Streszczenie rozprawy doktorskiej p.t.:

Nowe metody identyfikacji mikroRNA

(ang. *New methods for identification of microRNAs*)

Michał Szcześniak

Promotor:
prof. UAM dr hab. Izabela Makałowska

Pracownia Bioinformatyki
Instytut Biologii Molekularnej i Biotechnologii
Uniwersytet im. Adama Mickiewicza w Poznaniu

Poznań 2013

1. Wprowadzenie

Odkrycie małych regulatorowych RNA było jednym z najważniejszych wydarzeń w biologii molekularnej ostatnich lat. Cząsteczki te podzielono na liczne klasy, np. miRNA, siRNA czy piRNA, odpowiedzialne za różnorodne procesy komórkowe (Ghildiyal i Zamore, 2009). Spośród nich prawdopodobnie miRNA (mikroRNA) zyskały najwięcej uwagi. Liczne eksperymenty i analizy bioinformatyczne znacząco poszerzyły naszą wiedzę o ich biogenezie i pełnionych funkcjach. Jednocześnie niezwykle szybko wzrosła liczba znanych miRNA. miRNA zidentyfikowano już u setek gatunków roślin i zwierząt, u wirusów, a ostatnio również u protistów i grzybów.

U roślin miRNA uczestniczą w procesach wzrostu i rozwoju, w tym np. powