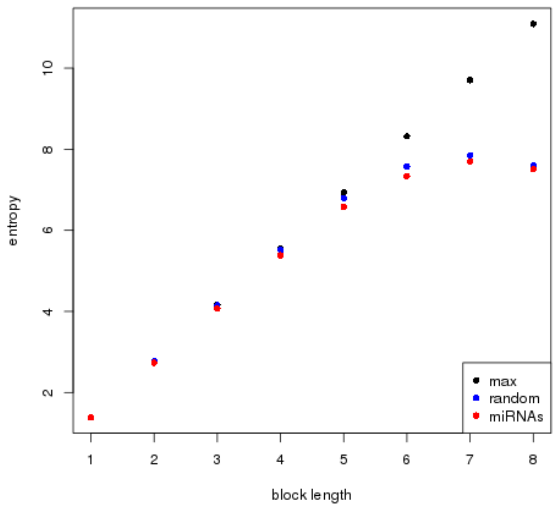
Rys. 5.7. Blokowa entropia zbioru miRNAs dla okna co długość bloku



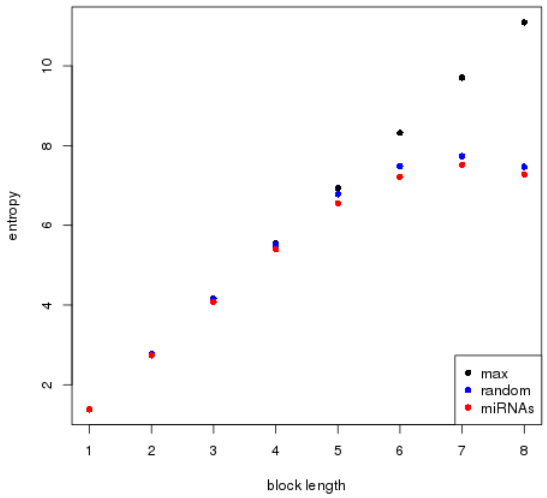
Rys. 5.8. Blokowa entropia zbioru miRNAs (pierwsze 8nt od strony 5') dla okna o długości bloku



Rys. 5.9. Blokowa entropia zbioru miRNAs (dla subsekwencji 8-15nt od strony 5') dla okna o długości bloku.



Rys. 5.10. Blokowa entropia zbioru intronicznych miRNAs dla okna o długości bloku.



Rys. 5.11. Blokowa entropia zbioru egzonicznych miRNAs dla okna o długości bloku.

5.3 Implementacja modelu biTargetScore

Implementacja modelu biTargetScore opiera się na skryptach napisach w języku R (www.r-project.org). Język R jako język interpretowany, wykorzystywany jest do obliczeń matematycznych i statystycznych(także tych czasochłonnych) ponieważ korzysta z bibliotek napisanych m.in. w języku C. Posiada on olbrzymie zasoby bibliotek implementujących techniki obliczeniowe oraz graficzne. Głównym powodem wyboru tego języka programowania w niniejszej pracy był jego związek z projektem **Bioconductor**, który został opary na platformie programistycznej R. **Bioconductor** posiada rozwinięte zasoby operowania na danych biologicznych, w tym przede wszystkim danych genomicznych, których dostępność przeważyła nad innymi językami z rodziny "bio".

Do poprawnego działania skryptu biTargetScore.R potrzebne są zainstalowane następujące składniki – pakiety: "Bioconductor", "TargetScore", "TargetScoreData", "gdata", "plotrix".

5.3.1 Wstępne przetworzenie danych wejściowych

Do przeprowadzenia analizy targetów potrzebne są dwa pliki wejściowe w formacie tekstowym zawierające tabelaryczne dane dotyczące zmiany transkrypcji (*fold change*) w porównywalnych próbach z eksperymentu biologicznego. Jeden plik powinien dotyczyć zmienności ekspresji transkryptów wraz informacją o relacjach między oznaczeniem - identyfikatorem sond a identyfikatorami transkryptów oraz genami. Fragment poprawnego pliku wejściowego Rys. 5.12.. Drugi plik powinien zawierać zmiany ekspresji miRNAs.

```
"1007_s_at#U48705#DDR1" 0.707012017257396 -0.0295997834523467
"1053_at#M87338#RFC2" 0.0744953682241937 0.540593812829006
"117_at#X51757#HSPA6" 0.994946322101304 0.00166070661223028
"121_at#X69699#PAX8" 0.300583207372982 -0.231801710812097
"1255_g_at#L36861#GUCA1A" 0.974644212660714 -0.00403802991786328
"1294_at#L13852#UBA7" 0.11568930997613 -0.551283301465054
```

Rys. 5.12. Fragment pliku wejściowego uzyskanego po analizie różnicowej ekspresji transkryptów przeprowadzonej testem statystycznym T. Zawiera on (czytając od lewej): identyfikator sondy, id transkryptu, id genu, wartość testu T, zmienność ekspresji (*fold change*).

Przetworzenie surowych danych mikromacierzowych można przeprowadzić z użyciem różnych narzędzi. Autor do celów testowych wykorzystywał publicznie dostępną Platformę integromicznych analiz danych z mikromacierzy DNA https://lifescience.plgrid.pl/pl/users/sign_in. Ta platforma aktualnie pozwala ona na analizę następujących typów mikromacierzy DNA:

- Affymetrix Human Genome U133 (A, A2, B, Plus2, ST)
- Affymetrix GeneChip Mouse Genome (MOE430A, ST, ST2)
- Affymetrix GeneChip Murine Genome (MGU74Av2)
- Agilent SurePrint G3

Platforma oferuje ona dwa rodzaje analiz: niskiego poziomu -obejmującą normalizację, sumaryzację i korekcję tła oraz przyporządkowanie identyfikatorów transkryptów i genów odpowiednim identyfikatorom sond. W toku tej analizy dane surowe mikromacierzy zostają przetworzone do postaci umożliwiającej porównywanie ich między sobą. Każdy producent mikromacierzy posługuje się własnymi identyfikatorami sond. Przyporządkowanie sondy do transkryptu i genu można przeprowadzić przez udostępnione biblioteki notacji oddzielne dla każdego producenta i rodzaju mikromacierzy:

http://www.bioconductor.org/packages/release/BiocViews.html#___AnnotationData.

Druga analiza jest tzw. wysokiego poziomu, która dla porównywanych zbiorów danych mikromacierzowych pozwala wyznaczyć istotność zmienności ekspresji dla każdej sondy (transkryptu). Można to przeprowadzić z użyciem testu statystycznego T lub testu SAM.

Celem analizy niskiego i wysokiego poziomu jest uzyskanie tabelarycznych danych w postaci pliku tekstowego zawierającego informację o zmienności ekspresji.

Oprócz dwóch zbiorów tekstowych dotyczących zmienności ekspresji potrzebne jest punktacja cech sekwencji, czyli parametry CS oraz P_{CT} pochodzące z bazy TargetScan [71][165]. W tym celu wykorzystano wstępne rozpakowanie odpowiednich zbiorów pochodzących z TargetScan w strukturze danych języka R. Umożliwia to biblioteka TargetScoreData [104]. Zgromadzone dane w TargetScan uwzględniają tylko regiony 3'UTR transkryptów.

Ze względu na ciągłą ewolucję liczebności poznawanych cząsteczek miRNAs baza miRBase wprowadziła system oznaczeń wersji. Analiza danych dotyczących miRNAs wymaga zatem uwspólnienia wykorzystywanej wersji miRBase. W tym celu wygenerowano zbiór tekstowy "miRBase_conv_org.txt" na podstawie konwertera wersji miRNA Converter miRandola (http://atlas.dmi.unict.it/mirandola) [144]. Zbiór ten umożliwia konwersję identyfikatorów miRNAs od wersji 15 do 20.

Skrypt napisany w języku R realizujący opracowany model "biTargetScore.R" został przystosowany do pracy lokalnej gdzie potrzebne parametry: ścieżki do plików, parametry obliczeń należy podać

bezpośrednio w nim. Druga możliwość – plik współpracuje z webową aplikacją i oczekuje na wywołanie z deklaracją parametrów z zewnątrz.

Parametry wywołania skryptu są następujące:

1. Ścieżka do pliku zawierającego identyfikatory transkryptów, genów oraz wartości *fold change*.
2. Ścieżka do pliku zawierającego identyfikatory miRNAs oraz wartości *fold change*.
3. Nazwa katalogu plików wynikowych.
4. Identyfikator język ("pl"/"en").
5. Wersja bazy miRBase używana przez mikromacierz miRNA.
6. Wersja bazy miRBase używana przez zasoby TargetScan.
7. Ścieżka do pliku konwersji miRBase.
8. Parametr ograniczający liczbę komórek w wynikowych tabelach prawdopodobieństwa.
9. Wybór poziomu przeprowadzanej analizy: transkrypt lub gen.
10. Parametr modelu VBGMM. Wartość progowa maksymalnej rozbieżności algorytmu VB-EM.
11. Parametr modelu VBGMM. Liczba ograniczająca maksymalną liczbę iteracji algorytmu *Expectation-Maximalization*.

5.3.2 Opis przebiegu skryptu

Poniższe punkty opisują rzeczywisty przebieg działania interpretowanego skryptu "biTargetScore.R".

1. Odczyt plików wejściowych i parametrów. Parametry są zapisane bezpośrednio w pliku lub jako parametry wywołania skryptu przez proces nadrzędny. Pliki z danymi dot. poziomu transkrypcji (*fold change*) odczytywane są przez funkcję *readMulti()*, która w stosunku do standardowych funkcji np. *read.table()*, pozwala na deklarowanie własnych znaków separatorów tekstu.
2. Załadowanie bibliotek dodatkowych: TargetScore (implementacja modelu VB-GMM), TargetScoreData (struktury danych zawierających bazę TargetScan), gdata (parser formatu XLS), plotrx (biblioteka graficzna), plyr(funkcja *join*()).
3. Przygotowanie macierzy danych: macierz zmienności transkrypcji transkryptów/genów logFC zawierająca wartości *fold change*. W tym miejscu następuje procedura uśredniania poziomów transkryptów: wiele transkryptów – jeden gen lub wiele genów – jeden transkrypt.
4. Przygotowanie macierzy zmienności miRNAs. Po odczytaniu z pliku wejściowego odpowiednich kolumn: "miRNA id" oraz logFC, skrypt dokonuje konwersji firmowych identyfikatorów miRNAs (probeSet ID) na identyfikatory miRBase. Następnie dokonuje konwersji wersji identyfikatorów tak by były one zgodne z wersją wykorzystaną w TargetScan. Na tym etapie następuje także odrzucenie tych miRNA, które były pominięte w toku aktualizacji kolejnych wersji miRBase, lub w przypadku, gdy wykorzystujemy mikromacierze z nowszą wersją miRBase to następuje redukcja ilości miRBase do tych, jakie były znane w wersji TargetScan.
5. Analiza algorytmem VB-GMM samych danych zmienności ekspresji. Wykorzystując macierze transkryptów/genów i miRNAs można bez wykorzystania informacji kontekstowej oraz filogenetycznej przeprowadzić obliczenia prawdopodobieństwa

interakcji miRNA/transkrypt. Przeprowadzenie tego obliczenia ma dwa zastosowania: w sytuacji, gdy jakaś para miRNA/transkrypt nie posiada informacji kontekstowej, wówczas wartość prawdopodobieństwa jest przepisywana z tego obliczenia. Drugie – uzyskana macierz prawdopodobieństwa pozwala porównać wyniki i wpływ na nie informacji kontekstowej.

6. Przygotowanie macierzy punktacji kontekstowej (CS). Biblioteka TargetScore pozwala na integrację informacji z różnych źródeł. W tym przypadku następuje przygotowanie macierzy danych informacji o kontekście lokalizacji *site*. Macierz CS zostaje utworzona na podstawie zasobów TargetScore, które wstępnie zostały załadowane w *data frame* przez bibliotekę TargetScoreData. Odpowiednia pętla "sapply" w ciele funkcji pozwala na odczytywanie punktacji kontekstowej dla kolejnego miRNAs pochodzącego z wcześniej przygotowanego zasobu. Użycie standardowej funkcji "merge()", która służy do scalania obiektów typu *data frame* w jedną wspólną strukturę, okazało się niewłaściwe. Pomimo wywołania jej z opcją "sort=False" nie pozostawia ona domyślnej kolejności wierszy. Dlatego zdecydowano się na wybór innej funkcji tj. "join()" z pakietu "plyr". Uzyskana macierz CS posiada wymiar: liczba genów/transkryptów x liczba miRNAs.
7. Przygotowanie macierzy konserwatywności (P_{CT}). Parametry dotyczące konserwatywności przechowywane w pliku "Summary_Counts.txt" odnoszą się do całych rodzin miRNA. Z tego względu oprócz struktury z danymi o konserwatywności należy także odczytać zasób "miR_Family_Info.txt", który pozwoli przyporządkować wartość P_{CT} indywidualnemu miRNA. Analogicznie jak w przypadku macierzy CS odpowiednia pętla pozwoli utworzyć macierz konserwatywności. Uzyskana macierz P_{CT} posiada wymiar: liczba genów/transkryptów x liczba miRNAs.
8. Zgodnie z proponowaną rozbudową modelu TargetScan analizę w modelu bioTargetScan przeprowadza się w dwóch kierunkach: mRNA->miRNA i miRNA->mRNA. Kierunek mRNA->miRNA polega na tym, że pętla iterująca odczytuje kolejne kolumny (czyli miRNA) i dla każdego uzyskanego miRNA uruchamiany jest model VB-GMM. Modelowi temu jako dane wejściowe podaje się wartości *fold change* dla wszystkich transkryptów oraz wartości CS i P_{CT} dla wszystkich genów/transkrytpów i danego miRNA. Wynikiem obliczenia jest wektor prawdopodobieństw interakcji danego miRNA z każdym z puli genów/transkryptów. W każdym z przeprowadzanych iteracji sprawdzane są wartości kolumn macierzy CS i P_{CT}. Jeśli któraś kolumna danych jest wypełniona samymi zerami to wówczas jest ona w trakcie obliczeń pominięta. W sytuacji, gdy obydwie kolumny z macierzy CS lub P_{CT} mają wartości zerowe, to wynik analizy jest brany z p.5.
9. Kierunek miRNA->mRNA realizowany jest analogicznie, tylko tym razem iteracja dotyczy wierszy macierzy (czyli genów/transkryptów).
10. Uzyskane macierze prawdopodobieństw cząstkowych zostają wymnożone iloczynem Hadamarda dając ostateczną macierz prawdopodobieństw.

5.3.3 Opracowanie wyników analizy

Uzyskana macierz prawdopodobieństw z reguły charakteryzuje się znacznymi wymiarami, które nie pozwalają na poglądową interpretację uzyskanych wyników. Dlatego w celu usprawnienia oceny rezultatów zaproponowano szereg graficznych rozwiązań. Opracowanie wyników powinno ułatwić odpowiedź na następujące pytania:

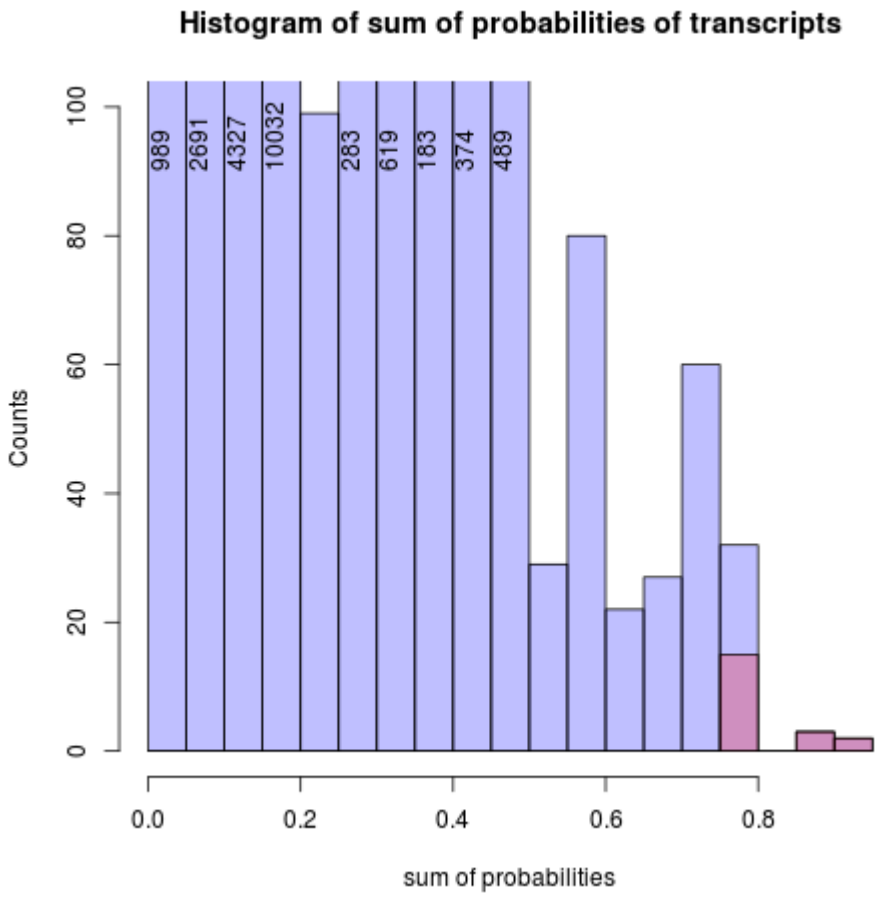
1. Który transkrypt/gen wykazuje największe prawdopodobieństwo bycia targetem jakiegoś miRNA?
2. Który miRNA wykazuje największe prawdopodobieństwo interferencji z jakimś transkryptem?
3. Która para miRNA – transkrypt/gen wykazuje największe prawdopodobieństwo wzajemnej interakcji?

Wynikowe tabele prezentują pary interakcji miRNA-transkrypt/gen, które uzyskują największych wartości. Ilość tych komórek zależna jest od deklarowanego parametru wejściowego. Dodane histogramy pozwalają ocenić stopień rozróżnienia prezentowanych topowych komórek od całej pozostałej populacji interakcji. W przypadku deklaracji analizy na poziomie transkryptu do każdego z topowych transkryptów dodawany jest każdy inny transkrypt, jeśli przynależy do tego samego genu niezależnie od wartości prawdopodobieństwa. Pełny spis plików i grafiki wynikowej:

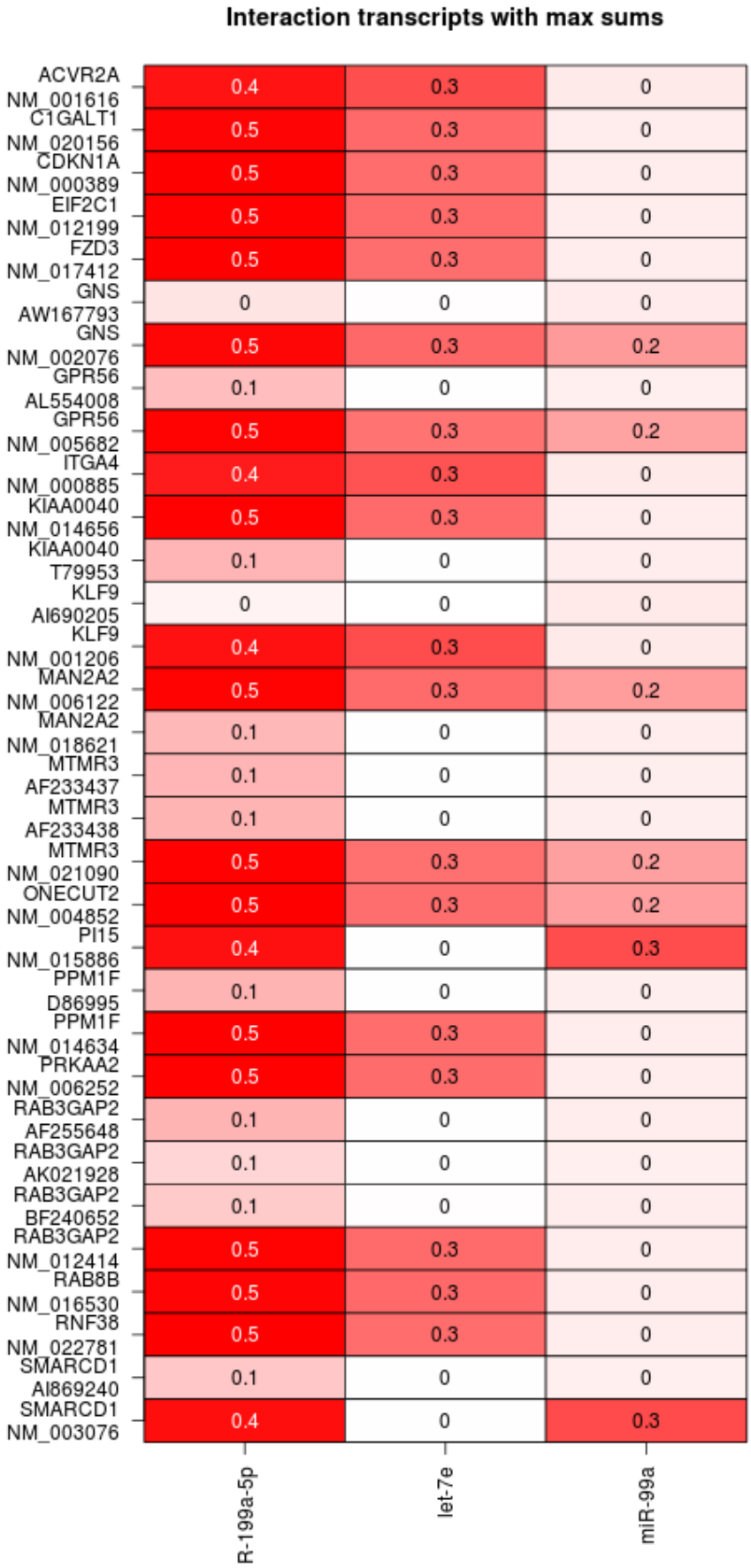
1. lista miRNAs, które nie posiadają żadnej wartości kontekstowych i konserwatywnych;
2. lista transkryptów, które posiadają więcej niż jeden gen;
3. lista topowych genów/transkryptów otrzymana dla samych wartości *fold change*;
4. kompletny plik wyników uzyskanych z analizy, zawiera pełną macierz w postaci pliku tekstowego;
5. histogram prawdopodobieństw wyników z analizy tylko *fold change*;
6. tabela zawierająca największe wartości prawdopodobieństw i odpowiadający im histogram;
7. tabela zawierająca wiersze (transkrypty/geny) wykazujące największe sumaryczne prawdopodobieństwo;
8. tabela zawierająca kolumny (miRNAs) wykazujące największe sumaryczne prawdopodobieństwo;

Opracowanie rezultatów realizuje funkcja *addMultiTrans()*, która uzupełnia tabele wyników o transkrypty przynależące od tego samego genu. Funkcja *hist2file()* pozwala zaznaczenie podzbioru danych wejściowych na histogramie.

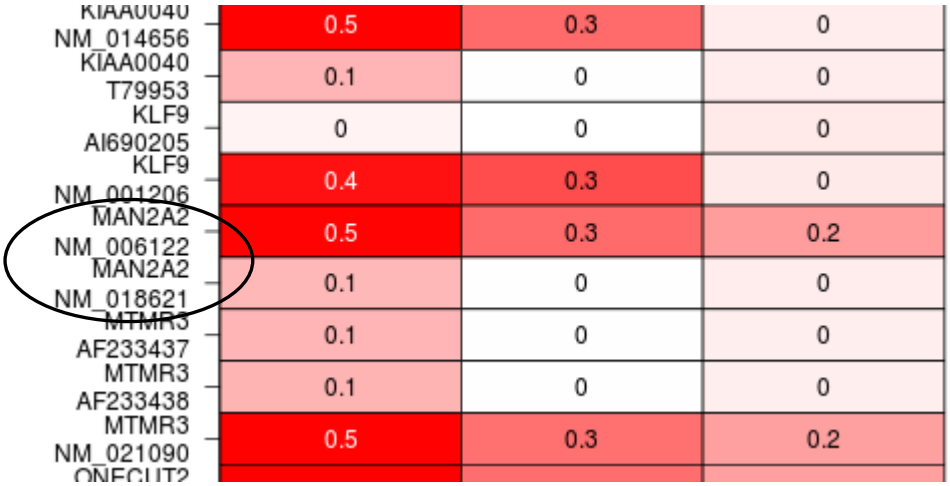
Opisywany skrypt umożliwia przeprowadzenie analizy na poziomie genu lub transkryptów. Większość oferowanych rozwiązań analizy targetów operuje na poziomie genów, poprzez różne sposoby powiązania wielu sond mikromacierzy i obliczenie średniej ekspresji dla różnorodnych transkryptów tego samego genu. W pewnych przypadkach lepszym rozwiązaniem jest uwzględnienie każdego rodzaju transkryptu oddzielnie. Ma to miejsce na przykład w sytuacjach zainteresowania izoformami danego genu lub badając alternatywny splicing. Wybór poziomu transkryptów oznacza, że powiązane wspólnym genem transkrypty zostają uwzględnione w prezentowanych wynikach analizy. W sytuacji odwrotnej, gdy pojedynczy transkrypt związany jest z wieloma genami, skrypt wygeneruje zestawienie takich transkryptów i odpowiadających im genom. Przykładowe rezultaty obliczeń przestawiono na rysunkach Rys. 5.13., Rys. 5.14, Rys. 5.15.



Rys. 5.13. Histogram sum prawdopodobieństw transkryptów



Rys. 5.14. Zestawienie transkryptów o największej sumie prawdopodobieństw



Rys. 5.15. Fragment wyniku analizy 'wiersze z maksymalną sumą'. Zaznaczenie kolorem czarnym dotyczy prezentacji genu MAN2A2 który posiada dwa różne transkrypty: NM_006122 oraz NM_018621. Można zwrócić uwagę, że ten drugi transkrypt nie reprezentuje wartości prawdopodobieństwa przekraczającej próg istotności.

6 Walidacja opracowanego modelu

Weryfikację modelu przeprowadzono porównując rezultaty uzyskane biTargetScore z następującymi programami: TargetScore przystosowanym dla wielu miRNA oraz programem miRanda [48].

Wybór programów był podyktowany uzyskaniem wyników z "sekwencyjnego" wyszukiwania targetów wg podstawowych, uznanych parametrów wyszukiwania. Program miRanda realizuje wyszukiwanie w dwóch etapach: uliniowienia sekwencji cząsteczek miRNA i mRNA metodą programowania dynamicznego. Stopień dopasowania sekwencji oceniany jest na podstawie stopnia komplementarności lokalnego dopasowania. Oprócz par Watson-Cricka punktacja uwzględnia także inne występujące komplementarności nukleotydów: pary wobble G:U. Drugi etap działania programu polega na oszacowaniu stabilności termodynamicznej lokalnego dopasowania. Wyznaczenie energii swobodnej odbywa się na podstawie wygenerowanej fikcyjnej jednoniciowej sekwencji i obliczeniu struktury w pakiecie ViennaRNA [112]. Uzyskanie informacji o energii wiązania stają się szczególnie interesujące także do weryfikacji par określonych w bazie walidacyjnej.

Program TargetScore wybrano ze względów naturalnych, ponieważ został on już zweryfikowany przez jego autora oraz stał się podstawą modelu biTargetScore. Jednak zostanie on w pełni wykorzystany dopiero przy uzyskaniu danych o pełnej informacji nt. ekspresji.

Dane walidacyjne uzyskano z hsa_MTI.xls mirTarBase (Hsu et al., 2011) (http://mirtarbase.mbc.nctu.edu.tw). Porównanie modeli dokonano przez ocenę uzyskanych krzywych ROC i wartości pola pod krzywą AUC. Dane nt. ekspresji uzyskano z eksperymentu Astma [84], który został przeprowadzony m.in na użytek obecnego opracowania.

W ramach walidacji modelu przeprowadzono:

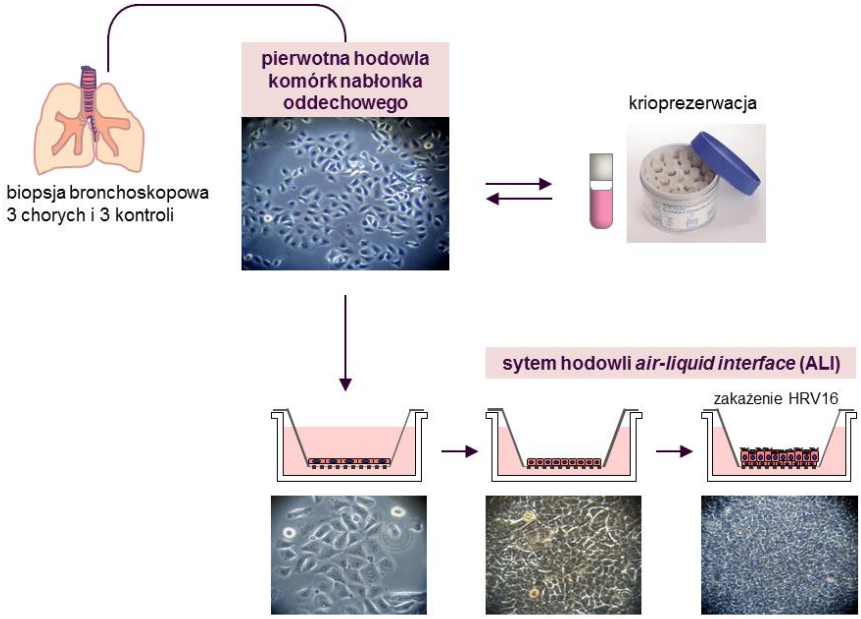
1. Pozyskano dane z eksperymentu "Astma" dla potrzeb walidacji modelu.
2. Zbadano charakterystykę bazy walidacyjnej.
3. Przeprowadzono wyszukiwanie targetów programem miRanda.
4. Wykonana została generacja i charakterystyka zbiorów Cs i P_{CT}..
5. Uruchomiono model biTargetScore dla całego zbioru miRNAs i wybranych genów.
6. Przeprowadzono porównanie metod.

W dalszej części opracowania pytanie o prawdopodobieństwo interakcji danego miRNA z każdym z transkryptów z puli przyjęto nazywać kierunkiem analizy miRNA->mRNA. Analogicznie prawdopodobieństwo interakcji danego transkryptu z każdym miRNA z puli nazwano kierunkiem mRNA->miRNA.

6.1 Eksperyment Astma

W Zakładzie Biologii Molekularnej i Genetyki Klinicznej CMUJ przeprowadzono eksperyment oszacowania wpływu eksperymentalnego zakażenia rinowirusem HRV16 komórek pierwotnych linii nabłonka oddechowego pobranych od osób z rozpoznaną astmą oraz od osób zdrowych [84]. Komórki były hodowane w warunkach umożliwiających ich pełne różnicowanie do fenotypu rzęskowego i kubkowego przez 3 tygodnie w specjalnych insertach na granicy faz pożywka hodowlana- powietrze (*air liquid interface* – ALI) Doświadczenie przeprowadzono dla 3 linii od

chorych na astmę (*patient*) i 3 linii kontrolnych (*control*), przy czym dla każdej z zakażonych linii prowadzono hodowlę bez zakażenia – stanowiąca sparowaną kontrolę do zakażenia. Po zakażeniu rinowirusem (miano 10^7 jednostek zakaźnych *plaque forming units* – PFU) i po upływie 24 godzin przeprowadzono izolację całkowitego RNA ze wszystkich 12 linii eksperymentalnych. Pomiaru profilu miRNA dokonano poprzez przeprowadzenie odwrotnej transkrypcji w obecności kontroli zewnętrznej (*spike-in*) z wykorzystaniem kinazy polinukleotydowej, która wydłużyła cząsteczki miRNA, a następnie odwrotnej transkryptazy ze starterem oligo-dT, która je przepisała do cDNA. W tym etapie został wprowadzony również znacznik fluorescencyjny dla każdej z cząsteczek poddanej odwrotnej transkrypcji. Uzyskano więc 12 prób zawierających odpowiednie wartości ekspresji miRNA. Schemat (Rys. 6.1) ilustruje przebieg eksperymentu.

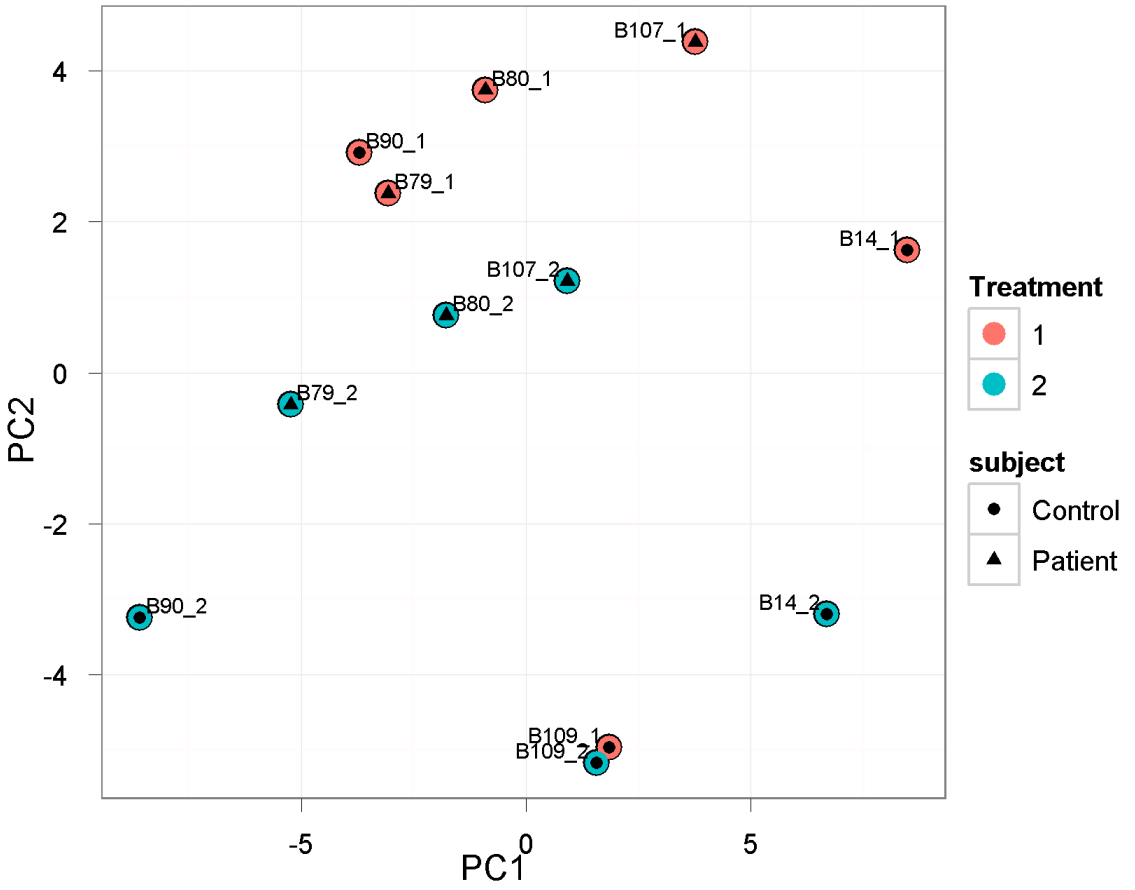


Rys. 6.1. Przebieg eksperymentu "Astma". (Publikacja za zgodą autorów: prof. dr hab. Marek Sanak i dr Bogumił Jakieła).

Profil miRNA oznaczono wykorzystując mikromacierz firmy Exiqon. Hybrydyzację przeprowadzono na matrycy miRCURY LNA™ microRNA Array (7th Gen). Po skanowaniu macierzy uzyskane dane zostały znormalizowane za pomocą pakietu Lowess (*Locally Weighted Scatterplot Smoothing*).

Firma Exiqon stosuje specjalny rodzaj macierzy miRNA, miRCURY LNA™ microRNA Array charakteryzującej się wysoką czułością i swoistością nawet dla dupleksów o dużej ilości par AT. Sondy wykrywające miRNA tą technologią charakteryzuje minimalny błąd niedopasowania. Sygnał hybrydyzacji pochodzi z cząsteczek, które wykazują niewiększą różnicę komplementarności niż 1 nukleotyd. Ponadto macierz charakteryzuje się dużą powtarzalnością, korelacja między macierzami jest na poziomie 99%. Również zakres tonalny (*dynamic range*) jest 5 rzędów wielkości większy w stosunku do innych mikromacierzy (http://www.exiqon.com). Znormalizowanie próby Tm gwarantuje równą detekcję miRNA bez względu na zawartość par GC w dupleksach.

Dla uzyskanych z analizy zbioru danych wartości ekspresji przeprowadzono analizę korespondencji - *Principal Component Analysis* (PCA). Analiza ta stosowana dla zmiennych ilościowych pozwala na graficzną prezentację na wspólnym wykresie każdego wymiaru (tut. próby) poprzez przeprowadzoną redukcję wielowymiarowości. Analiza wykresu (Rys. 6.2) uzyskanego w tej metodzie pozwala na proste i intuicyjne wnioskowanie o powiązaniach zachodzących pomiędzy kategoriami zmiennych, czyli wymiarami.



Rys. 6.2. Współrzędne prób na wykresie dwóch współrzędnych głównych. Symbole odpowiadają kodom hodowli komórkowych, 1 – w warunkach bez zakażenia, 2- po zakażeniu rinowirusem HRV16.

Do dalszych obliczeń wybrano zestawienie wszystkich komórek przed i po infekcji rinowirusem (grupa 2 vs 1 -Tabela 11). Rezultaty obliczeń zostały zawarte w zbiorze "ExpAnalysis1.txt".

Tabela 11. Zestawienie grup.w eksperymencie Astma. Kolorem czerwonym oznaczono porównywane grupy.

	Patient_1	Control_1	Patient_2	Control_2
Patient_1		ExpAnalysis2.xlsx	ExpAnalysis1.txt	
Control_1				
Patient_2				ExpAnalysis3.xlsx
Control_2				

Porównując ze sobą odpowiednie próby wyliczono różnicę średniej zmienności i uzyskano wartość logFC.

Uzyskane wartości logFC poddano testowi statystycznemu test- t na istotność zmienności ekspresji przy pomocy pakietu *limma* [141]. Uzyskane wartości istotności skorygowano poprawką Benjamin-Hochberga. Na podstawie skorygowanych wartości istotności wyselekcjonowano zbiór 30miRNAs o statystycznie istotnej ekspresji różnicowej. Podsumowując uzyskujemy wektor wartości logFC dla zbioru ok 500miRNAs. Całość przedstawionych powyżej obliczeń została przeprowadzona przez firmę Exiqon. W dalszym opracowaniu posłużono się całym zbiorem miRNAs i tym zawierającym wyselekcjonowane miRNAs – 30miRNAs.

6.2 Charakterystyka bazy walidacyjnej

Baza mirTarBase zawiera publicznie dostępne dane zweryfikowane eksperymentalnie. Podane liczby w nawiasach dotyczą wcześniejszej wersji, na której przeprowadzono analizy.

Baza zwiera 39110 (3597) rekordów: unikalnych wartości:

- 12104 (1959) genów
- 596 (432) miRNA w tym: 576 (414) ludzkie i 20 (18) wirusowych: np. ebv, hcmv, kshv, mmu, rno.

Zarejestrowane wirusowe cząsteczki miRNA wiążą się z obecnością ich w hodowlach ludzkich linii komórkowych. Występujące relacje w bazie są wzajemne: jeden miRNA reguluje wiele genów i na odwrót, jeden transkrypt może być regulowany przez różne miRNA.

Podane liczebności bazy zwracają uwagę na dwa aspekty. Pierwszy, że baza rejestruje zaledwie część ok 1/2 znanych genów, których liczbę szacuje się na ok 30 000. Bardziej natomiast uderza fakt, że w toku aktualizacji bazy obserwuje się bardzo mały przyrost liczebności miRNAs z 432 na 596. Informacje nt. statystyki prowadzonej ewidencji można znaleźć: http://mirtarbase.mbc.nctu.edu.tw/php/statistics.php.

Po przeskanowaniu zbioru walidacyjnego uzyskano macierz, której wiersze stanowią geny (1959), a kolumny – miRNAs (478). Lista 478 miRNA pochodzi ze eksperymentu Astma. Stopień wypełnienia macierzy to zaledwie 0,22%. Wymiar aktualnej macierzy (12104 x 576): 6 971 904 pozwala stwierdzić, że mamy macierz wypełnioną w zaledwie 0,54%. Ta informacja jest szczególnie istotna przy przeprowadzaniu oceny modelu.

Drugi aspekt z związany z wykorzystania tej bazy w celach walidacji modelu dotyczy redukcji puli danych ekspresji tylko do tych, które posiadają przynajmniej jeden rekord w bazie walidacyjnej.

Wersja notacji bazy miRBase v 20 (17). Dane w niej zawarte pochodzą z eksperymentów przeprowadzonych różnymi technikami: Reporter assay, qPCR, Western Blot, mikromacierze DNA i *next-generation sequencing experiments*.

6.3 Określanie targetów programem miRanda

Do przeprowadzenia obliczeń programem miRanda wykorzystano sekwencje miRNAs w wersji 19 miRBase [93][70][69][67][92]. Ze względu na przeznaczenie tej analizy skupiono się tylko na zbiorach miRNAs i mRNA, które występują w bazie zwalidowanej i w eksperymencie Astma. Czyli ok 500 miRNAs i 1959 genów.

Pierwszy krok dotyczył przyporządkowania transkryptów do genów i uzyskania odpowiednich sekwencji. W tym celu napisano skrypt Ruby, który operuje na lokalnie zaimplementowanej strukturze BioSQL (http://www.biosql.org) wraz załadowanym do niej zbiorem referencyjnych transkryptów w formacie GenBank "human.rna.gbf" (http://www.ncbi.nlm.nih.gov/). Na podstawie uzyskanych z bazy BioSQL transkryptów wraz z sekwencjami i symbolem genu przygotowano zbiór transkryptów w formacie FASTA. Oczywiście należy pamiętać, że zbiór uzyskanych transkryptów posiada relację: wiele transkryptów – jeden gen. Przy operowaniu na strukturze BioSQL uwzględniono tylko kodujące transkrypty oznaczone prefiksem "NM_*" oraz zachowano informacje o położeniu genomowym CDS każdego transkryptu.

Ze względu na zainteresowanie tylko regionem 3'UTR transkryptów rozważono sprawę relacji wiele transkryptów-jeden gen. Wyrywkowa analiza zbioru pozwala stwierdzić występowanie różnych wariantów transkryptów danego genu, które charakteryzują różne długości nie tylko regionów 3'UTR, ale także 5'UTR lub alternatywnego splicingu. Dlatego do dalszej analizy przyjęto wszystkie warianty (Tabela 12).

Tabela 12. Przykładowe warianty transkryptów

accession	version	length	start_pos	end_pos	gene
NM_001136042	2	2184	347	2107	AARSD1
NM_001142653	1	1923	347	847	AARSD1
NM_001142654	1	1908	347	832	AARSD1
NM_025267	3	1800	146	1723	AARSD1

Drugi krok dotyczy parametrów procedury wywołania programu miRanda. Napisano skrypt Ruby, wywołujący program zewnętrzny miranda z parametrem "–strict", który oznacza wymuszenie występowania w *miejscu wiązania* tzw. regionu *seed* (pozycja 2-8 miRNA). Parametr ten wymusza odrzucenie lokalnych dopasowań, które na pozycji 2-8 nukleotydów zawierają przerwy w dopasowaniu (*gaps*) lub niekanoniczne pary zasad.

Rezultaty uzyskane z analizy miRanda przeskanowano wyciągając informację o położeniu miejsca dopasowania, energii swobodnej, długości dopasowania, procentowego pokrycia i co istotne położeniu miejsca wiązania w jednym z obszarów: 5'UTR, CDS lub 3'UTR. Uzyskano 657 280 rekordy, które następnie konwertowano na tabelę (macierz) o wymiarze 478 miRNA x 1959 genów.

Konwersje na macierz dokonano pozostawiając rekordy zawierające lokalizację zgodną z regionem 3'UTR. Następnie dokonano agregacji rekordów z przeliczeniem energii na prawdopodobieństwa miejsca wiązania i przejściem z multitranskryptów na jeden gen. To przejście zrealizowano wybierając wszystkie niepokrywające się miejsca wiązań w obrębie wszystkich wariantów transkryptów danego genu. Energię wiązania przeliczono na prawdopodobieństwo wg wzoru:

$$p_i = \frac{e_i}{max\ (|e|)} \qquad \textbf{(6-1)}$$

gdzie:

e_i – energia wiązania danego dupleksu;

$max(|e|)$ – maksymalna wartość bezwzględnej energii w całym zbiorze potencjalnych dupleksów.

Agregację wielu miejsc wiązań dla danego genu uzyskano stosując wzór generatywny kumulujący wiele miejsc wiązań w obrębie tego samego genu:

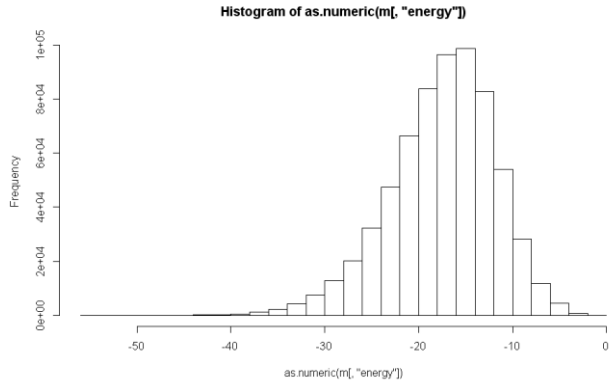
$$p_{tot} = 1 - (1 - p_1)(1 - p_2) \ldots . (1 - p_n) \quad \textbf{(6-2)}$$

gdzie:

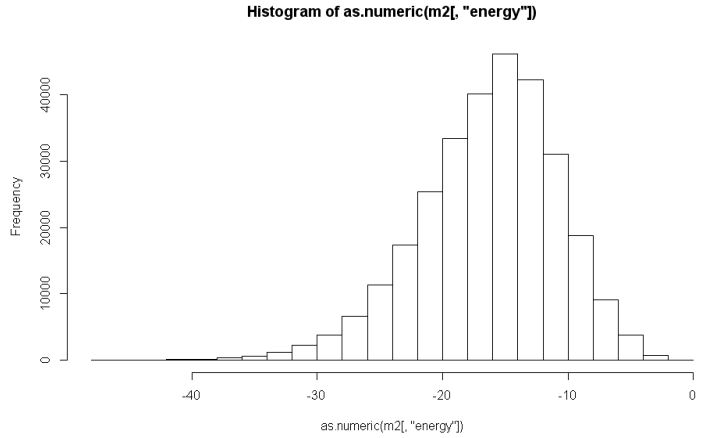
p_{tot} – skumulowane prawdopodobieństwo;

$p_1, \ldots, p_n$ – prawdopodobieństwa wszystkich miejsc wiązań w obrębie jednego genu.

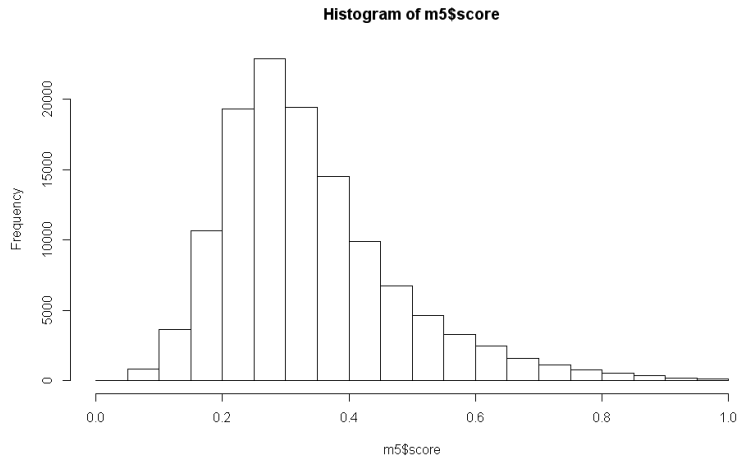
Pewną kontrolę przeprowadzanych obliczeń stanowią załączone histogramy wartości energii i prawdopodobieństwa (Rys. 6.3. , Rys. 6.4. , Rys. 6.5.).



Rys. 6.3. Rozkład energii swobodnej kompleksu dla wszystkich miejsc wiązań

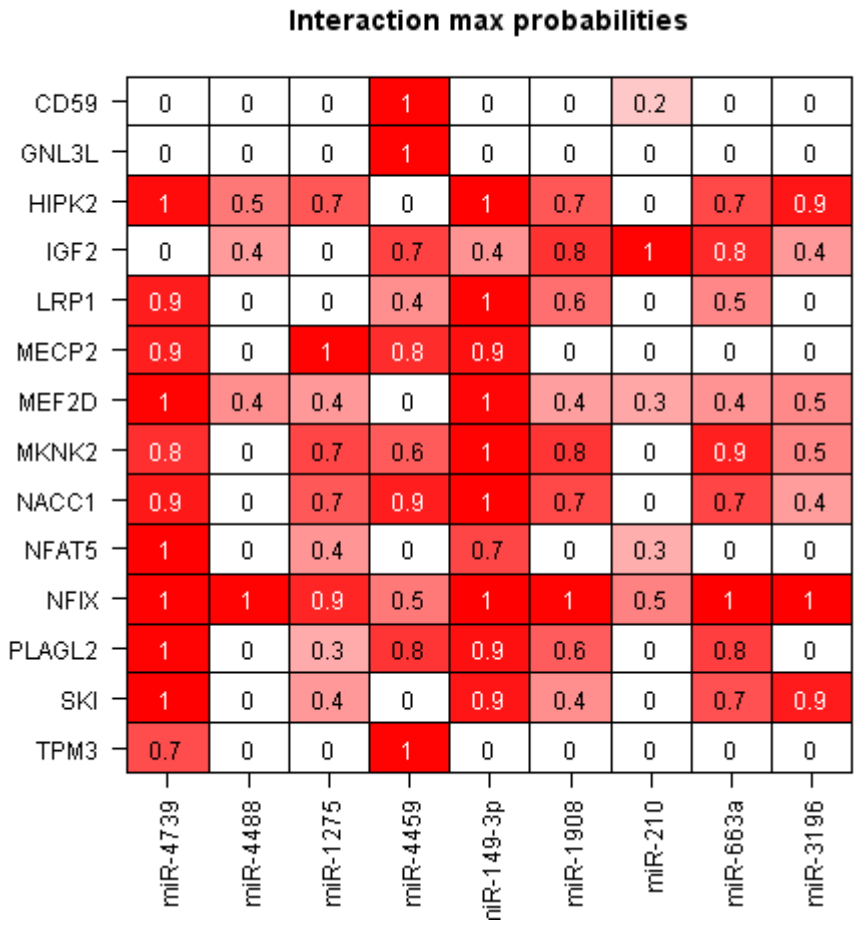


Rys. 6.4. Rozkład energii swobodnej kompleksu dla miejsc wiązań z regionów 3'UTR



Rys. 6.5. Rozkład prawdopodobieństwa po agregacji

Fragment uzyskanych rezultatów przedstawia Rys. 6.6..



Rys. 6.6. Rezultat miRanda. Pierwsze 20 najbardziej prawdopodobnych interakcji

6.4 Charakterystyka zbiorów punktacji kontekstowej i punktacji konserwatywności

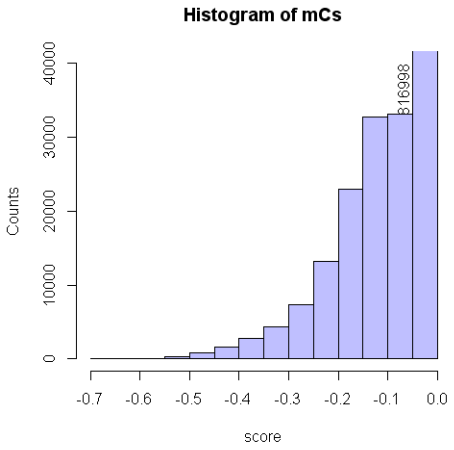
Celem tych obliczeń jest uzyskanie dwóch macierzy mCs i mPCT, które zawierają odpowiednio wartości odnoszące się do kontekstu położenia miejsca wiązania oraz do konserwatywności regionu położenia tego miejsca. Wymiar macierzy został wcześniej zdefiniowany na podstawie bazy walidacyjnej i wynosi 478miRNAs x 1959 genów. Niezależnie od kierunku analizy miRNA-> mRNA czy mRNA>miRNA potrzebne jest uzyskanie tych samych macierzy. Różnica kierunku analizy będzie dopiero uwzględniona na etapie wywoływania biblioteki TargetScan, kiedy odpowiednio zamiast kolumn będą użyte wiersze tych macierzy.

Biblioteka TargetScoreData zawiera wstępnie załadowane zasoby dotyczące punktacji cech sekwencji pochodzące z bazy TargetScanHuman 6.1. Zawierają one punktację kontekstu sekwencji miejsca wiązania Cs i prawdopodobieństwa konserwatywności targetów dla każdej pary miRNA/mRNA (P_{CT}).

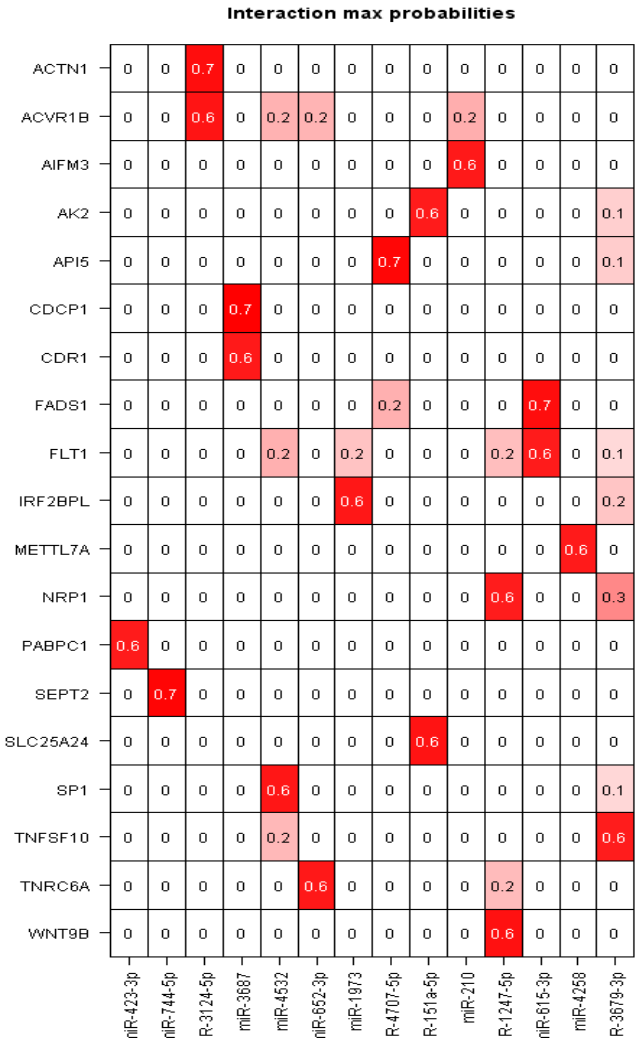
Uzyskanie macierzy mCs polega zatem na selekcji interesujących nas par miRNA/gen. Przy czym agregację multitranskrytpów zrealizowano wybierając najmniejszą wartość punktacji wariantów

transkryptowych genów. Najmniejsza wartość oznacza największe prawdopodobieństwo interakcji miRNA/mRNA.

Uzyskany zakres wartości macierzy mCs ∈ {-0,69...0}. Liczba wartości mniejszych od zera: 168650/933402 ok 18%. Rozkład punktacji Rys. 6.7. Fragment tej macierzy Rys. 6.8.



Rys. 6.7. Rozkład wartości macierzy mCs



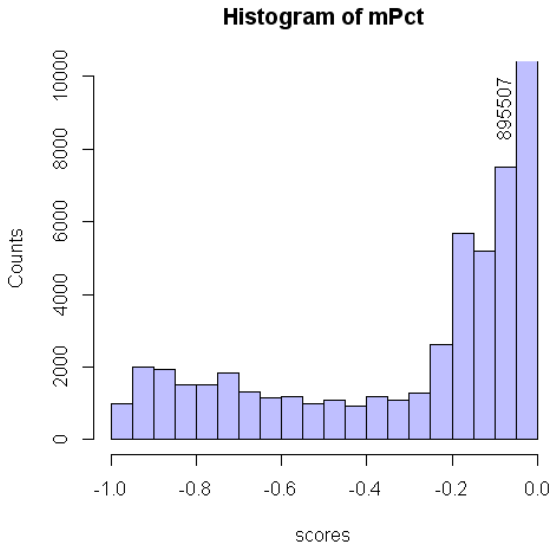
Rys. 6.8. Macierz mCs dla pierwszych 20 dopasowań

Uzyskanie macierzy mPCT chociaż przebiega bardzo podobnie jak macierzy mCs wymaga jednak jednego dodatkowego kroku. Parametr P_{CT} w bazie TargetScan podawany jest względem rodzin miRNA. TargetScan wprowadził własne pojęcie rodzin miRNA które bazuje na identyczności regionu *seed* 2-8 pozycja w dojrzałym miRNA [164]. Każda rodzina ma przypisany stopień konserwatywności:

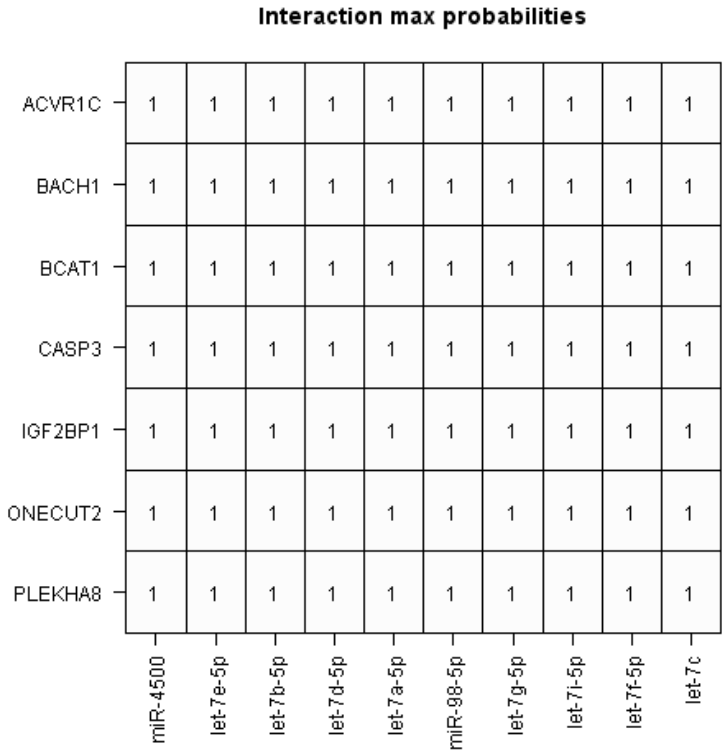
- szeroko konserwatywne (2) – konserwatywność u kręgowców, zwykle aż do zebrafish
- konserwatywne (1) – konserwatywność u ssaków, zwykle do ssaków łożyskowych
- słabo konserwatywne (0) – pozostałe.

Parametr P_{CT} obliczany jest tylko dla rodzin o szerokiej (dużej) konserwatywności [59]. Macierz mPCT uzyskano zatem przypisując wszystkim członkom w obrębie rodziny miRNA tą samą wartość P_{CT}.

Zakres wartości macierzy mPCT$\in$ {-0,99...0}. Liczba wartości mniejszych od zera: 45230/933402 ok 4,8%. Rozkład wartości Rys. 6.9. Przykładowe wartości macierzy Rys. 6.10.



Rys. 6.9. Rozkład wartości macierzy mPCT



Rys. 6.10. Macierz mPCT dla pierwszych 20 najlepszych interakcji

6.5 Analiza biTargetScore

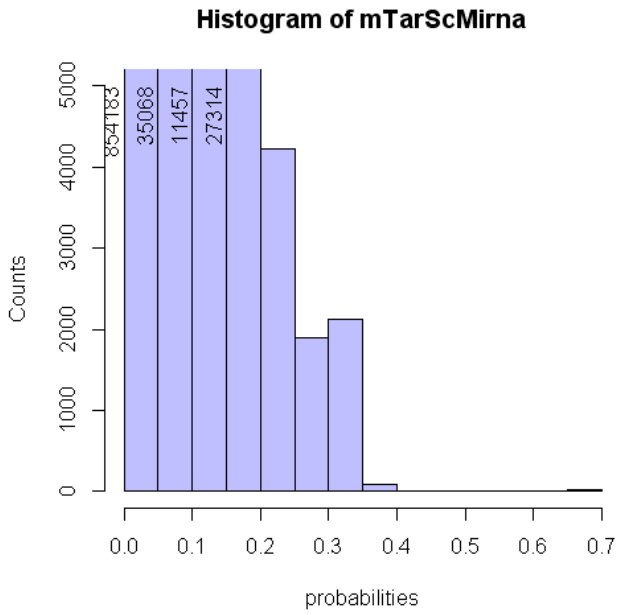
Wyszukiwanie interakcji miRNA/mRNA w przypadku biblioteki TargetScore przeprowadzono zadając pytanie, który miRNA wiąże się z danym targetem. Odwrotny kierunek, czyli jaki target wiąże się z danym miRNA wymagałby dysponowania danymi o zmienności ekspresji genów. Ze względu na brak tej informacji możliwa jest analiza tylko w jednym kierunku.

Modelem biTargetScore przeanalizowano zbiór zwalidowanych targetów (1959) i zbiór ok 500miRNA powiązanymi z wartościami poziomu ekspresji. W analizie uwzględniono wszystkie geny ze zbioru "hsa_MTI.xls", czyli takie, że każdy z nich posiada przynajmniej jeden zwalidowany target. Zbiór miRNAs pochodzi ze zbioru "ExpAnalysis1.txt".

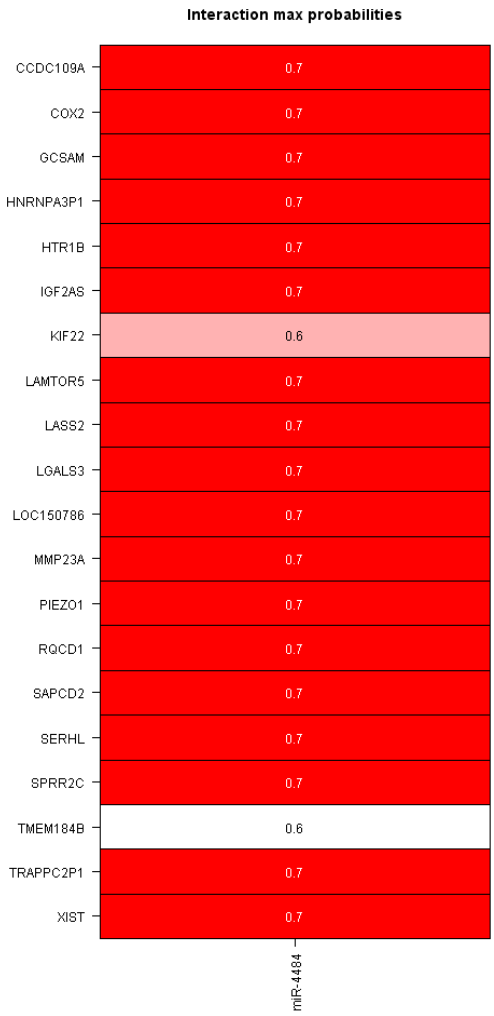
Tabela 13 przedstawia najważniejsze zbiory uzyskane z analizy biTargetScore oraz wcześniejszych analiz. Rys. 6.11. i Rys. 6.12. odpowiednio rozkład i rezultat biTargetScore.

Tabela 13. Zestawienie uzyskanych macierzy danych

macierz prawdopodobieństw	**wymiar**	**opis**
tSFCmiRNA	478x1	wynik biTargetScore uzyskany tylko na podstawie *fold change* miRNA
mValid	1959 x 478	macierz zwalidowana, okrojona do zmiennych istotnych w porównaniu
mRiRanda	1959 x 478	macierz wyników z analizy miRanda
mTarScMirna	1959 x 478	macierz wyników biTargetScore
mPCT	1959 x 478	macierz wyników z TargetScan
mCs	1959 x 478	macierz wyników z TargetScan



Rys. 6.11. Rozkład rezultatów biTargetScore



Rys. 6.12. Rezultat biTargetScore dla pierwszych 20 najbardziej prawdopodobnych interakcji

6.6 Porównywanie metod

Ze względu na kierunek mRNA->miRNA przeprowadzonych analiz, dalsza część pracy wymagała selekcji odpowiednich transkryptów. Preselekcja odpowiednich transkryptów podyktowana jest potrzebą wskazania, dla którego transkryptu zaobserwowano wektor wartości logFC miRNA. Nie dysponujemy odpowiednią informacją o logFC transkryptów. Drugi powód preselekcji transkryptów związany jest ze stosunkowo małą ilością zwalidowanych par miRNA/mRNA, jak również relatywnie niską ilością szacunkowych wartości w bazach Cs i P_{CT}. Można przypuszczać, że baza zwalidowana zawiera znaczą liczbę fałszywie negatywnych wyników, co wynika z zarejestrowanego progresu tej bazy.

Arbitralnej preselekcji genów można dokonać na podstawie:

1. Bazy walidacyjnej - wybieramy geny, które mają największe prawdopodobieństwo interakcji z pulą miRNAs, która istotnie zróżnicowała (30miRNA).
2. Wybieramy geny z największym prawdopodobieństwem uzyskanym w analizie biTargetScore.

3. Wykorzystując strukturę Gene Ontology, która na podstawie selekcji procesu biologicznego wskaże nam pule istotnych genów.

Zdecydowano się wykorzystać bazę zwalidowanych targetów i wybrać transkrypty, które posiadają co najmniej jeden miRNA zawarty w puli 30 miRNAs, wykazującej istotność statystyczną zmienności poziomu transkrypcji.

Z 478 miRNAs z eksperymentu Astma w bazie walidacyjnej 134 posiada co najmniej jedno miejsce wiązania. Z tych 134 miRNA tylko 11 stanowi część wspólną ze zbiorem 30miRNAs. Lista tych 11 miRNA zwiera Tabela 14.

Tabela 14. Lista 11 miRNAs, które posiadają co najmniej jeden target w bazie miRBase oraz zawarte są w zbiorze istotnie zróżnicowanych miRNA w eksperymencie Astma oraz odpowiadające im targety z bazy miRBase. Jednakowymi kolorami zaznaczono współwystępujące te same geny dla różnych miRNAs.

	miRNA	Target
1	hsa-miR-1285-3p	TP53
2	hsa-miR-15b-5p	BCL2 CCND1 CCNE1 VEGFA EIF4A1 RECK
3	hsa-miR-193a-3p	E2F6 MCL1 PTK2 PRAP1
4	hsa-miR-23a-3p	IL6R HES1 POU4F2 ATAT1 CXCL12
5	hsa-miR-23b-3p	E2F1 MET RB1 PLAU
6	hsa-miR-339-5p	BCL6
7	hsa-miR-34b-3p	MYC CDK6 MET VEGFA CDK4 ZAP70
8	hsa-miR-34b-5p	BCL2 MYC CDK6 MET CREB1 CAV1 MYB CDK4 CCNE2 SRSF2
9	hsa-miR-378a-3p	MYC VEGFA GALNT7 TOB2 NPNT
10	hsa-miR-423-3p	CDKN1A
11	hsa-miR-92b-3p	PRMT5 CDKN1C SLC15A1

Baza zwalidowana dla listy 11 miRNAs wskazuje 37 geny. Każdy z tych genów posiada rozpoznany przynajmniej jeden miRNA odpowiednio w zbiorze mCs i mPCT. Z powyższej tabeli wybieramy geny wskazujące w bazie zwalidowanej na największą ilość miRNA. Otrzymujemy 6 genów w tabeli poniżej.

Jest interesujące, że 5 spośród 6 genów uczestniczy w regulacji cyklu komórkowego. *CDK4* i *CDK6* kodują kinazy zależne od cyklin, niezbędne dla rozpoczęcia fazy replikacji DNA poprzedzającej podział komórki. MYC jest czynnikiem transkrypcyjnym uczestniczącym w pobudzeniu proliferacji komórki pod wpływem bodźców zewnętrznych i hamującym jej różnicowanie, MET jest receptorem i czynnikiem transkrypcyjnym pobudzanym przez czynnik wzrostu komórek wątroby (HGF), natomiast produktem *BCL-2* jest antyapoptotyczne białko o tej samej nazwie. VEGF jest cytokiną powodująca powstawanie nowych naczyń krwionośnych – angiogenezę. Hamowanie tych transkryptów pod wpływem zidentyfikowanych miRNA powoduje zatem jednokierunkową odpowiedź komórek pod wpływem zakażenia HRV16 manifestującą się zmniejszeniem ich proliferacji, apoptozą i potencjalnym hamowaniem indukcji nowotworzonych naczyń włosowatych w błonie śluzowej oskrzela u chorych na astmę.

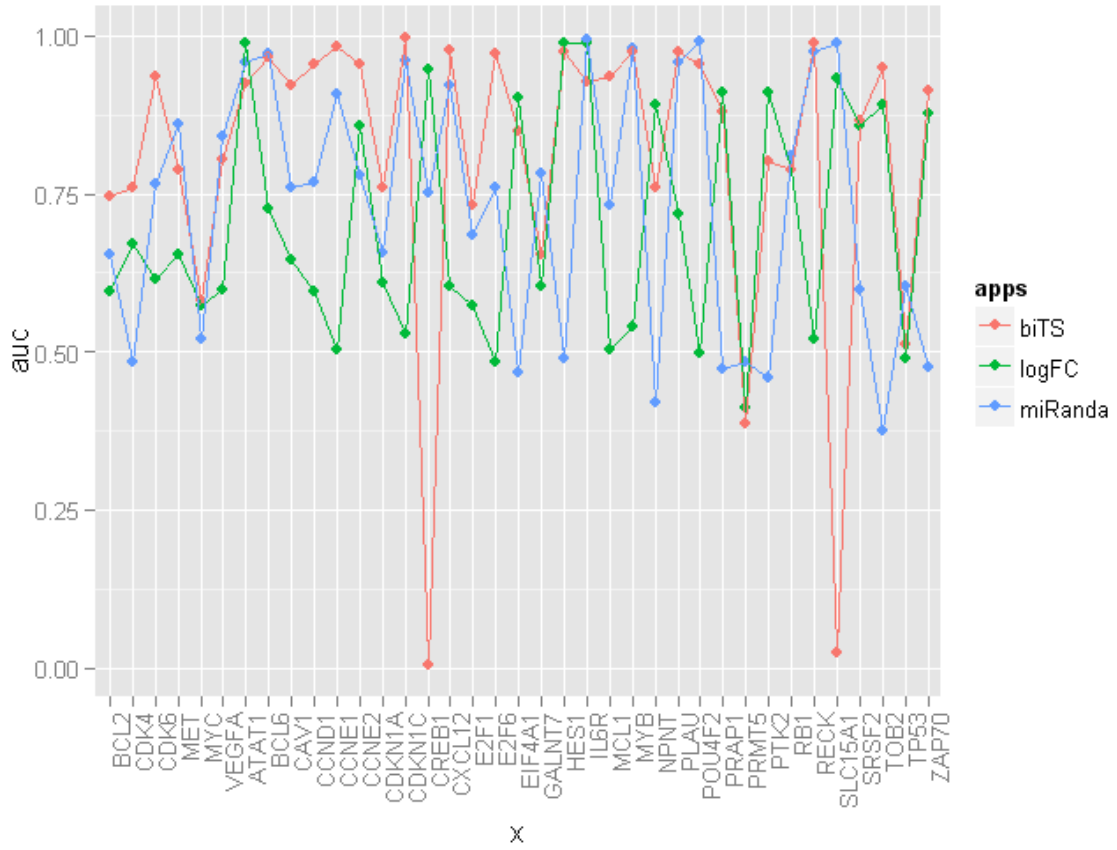
Tabela 15. Zestawienie wybranych 6 genów, które wykazują największą pule regulujących je miRNAs. Kolorami zaznaczono te miRNA, które występują w puli 30miRNAs.

	gen	miRNA
1	BCL2	hsa-miR-15b-5p, hsa-miR-34b-5p, hsa-miR-34c-5p, hsa-miR-449a, hsa-miR-20a-5p, hsa-miR-17-5p, hsa-miR-630, hsa-miR-34a-5p, hsa-miR-29b-3p, hsa-miR-16-5p, hsa-let-7a-5p, hsa-miR-15a-5p, hsa-miR-181d, hsa-miR-21-5p, hsa-miR-29a-3p, hsa-miR-181a-5p, hsa-miR-181b-5p, hsa-miR-29c-3p
2	CDK4	hsa-miR-34b-5p, hsa-miR-34b-3p, hsa-miR-34c-5p, hsa-miR-34a-5p, hsa-miR-302a-3p, hsa-miR-24-3p, hsa-miR-124-3p, hsa-miR-145-5p
3	CDK6	hsa-miR-34b-5p, hsa-miR-34b-3p, hsa-miR-449a, hsa-miR-34a-5p, hsa-let-7b-5p, hsa-miR-29b-3p, hsa-miR-185-5p, hsa-miR-203a, hsa-miR-16-5p, hsa-miR-424-5p, hsa-miR-107, hsa-miR-124-3p, hsa-miR-29a-3p, hsa-miR-29c-3p
4	MET	hsa-miR-23b-3p, hsa-miR-34b-5p, hsa-miR-34b-3p, hsa-miR-34c-5p, hsa-miR-449a, hsa-miR-34a-5p, hsa-miR-30a-5p, hsa-miR-206, hsa-miR-340-5p
5	MYC	hsa-miR-34b-5p, hsa-miR-34b-3p, hsa-miR-378a-3p, hsa-miR-34c-5p, hsa-miR-20a-5p, hsa-miR-17-5p, hsa-miR-34a-5p, hsa-miR-26a-5p, hsa-miR-24-3p, hsa-let-7a-5p, hsa-miR-98-5p, hsa-let-7g-5p, hsa-miR-145-5p, hsa-miR-21-5p, hsa-let-7c
6	VEGFA	hsa-miR-15b-5p, hsa-miR-34b-3p, hsa-miR-378a-3p, hsa-miR-93-5p, hsa-miR-125a-5p, hsa-miR-20a-5p, hsa-miR-302d-3p, hsa-miR-17-5p, hsa-miR-34a-5p, hsa-miR-29b-3p, hsa-miR-106a-5p, hsa-miR-361-5p, hsa-miR-106b-5p, hsa-miR-16-5p, hsa-miR-205-5p, hsa-miR-150-5p, hsa-miR-107, hsa-miR-15a-5p, hsa-miR-147a

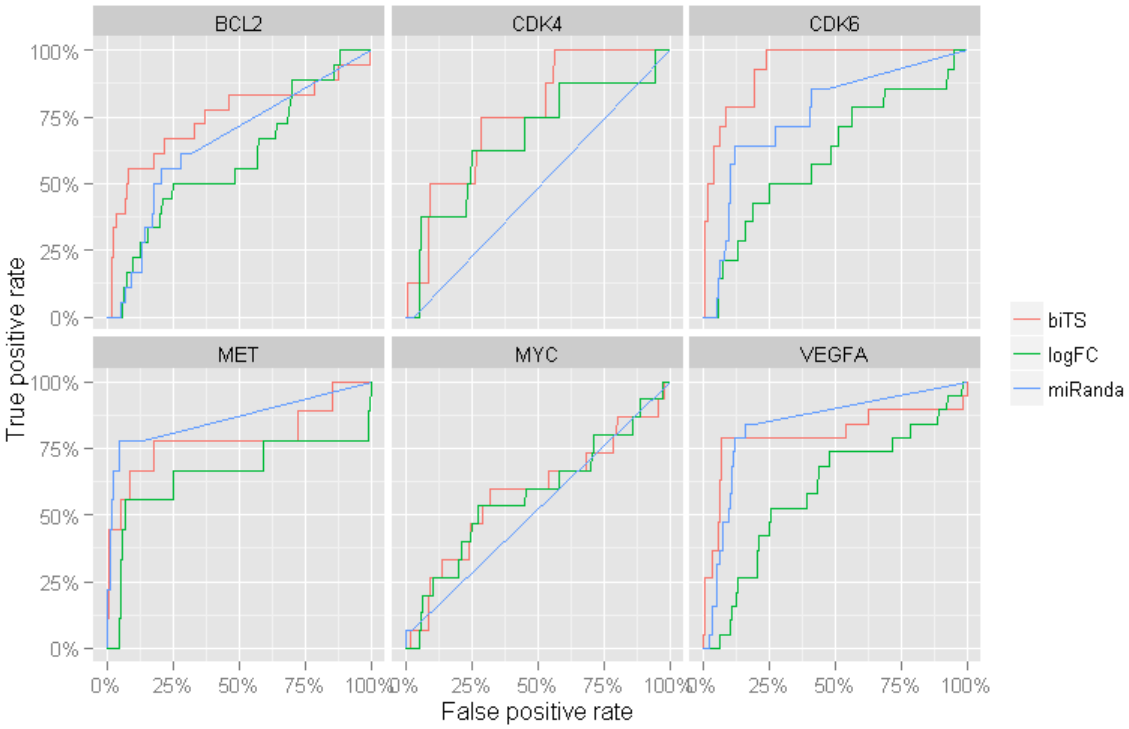
Dalsze analizy będą wykorzystywać obie grupy wybranych genów: szerszą 37 genów i tą mniejszą 6 genów.

Ocenę wartości metod wyszukiwania targetów dokonano wyznaczając dla każdego porównania wartość pola powierzchni pod krzywą ROC – AUC i następnie wykreślając krzywe ROC.

Pole powierzchni pod krzywą dla każdej metody i dla każdego genu z puli 37 genów przedstawia Rys. 6.13. Rzucające się w oczy dwie odstające wartości AUC dotyczą genów: CREB1, SLC15A1.

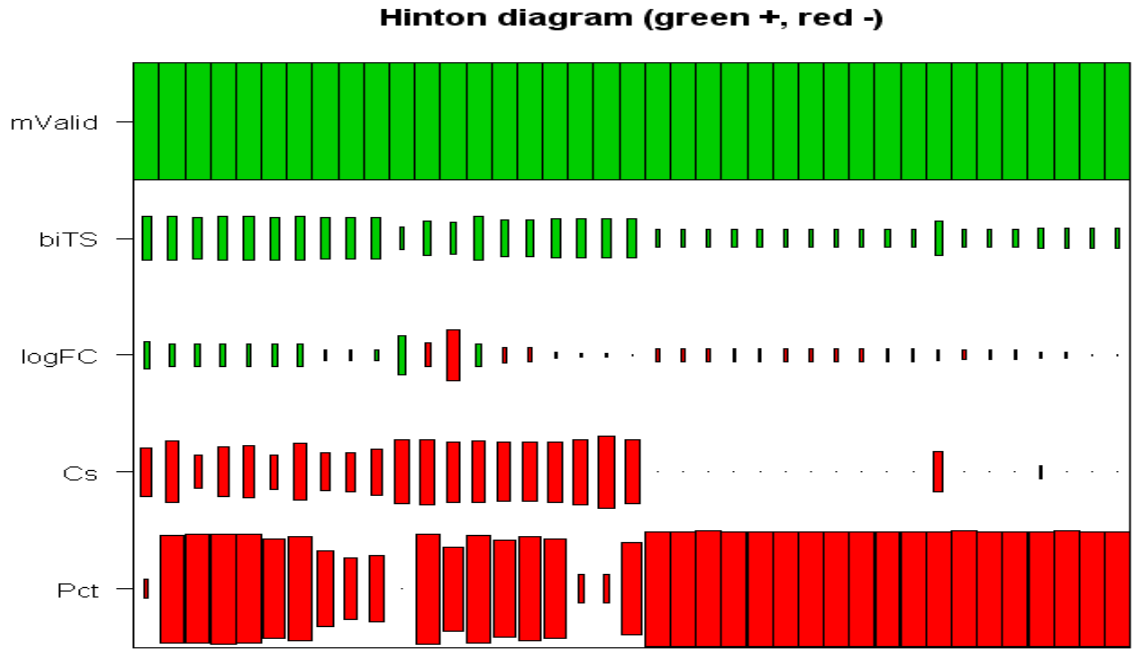


Rys. 6.13. Zestawienie wartości AUC uzyskanych różnymi metodami dla puli 37 genów



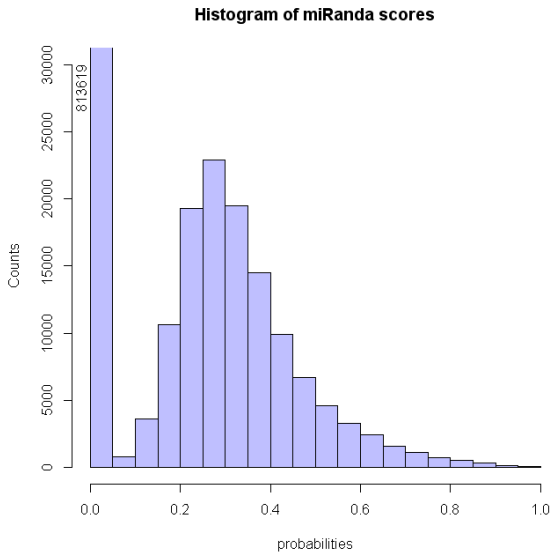
Rys. 6.14. Krzywe ROC dla wybranej puli 6 genów. Metody: biTS – biTargetScore, logFC – TargetScore wyznaczony tylko dla logFC miRNA, miRanda.

Rys. 6.14. przedstawia zestawienie krzywych ROC dla każdego genu z puli 6 genów. Rys. 6.15. przedstawia zestawienie wartości wejściowych modelu biTargetScore z wartością wyjściową – prawdopodobieństwem dla 10 par o największych wartościach prawdopodobieństwa miRNA/mRNA, które równocześnie występują w bazie zwalidowanej. Przedstawiony diagram Hintona ma na celu kontrole przeprowadzonych obliczeń. Kolor reprezentuje znak sygnału, wielkość prostokąta relatywną jego wielkość.

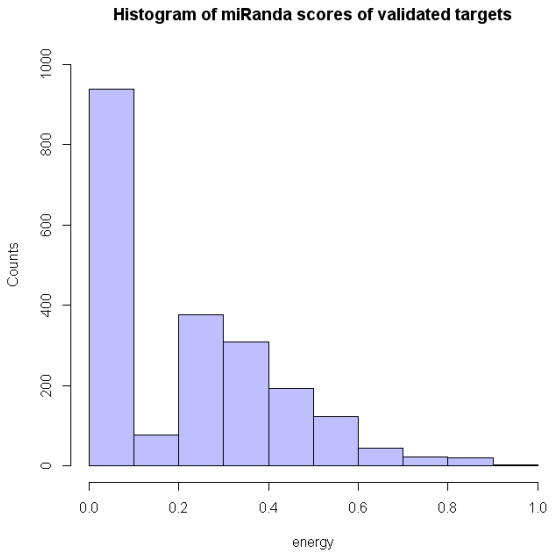


Rys. 6.15. Diagram Hintona dla pierwszych 10 "największych" wartości, kolejno: biTS, Cs, P_{CT}. Oznaczenia: mValid – wartość dodatnia (kolor zielony) wskazuje na występowanie tej pary miRNA/mRNA w bazie zwalidowanej, biTS – rezultat metody biTargetScore, logFC, Cs, P_{CT} – wektory danych wejściowych dla modelu biTargetScore.

Uzyskane w toku analizy dane pozwalają na weryfikację wartości energii swobodnej powstających dupleksów miRNA/mRNA. Szczególnie ciekawym wydaje się sprawdzenie tej energii w puli zwalidowanych par. Oficjalnie podaje się ze względu na mały stopień komplementarności sekwencji dupleksów relatywnie niską wartość tej energii. Przeprowadzając rozpoznawanie targetów metodą miRanda uzyskaliśmy informację o szacunkowej energii swobodnej znalezionych dopasowań. Na pytanie, jaka jest relacja między energią wiązania zwalidowanych par miRNA/mRNA oraz wszystkich rozpoznanych par przez miRanda odpowiadają dwa histogramy (Rys. 6.16., Rys. 6.17.).



Rys. 6.17. Rozkład energii miRanda wszystkich uzyskanych par miRNA/mRNA.



Rys. 6.16. Rozkład energii miRanda dla puli zwalidowanych par miRNA/mRNA.

Przedstawione rozkłady potwierdzają dominację relatywnie niskiej energii dla większości zwalidowanych par: ok 30% maksymalnej wartości modułu energii dopasowania. Wysokie pierwsze słupki wskazują jednocześnie na brak rozpoznania wielu par występujących w bazie zwalidowanej.

7 Podsumowanie

7.1 Dyskusja osiągniętych wyników

Zarówno dopasowywanie modelu, jak i jego wybór w świetle badanego zjawiska, ograniczają w sumie rozpatrywany poziom zjawiska. Zwykle uznaje się, że im mniejsza skala analizy zjawiska, tym głębsze wniknięcie w istotę i strukturę wyjaśnianego obiektu i większa moc jego wyjaśniania. Idąc tym tropem, prawa chemii kwantowej mają większy priorytet nieomylności niż "empiryczne formuły" reakcji chemicznych [129]. Inne modele adoptujemy dla poziomu molekularnego, inne dla submolekularnego i wreszcie inne dla systemowego ujęcia (sieciowego, zbiorowego). Złożoność mechanizmu interferencji RNA jak i wielu innych wspominanych w kontekście centralnego dogmatu w biologii molekularnej (patrz rozdział 2. Biologiczne podstawy regulacji genów) "zmusiły" biologów do uznania matematycznych i technicznych idei w biologii. Podobnie jak kiedyś miało miejsce połączenie "nieżywionej" chemii z biologią we wspólną kooperację – biochemię.

Technika w odniesieniu do układów mechanicznych, pneumatycznych czy elektronicznych posługuje się pojęciem **sprawności**, rozumianym, dla układów pasywnych, jako efektywność przekształcania energii wejściowej w wyjściową. Ze względu na przyjęty jakościowy sposób interpretacji mechanizmu RNAi bardziej trafne będzie użyć pojęcia z ekonomii i zarządzania, które **sprawność** w procesie produkcji określają przez jakość i efektywność opisanych relacją uzyskanych efektów do poniesionych nakładów. Rzeczywisty mechanizm biologiczny można oceniać tylko poprzez parametryzowanie przyjętego modelu. Nie chodzi tutaj o ocenę stopnia dopasowania modelu do rzeczywistości, tylko o sprawność rozwiązań, mechanizmów biologicznych. Okazuje się wtedy, że dobór konkretnego modelu narzuca nam z góry pewną jego jakość. Przyjmując model regulacyjny mechanizmu RNAi, *a priori* nisko oceniamy rzeczywisty mechanizm w kategorii sprawności i efektywności. Nic dziwnego, że budzi to pewien sprzeciw i refleksję u humanistów.

Pojawia się pytanie, jakimi mechanizmami w stanie fizjologicznym posługuje się biologia? Poprowadzenie granicy rozdzielającej patologię od fizjologii stanowi wyzwanie dla nauki, związane z metodologiczną umiejętnością analiz głównie stanów patologicznych, które są konsekwencją inwazyjności metod i ich analitycznego kierunku (brak możliwości rekonstytuowania całości z otrzymanych w toku analizy składników, tzn. uprzednio rozdzielonych na jednostki niższego rzędu, a później ponownie odtworzonej całości z tych jednostek [188]). Dopiero wtórnie, często autorytatywnie czy arbitralnie, część mechanizmów funkcjonujących w patologii uznaje się za normalne- fizjologiczne.

Zaprezentowany model nie jest *stricte* modelem mechanizmu rozpoznawania targetów przez kompleks miRNA-RISC. Stanowi on świadome, probabilistyczne – jako ograniczenie eksperymentalne - ujęcie zagadnienia rozpoznawania targetów, uwzględniające czynniki bezpośrednie (komplementarność) i pośrednie (filogeneza). Czynniki określane tutaj jako pośrednie nie występują w rzeczywistej sytuacji w obrębie organizmu. W przeciwieństwie do czynników bezpośrednich, które w połączeniu z kontekstowością oraz prawdopodobnie dodatkowymi czynnikami sterującymi tym procesem realizują stabilne połączenie częściowo komplementarnego dupleksu miRNA/RNA. To antynomiczne stwierdzenie, nierozwiązywalne na razie na gruncie biochemii, być może stanie się okazją do odkrycia nowych własności natury - umiejętności operowania immunologii na poziomie cząsteczek RNA. Jakie to są te czynniki, które

nadrabiają ten niski stopień komplementarności? Czyli, w jaki sposób następuje rzeczywiste rozpoznanie targetów? Poszukiwania w tym kierunku cały czas trwają.

Kolejna refleksja pojawia w związku ze sposobem oceny modeli, narzędzi bioinformatycznych służących do rozpoznawania targetów. Ocenę tą dokonuje się na podstawie danych eksperymentalnych, podczas gdy te same dane zostały wcześniej wykorzystane do usprawnienia własnego modelu, albo do lepszej parametryzacji obiektu. Na przykład model TargetScore w swoim działaniu wykorzystuje parametry kontekstowe sekwencji pochodzące z zasobów TargetScan. TargetScan z kolei swoje parametry wyznaczył na podstawie danych potwierdzonych eksperymentalnie, które z kolei służyły do walidacji TargetScore'a. Analogicznie sytuacja wygląda z parametrami filogenetycznymi, które w celu poprawy swojej jakości wykorzystują informację uzyskaną z bazy danych eksperymentalnych dotyczących kontekstowości. Ten problem metodologiczny pojawia się wtedy, gdy dysponujemy ograniczoną pulą metod i technik weryfikacji uzyskanych danych. Nie ominął on także obecnego rozwiązania. Konsekwencją takiej metodologii może być powielanie błędów lub założeń poczynionych w przeszłości, jeśli one występują.

Na podstawie wyników walidacji można stwierdzić, że zaproponowany w pracy model ma następujące zalety:

1. Uwzględnia profil ekspresji transkryptów oraz miRNAs.
2. Integruje dane kontekstowe i filogenetyczne.
3. Wykazuje przewagę pod względem czułości i swoistości względem porównywanych innych rozwiązań.

Predykcja funkcjonalności cząsteczek miRNAs zmaga się z próbą modelowania złożonego i nie w pełni poznanego mechanizmu regulacji genów. Do wyjaśnienia mechanizmu RNAi u zwierząt oprócz modelu biochemicznego, wykorzystuje się także metody z teorii grafów do wyjaśnienia wzajemnych relacji między elementami całego zbioru miRNAs. Hipoteza o kontroli jakości transkryptów w RNAi sugeruje przez analogię do technicznych modeli poszukiwanie rozwiązań na niższym poziomie organizacji materii. Skoro niższy poziom wyjaśniania mechanizmów uznaje się za bardziej podstawowy i jednocześnie bardziej "tolerancyjny", oznacza to, że każde tłumaczenie na wyższym poziomie powinno prowadzić do pewnych niewiadomych i niejasności, które wyjaśnia dopiero ten niższy poziom. Obecnie przyjęty model procesu interferencji RNA, jako mechanizmu selektywnej regulacji poziomu ekspresji transkryptów, budzi pewne wątpliwości wynikające choćby z usytuowania go pomiędzy różnymi mechanizmami kontroli jakości transkryptów lub kontroli potranslacyjnej (Rys. 2.3), które rozlokowuje się na poszczególnych etapach produkcji białka. Destrukcja półproduktów, jaka zachodzi w mechanizmie RNAi jest z punktu widzenia procedury produkcji najgorszym z możliwych realizacji regulacji poziomu transkryptów w cytoplazmie. Aktywność regulacji RNAi może mieć uzasadnienie przy zaburzeniach procesu transkrypcji lub zakłóceniach toru transmisji informacji. Tą wątpliwość poruszono już w rozdziale "Biologiczne podstawy regulacji genów". Nawet odmienne parametry dynamiki mechanizmu RNAi w porównaniu do regulacji transkrypcyjnej nie uzasadniają uznania tego mechanizmu za fizjologiczny element regulacyjności. W dziedzinie czasu regulacja RNAi może wykazywać pewną przewagę nad regulacją transkrypcyjną: np. reakcji komórki na nagłe zmiany środowiskowe.

W wyjaśnianiu mechanizmu RNAi nie można także wykluczyć potranslacyjnych mechanizmów, które mogą oddziaływać i wykorzystywać mechanizm RNAi. Nieprawidłowości wynikające z zaburzeń transkrypcyjnych stwierdzane na etapie filtracji białek, podczas kontroli ich

pofałdowania powinny być korygowane, jeśli nie bezpośrednio u źródła to przynajmniej na etapie RNAi. Będzie to możliwe tylko wtedy, jeśli "punkt pracy" RNAi będzie się znajdował pomiędzy pełną komplementarnością a całkowitym brakiem komplementarności umożliwiającej sterowalność *in situ* tego procesu. Cechy mechanizmu RNAi, które potwierdzają złożoność tego mechanizmu:

1. Brak pełnej komplementarności między miRNA i jego targetem wprowadza niejednoznaczność wiązania się par miRNA/mRNA, czyli umożliwia modulację poziomu dupleksów miRNA/mRNA innymi czynnikami. (Przez analogię do fizycznych właściwości półprzewodników, których przewodność lokuje się pomiędzy izolatorami i przewodnikami. Organizacja przestrzenna półprzewodników w układ np. tranzystora pozwala nawet na polową regulację jego przewodności np. w tranzystorach typu CMOS.)
2. Brak pełnej komplementarności między miRNA a mRNA pozwala na regulację jednym miRNA wielu różnych transkryptów, ale jak to zostało wykazane dla podzbiorów miRNAs o tym samym *seed*. Mechanizm RNAi u kręgowców wprowadza mniejszą specyficzność rozpoznawania targetów w porównaniu do roślinnego mechanizmu RNAi. Ten fakt trudno uznać za gradację jakości działania RNAi, raczej należy uznać niepełność wiedzy o tym mechanizmie i jego znaczeniu w dużym przeskoku jakościowym jaki się ujawnia w immunologii zwierząt.
3. Generalnie nie występują sytuacje pełnej komplementarności (patrz Rys. 6.17., Rys. 6.16.), sugeruje to, że pełna komplementarność jest zabroniona, ponieważ sama wprowadza jednoznaczną hybrydyzacje, a tym samym nie pozwala na regulację poprzez dodatkowe, hipotetyczne czynniki.
4. Konsekwencją relacji wiele do wielu miedzy miRNAs i zbiorem transkryptów jest to, że sekwencje części poza *seed*, od strony 3' miRNA roboczo nazwane 15'tką (od średniej ich długości) musi się znajdować w odległości "bezpiecznej" pomiędzy wszystkimi jej targetami – dokładnie ich części poza miejscem wiązania. Pojęcie odległości dotyczy stopnia podobieństwa między sekwencjami.

Analiza długości sekwencji miRNA (5.2 Analiza zbioru sekwencji miRNA) wskazuje na podobieństwo rozkładu tych długości do rozkładu normalnego (Rys. 5.2.). Średnia wartość długości to ok 21nt. Zbiór sekwencji miRNAs charakteryzuje zatem w miarę jednolita długość. Nie jest znana odpowiedź na pytanie, jakie czynniki decydują o rozrzucie długości. W miarę jednolita długość może świadczyć o trafnej klasyfikacji tych cząsteczek RNA do wspólnej grupy, jak również o wspólnym mechanizmie ich genezy i funkcyjności. Dalsze przypuszczalne znaczenie długości omówione zostało w rozdziale 7.2 Plan dalszych prac.

Badanie występowania częstości homologów miRNA w sekwencjach transkryptów wydaje się być interesujące ze względu na specyfikę wiązania się miRNA z targetem. W obrębie transkryptów przy relatywnie wysokim progu podobieństwa z miRNA na nici komplementarnej nie znaleziono większości dopasowań (Tabela 8). Komplementarnych homologów miRNAs na sekwencjach transkryptów znaleziono jedynie 378. Wartość ta odstaje i jest znacząco mniejsza od liczby znalezionych homologów (1760). Zestawione wynik (Tabela 8) potwierdzają specyfikę działania RNAi opartą na częściowej komplementarności tworzonych dupleksów. Równocześnie stanową one wskazówkę, jak może wyglądać kompozycja nukleotydowa niepoznanych jeszcze cząsteczek miRNAs. Większą liczbę dopasowań identyczności - homologów (1760) można tłumaczyć tym, że te lokalne dopasowania dotyczą genów cząsteczek miRNAs.

Jaka jest zależność między liczbą miRNAs, a liczbą odpowiadających im transkryptów? Na to pytanie odpowiadają Rys. 5.3. i Rys. 5.4. Zgodnie z oczekiwaniami obserwujemy szybki spadek liczby miRNAs dla zwiększającej się liczby transkryptów dla homologów. Mniejszy dla grupy częściowych dopasowań sekwencji. Tabela 9 zwraca uwagę te miRNAs, które posiadają dużą liczbę komplementarnych duplikatów np. jeden miRNA posiada 873 odpowiedników na nici komplementarnej. Te wysokie wartości odpowiedników wynikają z faktu, że wykorzystany w badaniu zbiór referencyjnych transkryptów zawiera także transkrypty z alternatywnych splicingów oraz paralogi (patrz rozdział 3.2. Dane i zasoby informacji).

Jedynie 205 na 1921 miRNAs posiada relatywnie dobre dopasowanie na nici komplementarnej. Z drugiej strony analiza specjalistycznym narzędziem miRanda (Rozdział 6.3. Określanie targetów programem miRanda) znajduje potencjalnych dopasowań znacznie więcej. Można, zatem wnioskować, że zbiór miRNAs charakteryzuje generalnie mały stopień komplementarności względem sekwencji transkryptów, być może w ten sposób zabezpieczając mechanizm RNAi przed możliwością jednoznacznego wiązania się każdego miRNA z jakimś transkryptem. Sekwencje miRNAs prawdopodobnie są tak dobrane, aby nie "kolidowały" ze sekwencjami transkryptów. Wykorzystany zbiór transkryptów ok 46 tyś. sekwencji zawiera kodujące i niekodujące RNA. Znalezione dopasowania dla nici dominującej mogą stanowić fragmenty genów ulokowanych, czy to w obszarze egzonów genów kodujących, czy to własne geny miRNAs.

Przeprowadzone badanie duplikacji homologów miało cel poglądowy (rozdział 5.2.2 Duplikacje homologów w obrębie transkryptów). Wykazuje ono, że sekwencje niektórych miRNAs znajdujemy w obrębie sekwencji komplementarnych pewnych transkryptów nawet wielokrotnie (Tabela 10), przy czym największa, stwierdzona krotność wynosi 4 (Rys. 5.5.). Podane w zestawieniu (Tabela 10) identyfikatory miRNAs można wykorzystać przy szczegółowej analizie konkretnego szlaku biologicznego.

Teoria informacji Shanonna dąży do pogodzenia dwóch przeciwstawnych celów: zwięzłości zapisu informacji i ochronę informacji przed zakłóceniami podczas transmisji. Komunikat zawiera tym więcej informacji, im mniejsze jest prawdopodobieństwo jego wystąpienia - jedno z podstawowych założeń ilościowej teorii informacji. Entropia jest maksymalna, gdy prawdopodobieństwa zdarzeń są takie same. Przeprowadzone wyliczenie entropii blokowej dla sekwencji zbioru miRNAs ma na celu ich porównanie ze zbiorem losowych sekwencji o takiej samej długości. Przedstawione zestawienia wykazują bardzo podobne wartości entropii dla wybranych podgrup (Rys. 5.6, Rys. 5.7, Rys. 5.8, Rys. 5.9, Rys. 5.10, Rys. 5.11). Oznacza to, że niepewność albo liczba możliwych kombinacji nukleotydowych jest podobna do maksymalnej entropii. Potwierdza to, że naturalne sekwencje miRNAs charakteryzuje losowość, ale trzeba pamiętać, że mamy do czynienia ze stosunkową małą liczbą miRNAs w obrębie testowanych grup. Idąc dalej, opierając się na podanej definicji entropii, informacyjność zbiorów jest bardzo mała. Ostatecznie selekcja ewolucyjna tych cząsteczek bardziej odpowiada procesom losowym aniżeli deterministycznym. Tym samym wykazując podobieństwo do analogicznych analiz przeprowadzonych na sekwencjach aminokwasowych [123][175][159].

Porównania różnych metod rozpoznania targetów dokonano w rozdziale 6.6. Uzyskane parametry AUC i krzywe ROC wskazują na słusznie obrany kierunek poszukiwań udoskonalonego rozwiązania. Poza dwoma odstającymi wartościami AUC pozostałe wartości dla biTargetScore są porównywalne albo lepsze w stosunku do tych znalezionych za pomocą miRanda. W celu pełniejszej weryfikacji modelu należy zmierzać do pozyskania odpowiednich danych

eksperymentalnych mikromacierzowych o sparowanych ekspresjach miRNA/mRNA. Powinny one umożliwić lepszą weryfikację modelu i dobór funkcji dopasowującej poziom ekspresji miRNA.

Diagram Hintona pozwala na naoczną obserwację zachowania się modelu dla danych z eksperymentu "Astma" (Rys. 6.15.). W pierwszej części (kolumnie) diagramu, która powstała dla największych uzyskanych wartości prawdopodobieństwa (biTS), obserwujemy dodatnie wartości logFC, oraz średnie wartości Cs i duże wartości P_{CT}. Co jest zgodne z oczekiwaniami. Druga kolumna diagramu dla wartości maksymalnych Cs wskazuje na przynajmniej dwa przypadki, kiedy pomimo ujemnego przyrostu poziomu ekspresji danego miRNA, model wyznaczył średnią wartość biTS. To świadczy, że nie można mówić, że największą determinantą w modelu jest wartość logFC. Ostatnia kolumna diagramu dla największych wartości P_{CT} - generalnie tam występujące wartości nie korelują (z jednym wyjątkiem) z dużymi wartościami Cs, stąd biorą się małe wartości biTS. Jedynie dla tego wyjątku obserwujemy większą wartość biTS.

Biblioteka TargetScore predysponuje do określania targetów, które zostały zwalidowane, aniżeli do wskazywania nowych, jeszcze nieodkrytych. Niewątpliwie, aby rozwijać kierunek rozpoznania nowych par miRNA/mRNA należałoby udoskonalać metodę "sekwencyjną" miRanda. Natomiast zastosowanie TargetScore w przypadku, gdzie szukamy odpowiedzi na występowanie potwierdzonych targetów jest wskazane ze względu na dobre parametry uzyskane względem bazy zwalidowanej.

Prezentowany model oprócz zdolności predykcji (*in situ*) można także wykorzystać w celu znalezienia korelacji poziomu ekspresji miRNA i mRNA w procesie interferencji RNA. Interesujące jest czy poziom tych miRNAs, które są aktywne w badanej próbce zmienia się w sposób zauważalny w pomiarach. W niektórych opracowaniach - modelach dla uproszenia przyjęto stały poziom miRNAs [107].

Zastosowanie w obliczeniach lokalnych repozytoriów posiada swoje wady i zalety. Złożoność systemów baz biologicznych znaczącą komplikuje administrację, w tym przede wszystkim aktualizację lokalnych zasobów. Zaletą natomiast jest swoboda generowania dowolnych zapytań i szybkość operacji zależna od własnego sprzętu. Duże zasoby udostępniają usługę zdalną API. Dostępne systemy obsługi zdalnej baz danych to: bioMart zainstalowany w Ensembl lub *Entrez Programming Utilities* (E-utilities) w NCBI.

Określenie relacji pomiędzy cząsteczkami miRNA. Relacje w obrębie podzbiorów miRNAs o tym samym *seed*, między miRNA o podobnych 15'tkach, ale różnych *seeds* wydaje się być istotna ze względu nie tylko na próbę odnalezienia ukrytych wzorców w sekwencjach, ale także ze względu na konsolidację i dopasowanie globalne na użytek organizmu.

Ekspresja cząsteczek miRNAs może być konstytutywnie lub przestrzennie i temporalnie regulowana. Oprócz analizy zmienności ekspresji istotna wydaje się być analiza poziomu ekspresji, która pozwoli wyłonić ten człon konstytutywnie produkowanych miRNAs. Relacje wiele do wielu między zbiorem miRNAs i targetami rysują sieć wzajemnych powiązań, a to oznacza, że analiza targetów powinna uwzględniać cały profil ekspresji zamiast założenia wzrostu ekspresji pojedynczego miRNA, jako odpowiedzi na transfekcję. Ta sieć relacji wpisuje się w całościową sieć regulacji genów. Wynika to np. z koekspresji sąsiadujących miRNAs i genów gospodarza [8].

Technika mikromacierzowa posiada ograniczenia. Metody mikromacierzowe nie wyłapują przejściowych i małych ilości par miRNA/mRNA. Nie rozróżniają transkryptów pseudogenów,

które też mogą stanowić geny miRNA. Ale te zagadnienia wykraczają już poza zamierzony zakres niniejszej pracy.

Przedstawione w pracy przegląd aktualnej wiedzy, wyniki analizy i wyciągnięte z nich wnioski potwierdzają prawdziwość sformułowanej tezy doktoratu. Zatem autor niniejszego opracowania uznaje, że cel doktoratu został osiągnięty.

7.2 Plan dalszych prac

Przeprowadzone w ramach niniejszej dysertacji analiza literaturowa i zawarte w niej rozwinięcie modelu predykcji targetów – podpowiadają kierunki dalszej pracy naukowej. Pojawiające się w toku rozprawy pytania oscylują wokół dwóch biegunów. Pierwszy dotyczy konsekwentnego rozwinięcia i parametryzacji zaproponowanego w pracy biocybernetycznego modelu **biTargetScore**. W tym celu niezbędne jest jednak pozyskanie odpowiednich danych eksperymentalnych: skorelowanych ze sobą poziomów ekspresji zbiorów miRNAs. Najlepiej, aby zmienność poziomów ekspresji miRNA i odpowiednich targetów dotyczyła tych par, które występują jednocześnie w dostępnych zasobach potwierdzonych eksperymentalnie targetów. W takiej sytuacji możliwa byłaby weryfikacja rodzaju korelacji między poziomem ekspresji miRNAs i ich targetów.

Dalsze rozwinięcie modelu i tym samym polepszenie jakości predykcji, związane jest z próbą eliminacji błędu jaki "popełnia" proponowany model, zakładając brak korelacji wzajemnej między poziomami ekspresji w zbiorze miRNAs. Jak wiadomo, geny części miRNAs występują w tak zwanych skupiskach, zlokalizowanych w tych samych regionach chromosomalnych. Sugeruje to wzajemny transkrypt obejmujący ten region i oznacza, że ten sam czynnik transkrypcyjny może mieć równoczesny wpływ na poziom ekspresji pri-miRNA. Również można się spodziewać, że jeden blokowany target w RNAi może modulować poziom transkrypcji czy to genów kodujących czy wręcz genów miRNA. Wszystkie te wymienione zastrzeżenia można uwzględnić w biocybernetycznym modelu, jeśli wykorzystamy stosowne zasoby systemowe posiadające informację o korelacjach i relacjach między genami.

Oddziaływanie jednej cząsteczki miRNA na wiele różnych targetów sugeruje podejście systemowe, uwzględniające **sieciowe** relacje miRNA i targetów. Cząsteczki miRNAs regulują sieci komórkowe, jako sieciowy komponent w wielu funkcjach komórkowych. Efekt działania miRNAs na ekspresję ich targetów w rzeczywistości wykazuje różnorodność i na tej podstawie interakcja miRNA-mRNA została skategoryzowane w trzy klasy [180][45]:

1. Cząsteczka miRNA może obniżać poziom targetów w niekonsekwentny sposób, działając jak przełącznik *switch*. Pomaga on w rozwoju i podtrzymywaniu przeznaczenia komórek.
2. Cząsteczka miRNA pomaga utrzymać średni poziom ekspresji jego targetów na niższym poziomie: *fine-tuner* precyzyjny regulator poziomu ekspresji.
3. Cząsteczka miRNA może redukować wariancje ekspresji targetów niwelując fluktuacje w ekspresji genów.

Cząsteczka miRNA, która ma większą liczbę aktualnych swoich targetów w komórce hamuje każdy swój target w mniejszym stopniu, niż gdyby miała tylko jeden. Nazwano to *dilution effect* [3]). Scharakteryzowano także nową klasę funkcyjnych cząsteczek *competitive endogenous RNAs* (ceRNAs), które komplikują ilościową charakterystykę regulacji miRNA [146].

Drugi biegun zainteresowań dotyczy interpretacji i wyjaśniania zwierzęcego mechanizmu RNAi, który zdaniem autora niniejszej pracy, jest bardziej złożony i stanowi ewolucyjne udoskonalenie układu immunologicznego u roślin. Jest to teza "przebijająca" się u niektórych autorów np. analizujących ten mechanizm w wirusowych ekspozycjach komórek gospodarza. Brak postępu i trudności ustalenia tak precyzyjnego i hipotetycznego mechanizmu może być związane ze wspominanym już poziomem oraz skalą rozpatrywania obiektu badania.

Do tej pory udało się zaobserwować podział funkcjonalny cząsteczki miRNA. Wyróżniono pierwsze około ***7nt,*** nazywając je *seed*. Jest to region prawie całkowicie komplementarny do miejsca wiązania na transkrypcie, determinujący wiązanie się z określonym regionem targetu. Precyzyjnie należałoby napisać poprzednie zdanie w liczbie mnogiej "targetów", ze względu na małą specyficzność tak krótkiej sekwencji. Natomiast pozostały fragment cząsteczki miRNA (15-stka) stwarza najwięcej problemów w biochemicznej interpretacji. W jego obrębie mieści się uzasadnienie stabilizacji czy współudział w tworzonym dupleksie tej części cząsteczki.

Tezę o funkcjonalności 15-stki wspierają:

- ewolucyjność (a nie inwolucyjność) mechanizmu RNAi, obserwowana na przykład przy porównaniu RNAi roślinnego i zwierzęcego;
- przesłanki racjonalne - na podstawie zestawienia tego mechanizmu z procesami sterowania produkcją lub rozwiązaniami technicznymi stosowanymi w teorii kodowania i informacji;
- próba wyjaśnienia zestawionych poniżej czynników.

Czynniki, na jakie warto zwrócić uwagę, związane są z długością cząsteczek miRNAs. Jest ona zbieżna, określona lub zdeterminowana przez:

1. szerokość kieszeni białkowej Ago w RISC'u;
2. dwa skoki podwójnej helisy – ok 21nt;
3. stanowi minimalną długość słowa charakterystyczną dla gatunku;
4. 7 nt jest minimalną długością *seed*;
5. sekwencja 15-stki, ponieważ nie jest komplementarna względem miejsc wiązania, więc może posiadać cechy kodującej sekwencji, która odpowiada peptydowi o długości 5 aminokwasów.

Zbieżność średniej długości miRNAs z podwójnym skokiem helisy DNA może nie być przypadkowy. W mechanizmach regulacji genów przez czynniki transkrypcyjne znana jest okresowość odległości miejsca wiązania i jego wpływie na efektywność transkrypcji. Również w mechanizmach regulacji i kontroli jakości transkryptów, długość ta jak można przypuszczać może być minimalną długością charakterystyczną czy reprezentatywną dla konkretnego genu czy grupy genów.

W immunologii bezpośrednio rozpoznawaną strukturą przez przeciwciało, a dokładniej przez liniowy epitop, stanowi fragment o długości ok 5 aminokwasów. Antygen najczęściej jest dużo większy i zostaje on związany przez dopasowane ramiona Fab przeciwciała. Z kolei częściej występujące są tzw. epitopy konformacyjne rozpoznające konkretną, trójwymiarową strukturę, który można traktować, jako skupisko kilku liniowych fragmentów sekwencji aminokwasowych, które rozpoznają region w strukturze sfałdowanego antygenu.

Już te kilka spostrzeżeń generują pulę bioinformatycznych zadań do realizacji. Dobór odpowiednich genomów wirusowych, oszacowanie podobieństw ludzkich miRNA i genów

wirusowych, rozwiniecie badań w dziedzinie konserwatywności miRNAs: te same miejsca wiązań, ale różne 15-stki.

Badanie konserwatywności miejsc wiązań w przekroju różnych organizmów rozpatruje tylko region *seed*, który się wiąże ze regionem targetu – *site*. Oznacza to, że region sąsiadujący z miejscem wiązania będzie różny, ale czy też będzie konserwatywny?

Najistotniejszą luką opisu mechanizmu RNAi stanowi z punktu widzenia biochemicznego niestabilność wiązania się pary miRNA/target spowodowane częściową komplementarnością sekwencji obu cząsteczek. Niestabilność stanowi z jednej strony możliwość regulacyjną – tutaj model kontroli jakości transkryptów, ale z drugiej nie jest znany mechanizm stabilizujący to wiązanie.

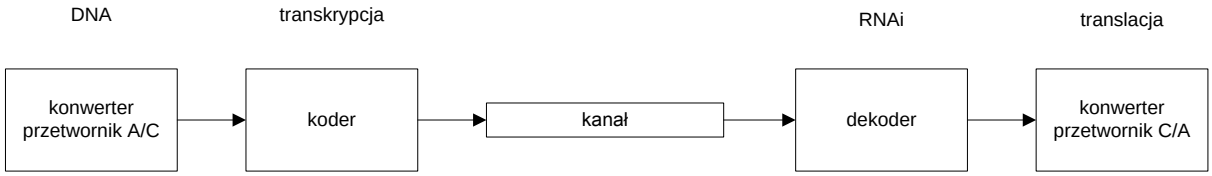
Przeprowadzone badania częstości homologów miRNAs w zbiorze transkryptów może zostać dalej rozwinięte w kierunku analizy częstości dla wydzielonych grup miRNAs. W rozprawie przedstawiono wiele różnych kryteriów różnicowania miRNAs. Są to:

1. podział zbioru miRNAs ze względu na geny (introniczne/egzoniczne/własne geny) (patrz Tabela 2);
2. rodziny miRNA wynikające ze wspólnego *seed* (TargetScan – rozdział 3.2.3) lub z podobieństwa (miRBase – rozdział 3.2.1)
3. konserwatywne i niekonserwatywne miRNAs (TargetScan – rozdział 3.2.3)

Wspominane już ograniczenia modelowania wynikające m.in. z przyjętego poziomu czy skali rozpatrywania zjawiska same w sobie także stają się inspiracją dalszych badań. Podobieństwo między schematem przepływu informacji biologicznej (centralny dogmat biologii molekularnej) a schematem kanału przepływu informacji w telekomunikacji sugeruje wykorzystanie w modelowaniu także innych metod rozwiniętych w teorii kodowania. Hipoteza o związku między degeneracją kodu genetycznego i jego odpornością na zakłócenia powstała spontanicznie od czasu odkrycia kodu genetycznego, ale opierając się na teorii kodowania. Nadmiarowość kodu genetycznego uznano, jako naturalny sposób ochrony przesyłanej informacji. W tym właśnie kierunku poszły prace Sergey Petoukhov (С. В. Петухов). Po wnikliwej analizie kodu genetycznego, a więc podstawowej własności kodujących sekwencji nukleotydowych stwierdził on, że właściwości kodu genetycznego są skorelowane z właściwościami kwasów nukleinowych. Jego osiągnięcie dotyczy formalnego opisu degeneracji kodu genetycznego. Wykorzystując metody algebry macierzy i teorii kodowania sygnałów wykazał zbieżność organizacji kodonów z kodowaniem korekcyjnym opartym na macierzy Hadamarda. Autor podał procedurę quasi algebraiczną, która doprowadziła go do tego odkrycia. Składa się ona z rekurencyjnej metody, która prowadzi do uzyskania macierzy genetycznych multipletów (genomacierzy), następnie jej konwersji oryginalną metoda autora, w wyniku której uzyskuje się prezentację multipletów w postaci macierzy Hadamarda [76].

Prawdopodobieństwo przypadkowego uzyskania przez naturę macierzy takiej organizacji tripletów jest bardzo małe i wynika z dużej liczby 64! istniejących wariantów ułożenia 64 tripletów w macierzy o wymiarze 8 x 8. Idąc dalej tym tropem możemy się spodziewać występowania w komórce biologicznej koderów i dekoderów wykorzystujących tą własność tripletów. Dla kanału komunikacyjnego, jednokierunkowego można zastosować kod korekcji błędu, który nie wymaga kanału zwrotnego (Rys. 7.1). Rozpoznanie błędu, jego detekcja jest znacznie prostsza niż realizacja systemu korekcji błędu. Detekcja wymaga skromniejszego zaplecza. Niemniej brak korekty błędów

oznacza w systemach technicznych retransmisję – transmisja redundantnej informacji wzrasta, gdy występują błędy [133]. Badany mechanizm RNAi przypomina układ kontroli jakości sekwencji transkryptu, być może zintegrowany z procesem translacji RNA, za czym przemawia jego immunologiczno-ewolucyjne pochodzenie. Hipoteza ta wydaje się wiarygodna głównie ze względu na usytuowanie się kompleksu miRISC blisko końca 3' transkryptu. Podobnie w przesyłanych ramkach transmisji na końcu znajduje się bit parzystości lub suma kontrolna w długich łańcuchach znaków. Funkcję dekodera informacji genetycznej wówczas pełniłby mechanizm degradacji transkryptów. Informacja uzyskana z tego dekodera, czyli wyłapane przez RNAi zakłócenie mogłyby być wykorzystane przez mechanizm fałdowania białka dążący do uzyskania prawidłowej struktury cząsteczki aminokwasowej pomimo występującego zaburzenia sekwencji transkryptu. Druga refleksja dotyczy przejścia cyfrowego – 4 literowego sygnału na postać analogową. Ten swoisty przetwornik A/C potrzebny jest wtedy, kiedy sekwencję nukleotydową transkryptu (cyfrową) konwertujemy na postać analogową, czyli przestrzenną strukturę białka, która w tego typu regulacji musi posiadać też postać wirtualną.

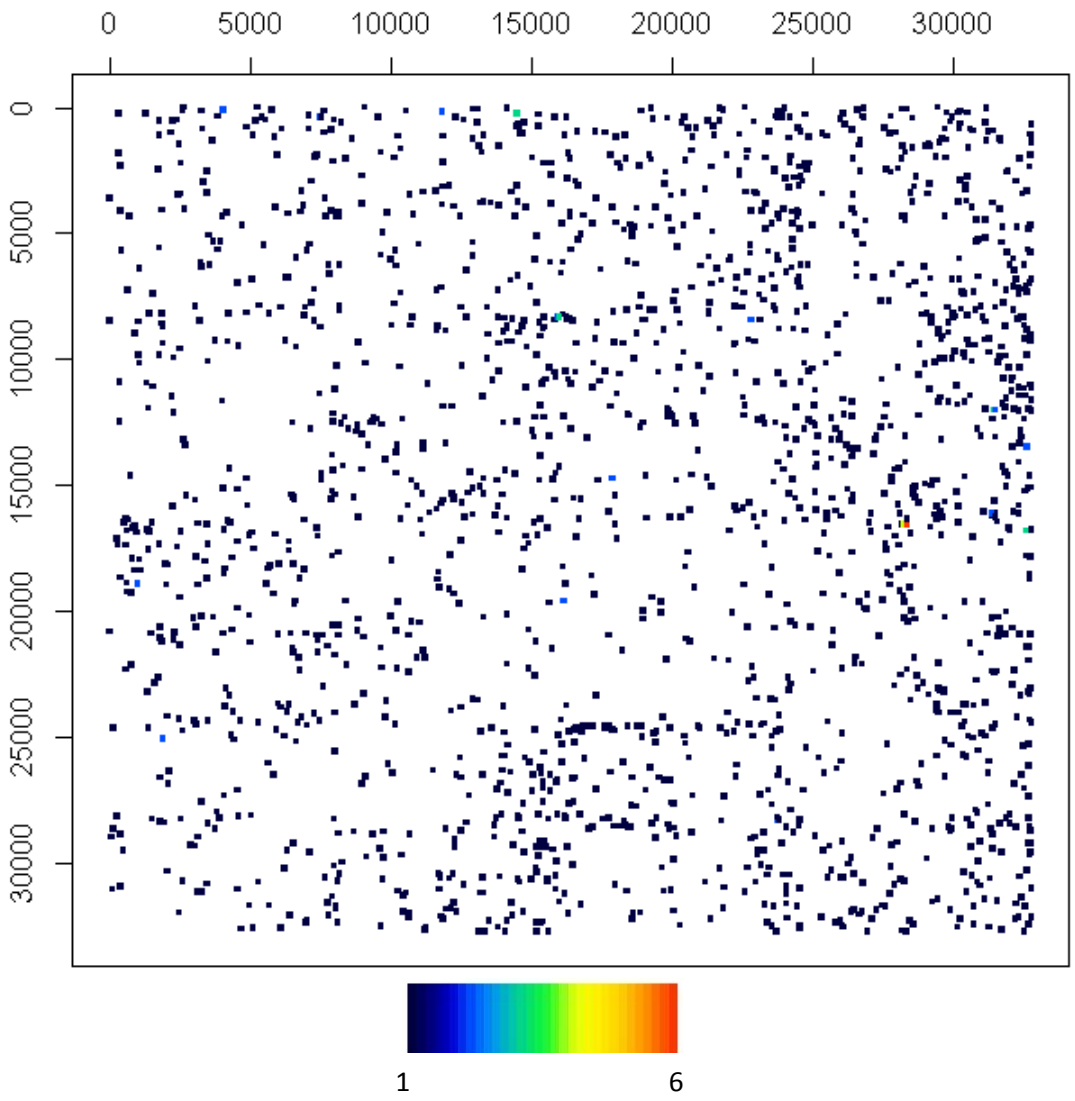


Rys. 7.1. Schemat typowego układu transmisji danych z wykorzystaniem korekcji błędu.

Postać cyfrową informacji (sygnał) stanowią symbole sekwencji nukleotydowej. Źródło informacji (sygnału) przyjmujemy, jako wirtualny, poprawny, bezbłędny zapis nukleotydowy genu. Jest to zapis czysto hipotetyczny. Sekwencja nukleotydowa w chromosomach stanowi już jego zakłócenia poprzez różne czynniki mutagenne. Oznacza to, że sygnał, źródło informacji nie jest nam znane, ale pomimo tego komórka włącza mechanizmy regulacyjno-kontrolne, chociaż nie dysponuje tym poprawnym, nukleotydowym zapisem. Wykorzystuje prawdopodobnie zapis wirtualny. Znamy jedynie jego zniekształcenie w postaci genomowej sekwencji. Realizacja kodera jest albo wynikiem losowej zmienności zapisu w genomie, albo pewną optymalizacją regulacyjną.

Autor niniejszego opracowania zadaje sobie sprawę, że propozycje i hipotezy teraz poruszane wymagają analizy na poziomie kwantowym, który wg obecnej wiedzy wykorzystuje, jako jeden z podstawowych elementów budulcowy komputera kwantowego, bramkę Hadamarda.

Genomacierze zdefiniowane przez Petoukhova są ciekawym podejściem do oceny całego zbioru sekwencji. Umożliwiają one naoczną obserwację relacji między sekwencjami i poszukiwanie elementów symetrii. Ich wadą jest to, że długość sekwencji narzuca odpowiedni rząd macierzy. Macierzy trudnej do percepcji wzrokowej dla już nawet relatywnie krótkich sekwencji. Ograniczeniem jest także wymaganie jednakowej długości porównywanych sekwencji. Dla długości sekwencji k=15nt wymiar macierzy wynosi: 32768 x 32768. Na Rys. 7.2 przedstawiono cały zbiór sekwencji miRNAs 15tek od strony 3'.



Rys. 7.2. Genomacierz miRNAs dla kierunku 3'->5' dla ostatnich 15 nukleotydów. Każdy punkt na macierzy reprezentuje daną sekwencję miRNA. Skala pod rysunkiem kolorami wyróżnia punkty, które skupiają więcej niż jedną sekwencję różnych miRNAs.

7.3 Podsumowanie

Odkrycie funkcyjnych cząsteczek RNA – miRNAs oraz wyjaśnienie podstawowego mechanizmu ich funkcjonalności - interferencji RNA umożliwiało powstanie odpowiednich narzędzi bioinformatycznych do predykcji targetów, które to narzędzia z kolei usprawniają procedury weryfikacji i walidacji eksperymentalnej wytypowanych targetów. Rozpoznanie rzeczywistych targetów *in situ* stanowi podstawę poznania i pełnego zrozumienia mechanizmu RNAi. Zaprezentowane w pracy podejście ilościowego wyjaśnienia mechanizmu regulacji genów, jakie realizuje przedstawiony model **biTargetScore** predykcji targetów, nie wyklucza istnienia różnych, ale jakościowych interpretacji opisywanego zjawiska. W pracy zrealizowano implementacje biocybernetycznego modelu **biTargetScore**, a następnie przeprowadzono jego walidację.

Zaproponowany w pracy model **biTargetScore** został porównany z innymi narzędziami predykcji, wykazując nad nimi przewagę, wyrażającą się przez poprawę jakości rozpoznania targetów. Do weryfikacji poprawności budowanych modeli wykorzystano zasób danych, który został potwierdzony eksperymentalnie.

W pracy zwrócono uwagę na istotność założeń modelu jakościowego opisującego mechanizm RNAi, na sposób przeprowadzenia wnioskowania z uzyskanych wyników predykcji. W pracy podjęto próbę wyjaśnienia przyczyn i wskazania trudności przetwarzania danych surowych, pochodzących z eksperymentów mikromacierzowych, które charakteryzuje stochastyczność. Wieloetapowość ich statystycznego przetwarzania wymaga staranności i świadomości stosowanych metod uproszczeń czy uogólnień. Analiza tych uproszczeń i uogólnień została przedstawiona w pracy.

Złożoność funkcjonowania organizmu żywego polega między innymi na jego zdolnościach adaptacyjnych i regulacyjnych. Logiczną konsekwencją zaburzenia stanu, w którym się organizm znajdował, jest automatyczne, samoistne uruchomienie mechanizmów korygujących tą dysfunkcję - zakłócenie. W świetle tych rozważań stan patologii staje się o wiele bardziej złożonym mechanizmem, ze względu na występowanie w nim dodatkowych, zależnych od siebie, hierarchicznych układów regulacyjnych.

Modelowanie biocybernetyczne stawiając na pierwszym miejscu **regulację**, wymaga postawienia nadrzędnego celu. Nie można przyjąć, że zapis nukleotydowy w DNA stanowi wzorzec, najwierniejszy jego zapis. Cel jest wirtualny i rozproszony pomiędzy wyspecjalizowane odłamy, jakie powstały w ewolucji. Kiedyś, w prehistorii, był bezpośrednio związany z obiektem. Każda niepożądana mutacja w sekwencji DNA wywołuje uruchomienie mechanizmów regulacji, których celem jest dążenie do homeostazy. Dzieje się tak nawet wtedy, gdy nie ma poprawnego wzorca w bibliotece matryc sekwencji aminokwasowych. Przedstawiana praca rzuca nowe światło na rozważane zagadnienia poprzez użycie modelu biocybernetycznego.

8 Piśmiennictwo

[1] Andronescu, M.; Bereg, V.; Hoos, H.H. et al (2008). 'RNA STRAND: the RNA secondary structure and statistical analysis database'. BMC Bioinformatics 9:340.

[2] Aravind, L.; Watanabe, H.; Lipman, D.J.; Koonin, E.V. (2000) 'Lineage-specific loss and divergence of functionally linked genes in eukaryotes.', Proceedings of the National Academy of Sciences 97 (21): 11319–11324.

[3] Arvey, A.; Larsson, E.; Sander, C.; Leslie, C.S. & Marks, D.S. (2010). 'Target mRNA abundance dilutes microRNA and siRNA activity'. Mol. Syst. Biol., 6, 215–233.

[4] Asangani, I.A.; Rasheed, S.A.; Nikolova, D.A.; Leupold, J.H.; Colburn, N.H.; Post, S.; Allgayer, H. (2008). 'MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer.' Oncogene 27(15):2128-36.

[5] Bagga, S., et al. (2005). 'Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation'. Cell 122, 553–563.

[6] Ball, P. (2013), 'DNA: Celebrate the unknowns', Nature 496, 419–420.

[7] Bartel, D.P. (2009). 'MicroRNAs: target recognition and regulatory functions'. Cel , 136:215-233.

[8] Baskerville, Scott & Bartel, David P. (2005). 'Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes'. RNA 11:241–247.

[9] Bayes, Thomas (1764). ‘Essay towards solving a problem in the doctrine of chances’, the Philosophical Transactions of the Royal Society of London.

[10] Benne, R. (1994), 'RNA editing in trypanosomes.', Eur. J. Biochem 221 (1): 9–23.

[11] Bentwich, I. (2005). 'Prediction and validation of microRNAs and their targets'. FEBS Lett 579:5904–5910.

[12] Berger, John A.; Hautaniemi, Sampsa; Järvinen, Anna-Kaarina; Edgren, Henrik; Mitra, Sanjit K. & Astola, Jaakko (2004). 'Optimized LOWESS normalization parameter selection for DNA microarray data'. BMC Bioinformatics 5:194.

[13] Berkhout, B.; Haasnoot, J. (2006), 'The interplay between virus infection and the cellular RNA interference machinery'. FEBS Lett 580 (12):2896–902.

[14] Berman, H.M.; Olson, W.K.; Beveridge, D.L. et al (1992). 'The nucleic acid database. A comprehensive relational database of threedimensional structures of nucleic acids'. Biophys J 63:751–759.

[15] Bishop, C. M. (2006), 'Pattern recognition and machine learning'. Springer - Information Science and Statistics, NY, USA.

[16] Bland, J.M.; Altman, D.G. (1986). 'Statistical methods for assessing agreement between two methods of clinical measurement'. Lancet 327 (8476): 307–10.

[17] Blevins, T.; Rajeswaran, R.; Shivaprasad, P.; Beknazariants, D.; Si-Ammour, A.; Park, H.; Vazquez, F.; Robertson, D.; Meins, F.; Hohn, T.; Pooggin, M.; (2006). 'Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing'. Nucleic Acids Res 34 (21): 6233–46.

[18] Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. (2003). 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias'. Bioinformatics 19 (2): 185–193.

[19] Brennecke, J.; Stark, A.; Russell, R.B. & Cohen, S.M. (2005). 'Principles of microRNA-target recognition'. PLoS Biol. 3: e85.

[20] Buchon, N.; Vaury, C., (2006), 'RNAi: a defensive RNA-silencing against viruses and transposable elements'. Heredity 96 (2): 195–202.

[21] Bueno, M.J.; de Castro, I.P.; Malumbres, M. (2008), 'Control of cell proliferation pathways by microRNAs'. Cell Cycle 7:3143–3148.

[22] Burnette, W.N. (1981). 'Western blotting': electrophoretic transfer of proteins from sodium dodecyl sulfate--polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A.'. Anal Biochem. 112(2):195-203.

[23] Cambronne, X.A.; Shen, R.; Auer, P.L.; Goodman, R.H. (2012). 'Capturing microRNA targets using an RNA-induced silencing complex (RISC)-trap approach'. Proc Natl Acad Sci 109(50):20473-8.

[24] Carrington, J.; Ambros, V. (2003). 'Role of microRNAs in plant and animal development'. Science 301 (5631): 336–8.

[25] Cerutti, H.; Casas-Mollano, J. (2006). 'On the origin and functions of RNA-mediated silencing: from protists to man'. Curr Genet 50 (2): 81–99.

[26] Chalfie, M.; Tu, Y.; Euskirchen, G.; Ward, W.W.; Prasher, D.C. (1994). 'Green fluorescent protein as a marker for gene expression'. Science 263(5148):802-5.

[27] Chen, C.; Ridzon, D.A.; Broomer, A.J. et al (2005). 'Real-time quantifi cation of microRNAs by stemloop RT-PCR'. Nucleic Acids Res 33(20):e179

[28] Cole, K.; Truong, V.; Barone, D.; McGall, G. (2004). 'Direct labeling of RNA with multiple biotins allows sensitive expression profiling of acute leukemia class predictor genes'.Nucleic Acids Res. 32e86.

[29] Collins, M.L.; Irvine, B.; Tyner, D.; Fine, E.; Zayati, C.; Chang, C.; Horn, T.; Ahle, D.; Detmer, J.; Shen, L.P.; Kolberg, J.; Bushnell, S.; Urdea, M.S.; Ho, D.D. (1997). 'A branched DNA signal amplification assay for quantification of nucleic acid targets below 100 molecules/ml'. Nucleic Acids Res. 25(15):2979-84.

[30] Coronnello, C.; Benos, P.V. (2013). 'ComiR: Combinatorial microRNA target prediction tool'. Nucl Acids Res 41 (Web Server issue): W159-64.

[31] Coronnello, C.; Hartmaier, R.; Arora, A.; Huleihel, L.; Pandit, K.V.; Bais, A.S.; Butterworth, M.; Kaminski, N.; Stormo, G.D.; Oesterreich, S.; Benos, P.V. (2012). 'Novel modeling of combinatorial miRNA targeting identifies SNP with potential role in bone density'. PLoS Comput Bio 8(12).

[32] Crick, F.H.C. (1956), 'On Protein Synthesis.', Symp. Soc. Exp. Biol. XII, 139-163.

[33] Cullen, B. (2006). 'Is RNA interference involved in intrinsic antiviral immunity in mammals?'. Nat Immunol 7 (6): 563–7.

[34] Cullen, B.R. (2004) 'Transcription and processing of human microRNA precursors'. Mol. Cell 16, 861–865.

[35] Cullen, Bryan R. (2006), 'Viruses and microRNAs',Nature Genetics 38, S25 - S30.

[36] Davidson, E.; Levin, M. (2005). 'Gene regulatory networks'. Proc. Natl. Acad. Sci. U.S.A. 102 (14): 4935.

[37] De Moivre, Abraham (1718). 'The Doctrine of Chances: a method of calculating the probabilities of events in play'.

[38] de Wet, J.R.; Wood, K.V.; Helinski, D.R.; DeLuca, M. (1985). 'Cloning of firefly luciferase cDNA and the expression of active luciferase in Escherichia coli'. Proc Natl Acad Sci 82(23):7870-3.

[39] Delongchamp R.R.; Bowyer, J.F.; Chen, J.J. et al (2004). 'Multiple-testing strategy for analyzing cDNA array data on gene expression'. Biometrics 60(3):774–782.

[40] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). 'Maximum Likelihood from Incomplete Data via the EM Algorithm'. Journal of the Royal Statistical Society, Series B 39 (1): 1–38.

[41] Didiano, D.; Hobert, O. (2006). 'Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions.' Nat Struct Mol Biol. 13(9):849-51.

[42] Djuranovic, S.; Nahvi, A.; Green, R. (2012), 'miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay'. Science 336 (6078): 237–40.

[43] Doench, J.G.; Sharp, P.A. (2004). 'Specificity of microRNA target selection in translational repression'. Genes Dev 18:504-511.

[44] Doma, M. K.; Parker R. (2006), 'Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation.', Nature 440, 561–564.

[45] Ebert, M.S.; Sharp, P.A. (2012). 'Roles for microRNAs in conferring robustness to biological processes'. Cell. 149(3):515-24.

[46] Edwards, D. (2003). 'Non-linear normalization and background correction in onechannel cDNA microarray studies'. Bioinformatics,19, 825–833.

[47] Elmen, J.; Lindow, M.; Silahtaroglu, A.; Bak, M.; Christensen, M.; Lind-Thomsen, A.; Hedtjarn, M.; Hansen, J.B.; Hansen, H.F.; Straarup, E.M.; McCullagh, K.; Kearney, P.; Kauppinen, S. (2008). 'Antagonism of microRNA-122 in mice by systemically administered LNA-antimiR leads to up-regulation of a large set of predicted target mRNAs in the liver'. Nucleic Acids Res 36:1153-1162.

[48] Enright, A.J.; John, B.; Gaul, U.; Tuschl, T.; Sander, C.; Marks, A.J. (2003). 'MicroRNA targets in Drosophila'. Genome Biology 5(1):R1.

[49] Erson-Bensan, A.E. (2014). 'Introduction to MicroRNAs in Biological Systems'. in: Yousef, M.; Allmer, J. (eds.) miRNomics: MicroRNA Biology and Computational Analysis. Methods in Molecular Biology 1107, Humana Press.

[50] Eulalio, A.; Huntzinger, E.; Nishihara, T.; Rehwinkel, J.; Fauser, M.; Izaurralde, E. (2009), 'Deadenylation is a widespread effect of miRNA regulation'. RNA 15 (1): 21–32.

[51] EXIQON http://www.exiqon.com

[52] EXIQON, Guidelines for setting up microRNA profiling experiments v2.0 December 2010 www.exiqon.com.

[53] Fagin, R.; Kumar, R. & Sivakumar, D. (2003). 'Comparing top k lists'. SIAM Journal on DiscreteMathematics, vol. 17, no. 1, pp.134–160.

[54] Fahlgren, N.; Jogdeo, S.; Kasschau, K.D. et al (2010). 'MicroRNA gene evolution in arabidopsis lyrata and arabidopsis thaliana'. Plant Cell 22:1074–1089

[55] Farh, K.K. et al. (2005) 'The widespread impact of mammalian microRNAs on mRNA repression and evolution'. Science 310, 1817–1821.

[56] Frank, F.; Sonenberg, N.; Nagar, B.(2010), 'Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2.', Nature, 465(7299):818-22.

[57] Friedländer, M.R. et al. (2014), 'Evidence for the biogenesis of more than 1,000 novel human microRNAs'. Genome Biology, 15 :R57.

[58] Friedman, R.C. et al. 'Supplemental Material. Most Mammalian mRNAs Are Conserved Targets of MicroRNAs'.in: Friedman, Robin C.; Farh, Kyle Kai-How; Burge, Christopher B.; Bartel, David P. Genome Research, 19:92-105 (2009).

[59] Friedman, R.C.;, Farh, K.K.; Burge, C.B. Bartel, D.P. (2009). 'Most mammalian mRNAs are conserved targets of microRNAs'. Genome Res. 19(1):92-105.

[60] Fryxell, K.J.;, Moon, W.J. (2005). 'CpG mutation rates in the human genome are highly dependent on local GC content'. Mol Biol Evol. 22(3):650-8.

[61] Fukuda, Y.; Kawasaki, H.; Taira, K. (2005). 'Exploration of human miRNA target genes in neuronal differentiation'. Nucleic Acids Symp Ser (Oxf):341-342.

[62] García, David M.; Baek, Daehyun; Shin, Chanseok; Bell, George W.; Grimson, Andrew; Bartel, David P. (2011). 'Weak Seed-Pairing Stability and High Target-Site Abundance Decrease the Proficiency of lsy-6 and Other miRNAs'. Nat Struct Mol Biol., 18:1139-1146.

[63] Gawroński, R. (1983). 'Biocybernetyka' in:Encyklopedia Fizyki Współczesnej PWN Warszawa, s. 779-790.

[64] Git, A.; Dvinge, H.; Salmon-Divon, M. et al (2010) 'Systematic comparison of microarray profi ling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression'. RNA 16(5):991–1006

[65] Gökmen, Z. & Erdal, C. (2014). 'Introduction to Statistical Methods for MicroRNA Analysis'. in: Yousef, M.; Allmer, J. (eds.), miRNomics: MicroRNA Biology and Computational Analysis. Methods in Molecular Biology 1107., Humana Press 2014.

[66] Greniewski, H. (1969), 'Cybernetyka niematematyczna'. Warszawa.

[67] Griffiths-Jones, S. (2004). 'The microRNA Registry'. Nucleic Acids Res. 32:D109-D111.

[68] Griffiths-Jones, S. (2010). 'miRBase: microRNA sequences and annotation'. Current Protoc Bioinform Chapter 12: Unit 12.9.1–10

[69] Griffiths-Jones, S.; Grocock, R.J.; van Dongen, S.; Bateman, A.; Enright, A.J. (2006). 'miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34:D140-D144.

[70] Griffiths-Jones, S.; Saini, H.K.; van Dongen, S.; Enright, A.J. (2008). 'miRBase: tools for microRNA genomics'. Nucleic Acids Res. 36:D154-D158.

[71] Grimson, Andrew; Farh, Kyle Kai-How; Johnston, Wendy K.; Garrett-Engele, Philip; Lim, Lee P.; Bartel, David P. (2007). 'MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing'. Molecular Cell, 27:91-105.

[72] Grosshans, H.; Filipowicz, W. (2008). 'Proteomics joins the search for microRNA targets'. Cell. 22;134(4):560-2.

[73] Haley, B.; Zamore, B. (2004). 'Kinetic analysis of the RNAi enzyme complex.' Nature Structural & Molecular Biology 11 (7): 599–606.

[74] Hamilton, A.; Baulcombe, D. (1999), 'A species of small antisense RNA in posttranscriptional gene silencing in plants.', Science 286 (5441) 29 October: Vol. 286. no. 5441, pp. 950 – 952.

[75] Hammond, S.M.; Bernstein, E.; Beach, D.; Hannon, G.J. (2000). 'An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells.', Nature 404 (6775): 293–296.

[76] He, M. & Petoukhov, S., (2011), 'Mathematics of Bioinformatics: Theory, Practice, and Applications'. John Wiley & Sons, p. 180-228.

[77] Higgs, P.H.; Attwood T.K. (2004), 'Bioinformatics and Molecular Evolution', Wiley-Blackwell.

[78] Hsu, S.D.; Tseng, Y.T.; Shrestha, S.; Lin, Y.L.; Khaleel, A.; Chou, C.H.; Chu, C.F.; Huang, H.Y.; Lin, C.M.; Ho, S.Y.; Jian, T.Y.; Lin, F.M.; Chang, T.H.; Weng, S.L.; Liao, K.W.; Liao, I.E.; Liu, C.C.; Huang, H.D. (2014). 'miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions'. Nucleic Acids Res.42(Database issue):D78-85.

[79] Huang, J.C.; Babak, T.; Corson, T.W.; Chua, G., Khan, S.; Gallie, B.L.; Hughes, T.R.; Blencowe, B.J.; Frey, B.J. & Morris, Q.D. (2007). 'Using expression profiling data to identify human microRNA targets'. Nature Methods, 4, 215–233.

[80] Huber, W.; von Heydebreck, A.; Sültmann, H.; Poustka, A.; Vingron, M. (2002). 'Variance stabilization applied to microarray data calibration and to the quantification of differential expression'. Bioinformatics 18 suppl. 1, S96-S104.

[81] International Structural Genomics Organization, ISGO http://www.isgo.org

[82] Irizarry, R.A. etal (2003). 'Exploration, normalization, and summaries of high density oligonucleotide array probe level data'. Biostatistics, 4, 249–264.

[83] Jacobson, A.; Peltz, S. W. (1996), 'Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells.' Annu. Rev. Biochem.65, 693−739.

[84] Jakieła, B.; Gielicz, A.; Plutecka, H.; Hubalewska, M.; Mastalerz, L.; Bochenek, G.; Soja, J.; Januszek, R.; Musial, J.; Sanak, M. (2013), 'Eicosanoid biosynthesis during mucociliary and mucous metaplastic differentiation of bronchial epithelial cells'.Prostaglandins Other Lipid Mediat. 2013 Oct; 106:116-23.

[85] Jing, Q.; Huang, S., Guth, S.; Zarubin, T.; Motoyama, A.; Chen, J.; Di Padova, F.; Lin, S.C.; Gram, H.; Han, J. (2005). 'Involvement of microRNA in AU-rich element-mediated mRNA instability'. Cell 120 (5): 623–34.

[86] Jovanovic, M.; Hengartner, M.O. (2006) 'miRNAs and apoptosis: RNAs to die for'. Oncogene 25:6176–6187.

[87] Kawahara, Y.; Zinshteyn, B.; Sethupathy, P.; Iizasa, H.; Hatzigeorgiou, A.G.; Nishikura, K. (2007). 'Redirection of silencing targets by adenosine-to-inosine editing of miRNAs.' Science 315:1137-1140.

[88] Kertesz, M.; Iovino, N.; Unnerstall, U.; Gaul, U.; Segal, E. (2007). 'The role of site accessibility in microRNA target recognition'. Nat Genet 39:1278-1284.

[89] Kiezun, Adam; Artzi, Shay; Modai, Shira; Volk, Naama; Isakov, Ofer; Shomron, Noam (2012), 'miRviewer: A multispecies microRNA homologous viewer', In BMC Research Notes 2012, 5:92.

[90] Klosterman, P.S.; Hendrix, D.K.; Tamura, M.; Holbrook, S.R.; Brenner, S.E. (2004). 'Three-dimensional motifs from the SCOR, structural classifiation of RNA database: extruded strands,base triples, tetraloops and U-turns'. Nucleic Acids Research, Vol. 32, No. 8, 2342-2352.

[91] Kooperberg, C.; Fazzio, T.G.; Delrow, J. J. and Tsukiyama, T. (2002). 'Improved background correction for spotted DNA microarrays'. Journal of Computational Biology9, 55-66.

[92] Kozomara, A.; Griffiths-Jones, S. (2011). 'miRBase: integrating microRNA annotation and deep-sequencing data'. Nucleic Acids Res. 39:D152-D157.

[93] Kozomara, A.; Griffiths-Jones, S. (2014). 'miRBase: annotating high confidence microRNAs using deep sequencing data'. Nucleic Acids Res. 42:D68-D73.

[94] Krol, J.; Sobczak, K.; Wilczynska, U.; Drath, M.; Jasinska, A.; Kaczynska D.; Krzyzosiak, W.J. (2004), 'Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design.', J Biol Chem 279 (40): 42230–9.

[95] Kullback, B.J., Leibler, R.A (1951). 'On information and sufficiency'. Ann Math Statist 22:79–86.

[96] Laederach, A. (2007). 'Informatics challenges in structured RNA'. Brief Bioinform 8:294–303.

[97] Leaman, D.; Chen, P.Y.; Fak, J.; Yalcin, A.; Pearce, M.; Unnerstall, U.; Marks, D.S.; Sander, C.; Tuschl, T.; Gaul, U. (2005). 'Antisense-mediated depletion reveals essential and specific functions of microRNAs in Drosophila development'. Cell 121:1097-1108.

[98] Lee, R.C.; Ambros, V.(2001), 'An extensive class of small RNAs in Caenorhabditis elegans.' Science. 294(5543):862-4.

[99] Lee, R.C.; Feinbaum, R.L.; Ambros, V. (1993). 'The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14'. Cell75 (5): 843–54.

[100] Lee, Y.; Kim, M.; Han, J.; Yeom, K.H.; Lee, S.; Baek, S.H.; Kim, V.N. (2004). 'MicroRNA genes are transcribed by RNA polymerase II'. EMBO J. 23 (20): 4051–60.

[101] Lee, Y.; Nakahara, K.; Pham, J.; Kim, K.; He, Z.; Sontheimer, E.; Carthew, R. (2004). 'Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways'. Cell 117 (1): 69–81.

[102] Lewis, Benjamin P.; Burge, Christopher B.; Bartel, David P. (2005). 'Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets'. Cell, 120:15-20.

[103] Li, W.; Ruan, K.; (2009). 'MicroRNA detection by microarray'. Anal Bioanal Chem 394(4): 1117–1124.

[104] Li, Y. TargetScoreData: TargetScoreData. R package version 1.3.1 (https://bioconductor.org/packages/).

[105] Li, Yang; Lu, Jinfeng; Han, Yanhong; Fan, Xiaoxu; Ding, Shou-Wei (2013). 'RNA Interference Functions as an Antiviral Immunity Mechanism in Mammals'. Science 342 (6155): 231–234.

[106] Li, Yue; Goldenberg, Anna; Wong, Ka-Chun & Zhang, Zhaolei (2014). 'A probabilistic approach to explore human microRNA targetome using microRNA-overexpression data and sequence information'. Bioinformatics 30 (5): 621-628.

[107] Li, Yue; Liang, Cheng; Wong, Ka-Chun; Jin, Ka-Chun & Zhang, Zhaolei (2014). 'Inferring probabilistic miRNA–mRNA interaction signatures in cancers: a role-switch approach'. Nucleic Acids Res.1.

[108] Liang, Ru-Qiang; Li, Wei; Li, Yang; Tan, Cui-yan; Li, Jian-Xun; Jin, You-Xin & Ruan, Kang-Cheng (2005). 'An oligonucleotide microarray for microRNA expression analysis based on labeling RNA with quantum dot and nanogold probe'. Nucl. AcidsRes. 33(2): e17.

[109] Lim, L.P.; Lau, N.C.; Garrett-Engele, P.; Grimson, A.; Schelter, J.M.; Castle, J.; Bartel, D.P.; Linsley, P.S.; Johnson, J.M. (2005), 'Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs'. Nature 433 (7027): 769–73.

[110] Lin, S.; Ding, J. (2009). 'Integration of Ranked Lists via Cross Entropy Monte Carlo with Applications to mRNA and microRNA Studies'. Biometrics, 65, 9-18.

[111] Liu, C.G.; Calin, G.A.; Volinia, S.; Croce, C.M. et al (2008). 'MicroRNA expression profi ling using microarrays'. Nat Protoc 3(4):563–578.

[112] Lorenz, Ronny; Bernhart, Stephan H.; zu Siederdissen, Christian Höner; Tafer, Haki; Flamm, Christoph; Stadler, Peter F. & Hofacker, Ivo L. (2011). 'ViennaRNA package 2.0'. Algorithms for Molecular Biology, 6(1):26.

[113] Lu, Chao (2004). 'Improving the scaling normalization for high-density oligonucleotide GeneChip expression microarrays'. BMC Bioinformatics5: 103.

[114] Lu,Y.;Zhou,Y.; Qu, W.; Deng, M. & Zhang, C.; (2011). 'A Lasso regression model for the construction of microRNA-target regulatory networks'. Bioinformatics, 27, 215–233.

[115] Lucy, A.; Guo, H.; Li, W.; Ding, S. (2000). 'Suppression of post-transcriptional gene silencing by a plant viral protein localized in the nucleus'. MBO J 19 (7): 1672–80

[116] Lujambio, A.; Ropero, S.; Ballestar, E.; Fraga, M.F.; Cerrato, C.; Setien, F.; Casado, S.; Suarez-Gauthier, A.; Sanchez-Cespedes, M.; Gitt, A.; Spiteri, I.; Das, P.P.; Caldas, C.; Miska, E.; Esteller, M. (2007). 'Genetic Unmasking of an Epigenetically Silenced microRNA in Human Cancer Cells'. Cancer Res 67:1424-1429.

[117] Mack, G.S. (2007), 'MicroRNA gets down to business'. Nature Biotechnology 25, 631 – 638.

[118] Maillard, P. V.; Ciaudo, C.; Marchais, A.; Li, Y.; Jay, F.; Ding, S. W.; Voinnet, O. (2013), 'Antiviral RNA Interference in Mammalian Cells'. Science342 (6155): 235–238.

[119] Malone, C.D. & Hannon, G.J., (2009), 'Small RNAs as Guardians of the Genome'. Cell 136(4): p. 656-668.

[120] Mathelier, A. & Carbone, A. (2013). 'Large scale chromosomal mapping of human microRNA structural clusters'. Nucleic Acids Res 41, 4392–4408.

[121] Mattick, J.S.; Makunin, I.V. (2006), 'Non-coding RNA'. Hum Mol Genet, 15 Spec No 1:R17-R29.

[122] Miranda, K.C.; Huynh, T.; Tay, Y. et al (2006). 'A pattern-based method for the identifi cation of microRNA binding sites and their corresponding heteroduplexes'. Cell 126:1203–1217.

[123] Monod J. (1969), 'On Symetry and Function in Biological Systems'. in: Enstrom, A.; Strandberd, B. (eds.) Nobel Symposium 11: Symetry and Function of Biological System at the Macromolecular Level. Almqvist & Wiksell Forlag AB, Scotckholm, p. 1527.

[124] Morin, R. D.; O'Connor, M. D.; Griffith, M.; Kuchenbauer, F.; Delaney, A.; Prabhu, A. -L.; Zhao, Y.; McDonald, H.; Zeng, T.; Hirst, M.; Eaves, C. J.; Marra, M. A. (2008). 'Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells'. Genome Research 18 (4): 610–621.

[125] Morin, Ryan D.; Bainbridge, Matthew; Fejes, Anthony; Hirst, Martin; Krzywinski, Martin; Pugh, Trevor J.; McDonald, Helen; Varhol, Richard; Jones, Steven J.M. & Marra, Marco A. (2008). 'Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing'. BioTechniques 45 (1): 81–94.

[126] Morozova, N.; Zinovyev, A.; Nonne, N.; Pritchard, L.L.; Gorban, A.N.; Harel-Bellan, A., (2012). 'Kinetic signatures of microRNA modes of action'. RNA 18 (9): 1635–55.

[127] Motameny, S.; Wolters, S.; Nürnberg, P. et al (2010). 'Next generation sequencing of miRNAs – strategies, resources and methods'. Genes 1:70–84.

[128] Nakamoto, M.; Jin, P.; O'Donnell, W.T.; Warren, S.T. (2005). 'Physiological identification of human transcripts translationally regulated by a specific microRNA'. Hum Mol Genet 14:3813-3821.

[129] Nikitin, E. (1975). 'Wyjaśnianie jako funkcja nauki'. tłum. z rosyjskiego, Warszawa: PWN.

[130] Orom, U.A., Nielsen, F.C.; Lund, A.H. (2008). 'MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation'. Mol Cell 30:460–471.

[131] Orom, U.A.; Lund, A.H. (2009). 'Experimental identification of microRNA targets'. Gene 451:1–515.

[132] Pasquinelli, A.E.; Reinhart, B.J.; Slack, F.; Martindale, M.Q.; Kuroda, M.I.; Maller, B.; Hayward, D.C.; Ball, E.E.; Degnan, B.; Müller, P.; Spring, J.; Srinivasan, A.; Fishman, M.; Finnerty, J.; Corbo, J.; Levine, M.; Leahy, P.; Davidson, E.; Ruvkun, G.; (2000). 'Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA'. Nature 408 (6808): 86–9.

[133] Peterson, W.W. & Weldon, E. (1972), 'Error-correcting Codes', MIT Press.

[134] Pillai, R. S.; Bhattacharyya, S. N.; Filipowicz, W. (2007), 'Repression of protein synthesis by miRNAs: how many mechanisms?' Trends Cell Biol. 17, 118.

[135] Pollard, K.S.; Dudoit, S.; van der Laan, M.J. (2004). 'Multiple testing procedures: R multtest package and applications to genomics. in: Gentleman, R.; Carey, V.; Dudoit, S.; Irizarry, R.; Huber, W. (eds.). Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer, New York, chapter 15.

[136] Pruitt, K.D.; Tatusova, T.; Ostell, J.M.; McEntyre, J.; Ostell, J. (eds.). (2005). 'The Reference Sequence (RefSeq) Project'. National Library of Medicine (US), NCBI; Bethesda, MD: The NCBI Handbook. Chapter 18.

[137] Rainer, J.; Ploner, C.; Jesacher, S.; Ploner, A.; Eduardoff, M.; Mansha, M.; Wasim, M.; Panzer-Grumayer, R.; Trajanoski, Z.; Niederegger, H. & Kofler, R. (2009). 'Glucocorticoid-regulated microRNAs and mirtrons in acute lymphoblastic leukemia'. Leukemia. 23(4):746-52.

[138] Rajewsky, N. (2006). 'L(ou)sy miRNA targets?' Nat. Struct. Mol. Biol. 13:754-755.

[139] Reczko, M.; Maragkakis, M.; Alexiou, P. et al (2012). 'Functional microRNA targets in protein coding sequences'. Bioinformatics 28:771–776.

[140] Reinhart, B.J.; Slack, F.J.; Basson, M.; Pasquinelli, A.E.; Bettinger, J.C.; Rougvie, A.E.; Horvitz, H.R.; Ruvkun, G.; (2000). 'The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans'. Nature 403 (6772): 901–6.

[141] Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; & Smyth, G.K. (2015). 'Limma powers differential expression analyses for RNA-sequencing and microarray studies'. Nucleic Acids Research 43.

[142] Ritchie, Matthew E.; Silver, Jeremy; Oshlack, Alicia; Holmes, Melissa; Diyagama, Dileepa; Holloway, Andrew & Smyth, Gordon K. (2007). 'A comparison of background correction methods for two-colour microarrays'. Bioinformatics Vol. 23 no. 20, pages 2700–2707.

[143] Robinson, K.; Beverley, S. (2003), 'Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite Leishmania.', Mol Biochem Parasitol 128 (2): 217–28.

[144] Russo, Francesco; Di Bella, Sebastiano; Nigita, Giovanni; Macca, Valentina; Laganà, Alessandro; Giugno, Rosalba; Pulvirenti, Alfredo; Ferro, Alfredo (2012). 'miRandola: Extracellular Circulating microRNAs Database'. PLoS ONE 7(10).

[145] Saito, Y.; Liang, G.; Egger, G.; Friedman, J.M.; Chuang, J.C.; Coetzee, G.A.; Jones, P.A. (2006). 'Specific activation of microRNA-127 with downregulation of the proto-oncogene BCL6 by chromatin-modifying drugs in human cancer cells.' Cancer Cell 9:435-443.

[146] Salmena, L.; Poliseno, L.; Tay, Y.; Kats, L.; Pandolfi, P.P. (2011), 'A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?', Cell. 146(3):353-8.

[147] Saumet, A.; Lecellier, C.H. (2006), 'Anti-viral RNA silencing: do we look like plants ?'. Retrovirology 3 (3): 3.

[148] Schürmann, Thomas; Grassberger, Peter (1996). 'Entropy estimation of symbol sequences'. Chaos: An Interdisciplinary Journal of Nonlinear Science, Volume 6, Issue 3, pp.414-427.

[149] Schütz, S.; Sarnow, P. (2006), 'Interaction of viruses with the mammalian RNA interference pathway'. Virology 344 (1): 151–7.

[150] Sen, G.L.; Wehrman, T.S.; Blau, H.M. (2005), 'mRNA translation is not a prerequisite for small interfering RNA-mediated mRNA cleavage.', Differentiation 73 (6): 287–293.

[151] Sethupathy, P.; Megraw, M. & Hatzigeorgiou, A. G. (2006). 'A guide through present computational approaches for the identification of mammalian microRNA targets'. Nature Methods, vol. 3, no. 11, pp. 881–886.

[152] Shabalina, S.A.; Koonin, E.V. (2008), 'Origins and evolution of eukaryotic RNA interference.', Trends Ecol Evol 23:578–587.

[153] Shannon, C.E (1948). 'A mathematical theory of communication'. Bell Labs Techn J 27:379–423.

[154] Shi, L.; Reid, L.H.; Jones, W.D. et al (2006) 'The MicroArray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements'. Nat Biotechnol 24:1151–1161.

[155] Smyth, G. K. (2005). 'Limma: linear models for microarray data'. in: Gentleman, R.; Carey, V.; Dudoit, S.; Irizarry, R.; Huber, W. (eds.). Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer, New York, pages 397–420.

[156] Sontheimer, E.J. (2005). 'Assembly and function of RNA silencing complexes.', Nature Reviews Molecular Cell Biology 6 (2): 127–138.

[157] Stark, A., Brennecke, J., Bushati, N., Russell, R.B. & Cohen, S.M. (2005), 'Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution'. Cell 123, 1133–1146.

[158] Stefani, G.; Slack, F.J. (2008), 'Small noncoding RNAs in animal development'. Nat Rev Mol Cell Biol 9:219–230.

[159] Szoniec G. & Ogorzalek, M. (2013). 'Entropy of never born protein sequences'. SpringerPlus2:200.

[160] Tadeusiewicz, R. (2009), 'Studia doktoranckie w dyscyplinie biocybernetyka i inżynieria biomedyczna' Acta Bio-Optica et Informatica Medica vol. 15, nr 1.

[161] Tadeusiewicz, R. (2009). 'Modelowanie cybernetyczne i symulacja komputerowa systemów biologicznych'. Prace Komisji Nauk Technicznych PAU, tom III.

[162] Tadeusiewicz, R. (2009). 'O potrzebie i możliwości stworzenia cybernetycznej teorii systemów biomedycznych'. Inżynierowie dla Biologii i Medycyny 5, 3-6.

[163] Tam, O.H.; Aravin, A.A.; Stein, P.; Girard, A. et al. (2008), 'Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes.' Nature, 453(7194):534-8.

[164] TargetScan (http://www.targetscan.org/cgi-bin/targetscan/mirna_families.cgi?db=vert_61).

[165] TargetScan HumanRelese 6.2, June 2012 (http://www.targetscan.org/)

[166] Tay, Y.; Zhang, J.; Thomson, A.M. et al (2008). 'MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation'. Nature 455:1124–1128.

[167] Terwilliger, T.C. (2011). 'The success of structural genomics'. J Struct Funct Genom 12:43–44.

[168] Thomson, D.W.; Bracken, C.P.; Goodall, G.J. (2011). 'Experimental strategies for microRNA target identification'. Nucleic Acids Res 39:6845–6853.

[169] Towbin, H.; Staehelin, T.; Gordon, J. (1979). 'Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications'. Proc Natl Acad Sci. 76(9):4350-4.

[170] Trąbka J. (2005), 'Cyberkultura', Stowarzyszenie Twórcze Artystyczno-Literackie, Kraków.s.11-12.

[171] Vigorito, E.; Perks, K.L.; Abreu-Goodger, C.; Bunting, S.; Xiang, Z.; Kohlhaas, S.; Das, P.P.; Miska, E.A.; Rodriguez, A.; Bradley, A.; Smith, K.G.; Rada, C.; Enright, A.J.; Toellner, K.M.; Maclennan, I.C.; Turner, M. (2007) 'microRNA-155 regulates the generation of immunoglobulin class-switched plasma cells'. Immunity 27:847-859.

[172] von Neumann, J. 'First Draft of a Report on the EDVAC', June 30, 1945.

[173] Waller, Tomasz; Gubała, Tomasz; Sarapata, Krzysztof; Piwowar, Monika; Jurkowski, Wiktor (2015). 'DNA microarray integromics analysis platform'. BioData Mining 8:18.

[174] Wang, Y.; Addess, K.J.; Chen, J. et al (2007). 'MMDB: annotating protein sequences with Entrez's 3D-structure database'. Nucleic Acids Res 35:D298–D300.

[175] Weiss, O. et al (2000). 'Information content of protein sequences'. J Math Biol 206:379–386.

[176] Wiener, N. (1960). 'Application of cybernetics to medicine and biology'.Proc. I° int. congr. of cyb. med., Napoli.

[177] Wilkins, C.; Dishongh, R.; Moore, S.; Whitt, M.; Chow, M.; Machaca, K.; (2005). 'RNA interference is an antiviral defence mechanism in Caenorhabditis elegans'. Nature 436 (7053): 1044–7.

[178] Wolter, J.M.; Kotagama, K.; Pierre-Bez, A.; Firago, M. & Mangone, M. (2014). '3'LIFE: a functional assay to detect miRNA targets in high-throughput.' Nucleic Acids Res. 42 (17).

[179] Woltering, J.M.; Durston, A.J. (2008). 'MiR-10 represses HoxB1a and HoxB3a in zebrafish. PLoS ONE 3(1).

[180] Wu, CI;, Shen, Y.; Tang, T. (2009). 'Evolution under canalization and the dual roles of microRNAs: a hypothesis'. Genome Res. 19(5):734-43.

[181] Xiao, F.; Zuo, Z.; Cai, G.; Kang, S.; Gao, X.; Li, T. (2009). 'miRecords: an integrated resource for microRNA-target interactions'. Nucleic Acids Res. 37: D105-D110.

[182] Xin, Y.; Olson, W.K. (2009). 'BPS: a database of RNA base-pair structures'. Nucleic Acids Res 37:D83–D88.

[183] Yeung, Man L .; Benkirane, Monsef; Jeang, Kuan-Teh (2007), 'Small non-coding RNAs, mammalian cells, and viruses: regulatory interactions?' Retrovirology 4:74.

[184] Zambon, R.; Vakharia, V.; Wu, L. (2006). 'RNAi is an antiviral immune response against a dsRNA virus in Drosophila melanogaster'. Cell Microbiol 8 (5): 880–9.

[185] Zamore, P.D.; Tuschl, T.; Sharp, P.A.; Bartel, D.P. (2000). 'RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals.', Cell 101 (1): 25–33

[186] Zeliaś, A.; Pawełek, B.; Wanat, S. (2002), 'Metody statystyczne. Zadania i sprawdziany'. Polskie Wydawnictwo Ekonimiczne, Warszawa.

[187] Zhang, B.; Pan, X.; Cobb, G.; Anderson, T. (2006). 'Plant microRNA: a small regulatory molecule with big impact'. Dev Biol 289 (1): 3–16.

[188] Zon, J. (1987). 'Topografia badań w dziedzinie bioelektroniki'. in: Sedlak, W.; Zon, J.; Wnuk, M. (eds.). Bioelektronika. Materiały VI Sympozjum, KUL, Lublin, 20-21 XI, p.11-34.

9 Wykaz rysunków i tabel

Rysunki

Tabele

10 Wykaz załaczonych plików i informacje techniczne

Załączona do opracowania płyta CD zawiera następujące zbiory:

1	biTargetScore.R	implementacja modelu biTargetScore.R
2	ExpAnalysis1.txt	zbiór miRNAs z różnicową ekspresją "Astma"
3	TargetScoreAstma.R	implementacja modelu biTargetScore dla danych "Astma"
4	hsa_MTI.xls	zbiór targetów potwierdzonych eksperymentalnie
5	miRBase_conv_org.txt	plik konwersji nomenklatury miRNAs
6	biosql-1.0.1.zip	struktura BioSQL

Obliczenia w pracy przeprowadzono z wykorzystaniem oprogramowania:

1. **R language** - version 3.1.2 (2014-10-31)

Platform: x86_64-unknown-linux-gnu (64-bit)

locale:

[1] LC_CTYPE=en_US LC_NUMERIC=C LC_TIME=en_US

[4] LC_COLLATE=C LC_MONETARY=en_US LC_MESSAGES=en_US

[7] LC_PAPER=en_US LC_NAME=C LC_ADDRESS=C

[10] LC_TELEPHONE=C LC_MEASUREMENT=en_US LC_IDENTIFICATION=C

attached base packages:

[1] stats graphics grDevices utils datasets methods base

other attached packages:

[1] plyr_1.8.1 plotrix_3.5-11 gdata_2.13.3

[4] TargetScoreData_1.0.0 TargetScore_1.2.0 Matrix_1.1-5

[7] pracma_1.8.3

loaded via a namespace (and not attached):

[1] Rcpp_0.11.4 grid_3.1.2 gtools_3.4.1 lattice_0.20-30

2. **Serwer bazodanowy MySQL**:
   wersja serwera: 5.5.16-log MySQL Community Server (GPL)
3. **Jezyk skryptowy Ruby**:
   wersja: ruby 1.9.2p136 (2010-12-25 revision 30365) [x86_64-linux]

Dodatek A. Reguły prawdopodobieństwa

Klasyczna definicja prawdopodobieństwa P.S. Laplace'a [186] podaje, że prawdopodobieństwo zdarzenia losowego A, złożonego z m zdarzeń elementarnych stanowiących podzbiór wszystkich możliwych rozłącznych i jednakowo możliwych zdarzeń elementarnych n, wyznaczamy ze wzoru:

$$P(A) = \frac{m}{n} \quad \textbf{(A-1)}$$

gdzie:

m - liczba zdarzeń elementarnych sprzyjających zdarzeniu A;

n - liczba wszystkich, rozłącznych zdarzeń elementarnych.

Ze względu na wady klasycznej definicji prawdopodobieństwa, która mówi o zdarzeniach jednakowo możliwych, skończonej liczbie wszystkich zdarzeń elementarnych i wymaganej znajomości wszystkich zdarzeń elementarnych wprowadzono częstościowe definicje prawdopodobieństwa. W tym ujęciu zamiast posługiwania się liczbą zdarzeń elementarnych wprowadzono liczbę wykonanych doświadczeń. Wówczas częstość względną można uznać, jako ocenę prawdopodobieństwa. Im większa jest liczba doświadczeń, tym ocena prawdopodobieństwa jest lepsza.

$$P(A) = \lim_{n\to\infty} \frac{m_A}{n} \quad \textbf{(A-2)}$$

gdzie:

m_A - liczba doświadczeń, w których uzyskano zdarzenie A;

n - liczba przeprowadzonych doświadczeń.

Prawdopodobieństwo rozumiane jako granica częstości względnej zdarzenia A przy rosnącej do nieskończoności liczbie doświadczeń wprowadza zatem eksperyment niemożliwy do realizacji, jak również nie określa liczby doświadczeń wystarczającej do wyliczenia prawdopodobieństwa. Dalszy rozwój teorii prawdopodobieństwa przeprowadził T. Bayes. Przystosował on pojęcie prawdopodobieństwa dla zdarzeń niepowtarzalnych – wprowadzając miarę niepewności albo wiarygodności, gdzie prawdopodobieństwo tzw. *a posteriori* powstaje przez przekształcenie założenia – prawdopodobieństwa *a priori* w świetle uzyskanych danych w przeprowadzonym eksperymencie.

Zmienną losową nazywamy dyskretną, gdy przyjmuje ona wartości skończone lub przeliczalne, a występowanie tych wartości jest zdarzeniem losowym o danym prawdopodobieństwie. Funkcją rozkładu prawdopodobieństwa p zmiennej losowej X nazywamy funkcję przyporządkowującą wartościom zmiennej losowej odpowiednie wartości prawdopodobieństwa.

Prawdopodobieństwo p zdarzenia przypisuje zdarzeniom losowym wartości liczbowe z przedziału [0,1]. Wartości te są przypisywane na podstawie przeprowadzanych serii identycznych prób, czyli takich, których nie potrafimy od siebie odróżnić; brak jest czynników różnicujących je od siebie. Dla doświadczeń wieloetapowych przy obliczaniu prawdopodobieństwa posługujemy się dwoma podstawowymi regułami:

Regułą sum:

$$p(X) = \sum_{Y} p(X,Y) \tag{A-3}$$

Regułą iloczynów:

$$p(X,Y) = p(Y|X)p(X) \tag{A-4}$$

gdzie X, Y – zmienne losowe.

Prawdopodobieństwem warunkowym $P(A/B)$ zdarzenia A pod warunkiem, że zaszło zdarzenie B, nazywamy iloraz prawdopodobieństwa łącznego zajścia zdarzenia A i B oraz prawdopodobieństwa darzenia B.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \tag{A-5}$$

Dwa zdarzenia są niezależne w przypadku, gdy $P(A/B) = P(A)$ lub $P(B/A) = P(B)$. Wówczas otrzymujemy:

$$P(A \cap B) = P(A) \cdot P(B) \tag{A-6}$$

Na podstawie reguły iloczynów oraz właściwości symetryczności $p(X,Y) = p(Y,X)$ uzyskujemy związek między prawdopodobieństwami warunkowymi zwany twierdzeniem Bayesa:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{A-7}$$

Uwzględniając regułę sum zmieniamy mianownik w powyższym równaniu (A-7) uzyskując:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)} \tag{A-8}$$

Tak przedstawiony mianownik ma właściwość stałej normalizującej, aby suma wszystkich prawdopodobieństw warunkowych była równa jeden.

Zmienną losową parametryzuje się niekiedy wyznaczając średnią μ i wariancją σ^2. Średnią μ nazywa się często wartością oczekiwaną zmiennej losowej $E(X) = \mu$. Definiujemy ją następująco:

$$E(X) = \sum_{i=1}^{n} p_i x_i \tag{A-9}$$

Tak definiowana wartość stanowi średnią ważoną przyjmowaną przez zmienną losową, gdzie wagi poszczególnych wartości stanowią prawdopodobieństwa ich wystąpienia. W tym wzorze wartość n oznacza liczbę wszystkich skończonych wartości zmiennej losowej X.

Wartość wariancji obliczamy według wzoru:

$$\sigma^2 = \sum_{i=1}^{n} p_i(x_i - \mu)^2 = E[X - E(X)]^2 \quad \textbf{(A-10)}$$

Dodatek B. Prawdopodobieństwo zmiennej losowej ciągłej

W przeciwieństwie do zmiennych losowych dyskretnych - nieciągłych, zbiór możliwych wartości zmiennej losowej ciągłej jest nieskończony i nieprzeliczalny. Funkcja gęstości określa rozkład prawdopodobieństwa zmiennej losowej ciągłej. Jeśli prawdopodobieństwo zmiennej rzeczywistej x należącej do przedziału $(x, x + \delta x)$ jest dane funkcją $p(x)\delta x$ dla $\delta x \to 0$, wówczas funkcje $p(x)$ nazywa się funkcją gęstości prawdopodobieństwa. Funkcja ta przyporządkowuje wartościom zmiennej losowej odpowiednie wartości prawdopodobieństwa.

$$p(x\epsilon(a,b))) = \int_a^b p(x)dx \tag{B-1}$$

Ponieważ wartość prawdopodobieństwa musi być nieujemna oraz wartość x musi leżeć gdzieś na osi rzeczywistej funkcja gęstości prawdopodobieństwa musi spełniać warunki:

$$p(x) \geq 0 \tag{B-2}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{B-3}$$

Reguły sum i iloczynów dla zmiennych ciągłych z użyciem funkcji gęstości prawdopodobieństwa mają postać:

$$p(x) = \int p(x,y)dy \tag{B-4}$$

$$p(x,y) = p(y|x)p(x) \tag{B-5}$$

Dodatek C. Rozkład Gaussa

Rozkład Gaussa zwany też rozkładem normalnym stanowi jeden z podstawowych rozkładów często stosowanych przy dopasowaniu krzywej do rozkładów zmiennych uzyskanych w doświadczeniach. Dla zmiennej rzeczywistej jedno x i wielowymiarowych $\boldsymbol{x}$ (D) opisują go wzory:

$$N(x|\mu,\delta^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \quad \textbf{(C-1)}$$

$$N(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \quad \textbf{(C-2)}$$

gdzie:

μ - wartość średnia;

σ – odchylenie standartowe;

$\boldsymbol{\mu}$ - wektor średnich D wymiarowy;

Σ - macierz kowariancji D x D.

Rozkład Gaussa jest często wykorzystywany jako rozkład a prioryczny. Należy on do tzw. rozkładów sprzężonych (*conjugate priors*). Rozkłady te posiadają tą własność, że rozkład *a posteriori* należy do tej samej rodziny rozkładów, co rozkład *a priori*. W takich przypadkach rolą funkcji wiarygodności jest uaktualnienie parametrów rozkładu aprioryczngo.

Dodatek D. Model graficzny - sieć Bayesa

Sieć Bayesa reprezentuje strukturę modelu probabilistycznego, czyli rozkład prawdopodobieństw. Wyraża ona zależności i niezależności zmiennych losowych. Sieć bayesowska koduje informacje za pomocą grafu, którego wierzchołki (węzły) stanowią zmienne losowe, a krawędzie obrazują probabilistyczne zależności między zmiennymi. Taki graf zawiera sposób, w jaki łączne prawdopodobieństwo wszystkich zmiennych losowych zdekomponować na czynniki, z których każdy zależny jest jedynie od pewnego podzbioru zmiennych. Graf, w którym połączenia mają wyróżniony kierunek za pomocą strzałek nazywa się skierowanym grafem acyklicznym. Taka struktura odwzorowuje przyczynowy aspekt dziedziny.

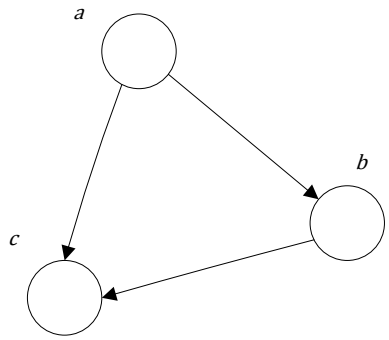
Rozważając łączny rozkład $p(a, b, c)$ trzech zmiennych *a, b* oraz *c* możemy ją korzystając z reguły iloczynów przedstawić w postaci:

$$p(a, b, c) = p(c|a, b)p(a, b) \qquad \textbf{(D-1)}$$

Ponownie stosując tą regułę dla drugiego członu prawej strony równania otrzymujemy:

$$p(a, b, c) = p(c|a, b)p(b|a)p(a) \qquad \textbf{(D-2)}$$

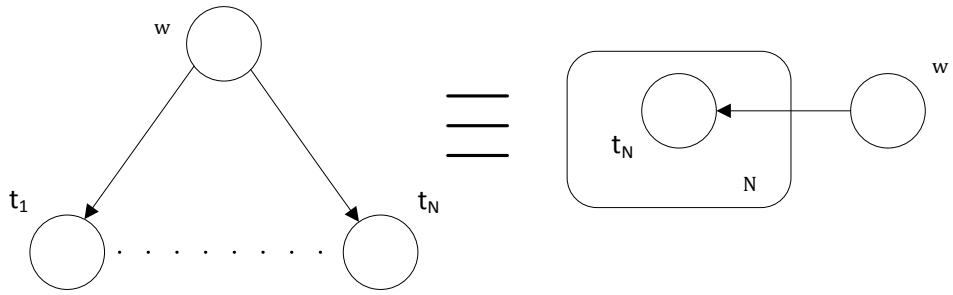
Reprezentacja graficzna takiego modelu może wyglądać następująco:



Rys. D.1 Graf skierowany reprezentujący łączne prawdopodobieństwo trzech zmiennych $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ uzyskane poprzez dekompozycje prawostronną.

W przypadku dużej liczby zmiennych o jednakowej zależności stosuje się wzory uproszczone. Przykład reprezentacji graficznej dla łącznego prawdopodobieństwa:

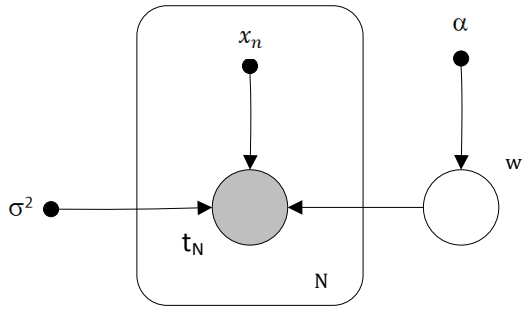
$$p(\boldsymbol{t}, \boldsymbol{w}) = p(\boldsymbol{w}) \prod_{n=1}^{N} p(t_n|\boldsymbol{w}) \qquad \textbf{(D-3)}$$



Rys. D.2. Równoważny graf dla N różnych zmiennych zależnych w postaci płytki oznaczonej N. Przedstawia on N węzłów z za pomocą jednego oznaczonego t_N.

Zmienne losowe przedstawia się dużym, pustym w środku okręgiem. Natomiast określone parametry oznacza się małym wypełnionym okręgiem. Obserwowane zmienne oznacza się poprzez zacieniowanie określonego węzła.

$$p(\boldsymbol{t}, \boldsymbol{w}|\boldsymbol{x}, \alpha, \sigma^2) = p(\boldsymbol{w}|\alpha) \prod_{n=1}^{N} p(t_n|\boldsymbol{w}, x_n, \sigma^2) \quad \textbf{(D-4)}$$

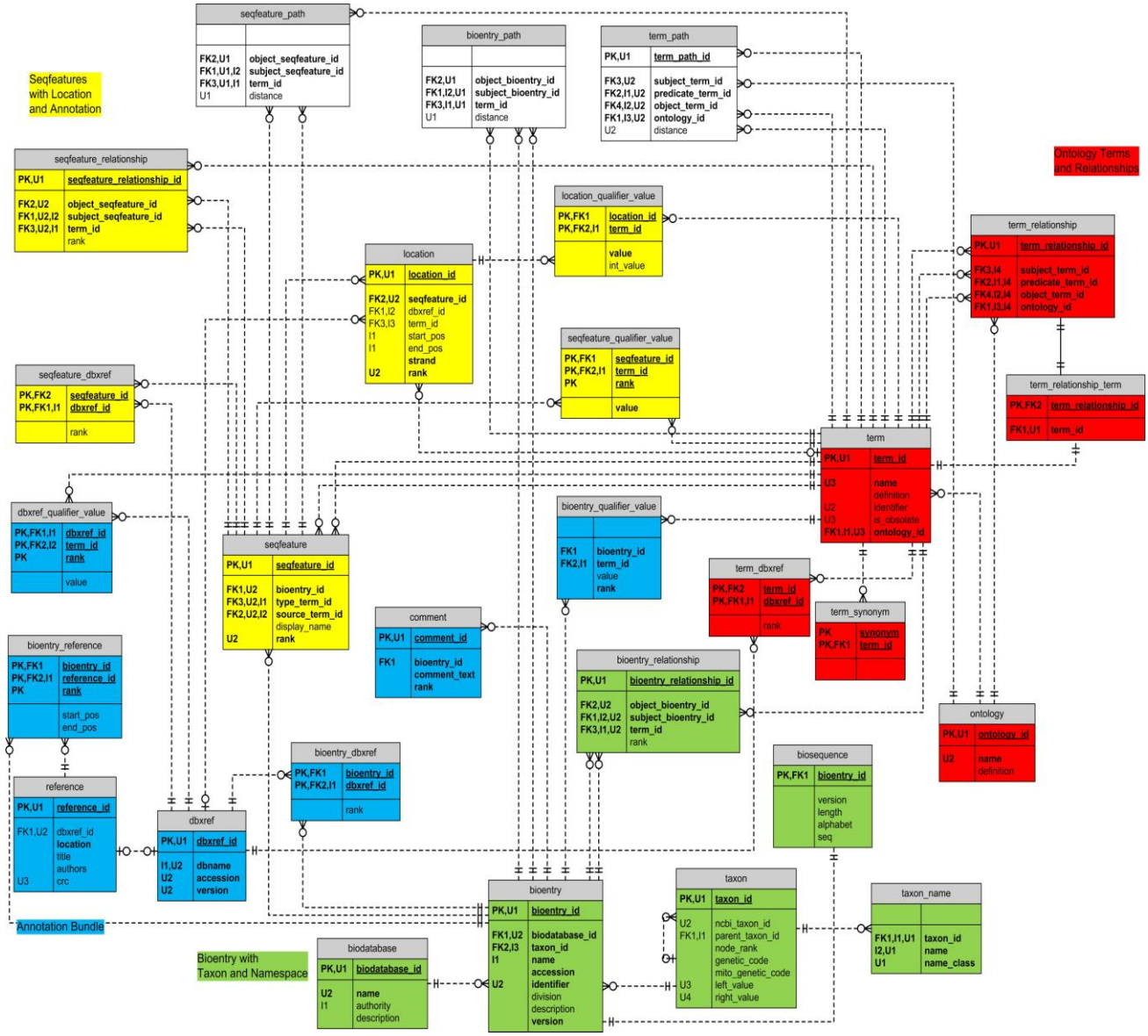


Rys. D.3. Graf ze zmienną obserwowaną t_n, oraz zmiennymi określonymi: x_n, α, σ^2.

Sieci Bayesa umożliwiają obliczenie rozkładów prawdopodobieństwa wybranych, interesujących nas zmiennych, pod warunkiem ustalenia stanów pozostałych zmiennych zależnych. Prawdopodobieństwo wystąpienia pewnego zdarzenia w określonych okolicznościach jest zatem prawdopodobieństwem tego zdarzenia zależnym (w sensie warunkowości) od tych okoliczności.

Dodatek E. Struktura modelu BioSQL

Lokalna implementacja BioSQL (http://www.biosql.org/) zawiera rozpakowaną strukturę relacyjną zawartą na płycie CD dołączonej do pracy.



Rys. E.1 Struktura relacyjna modelu BioSQL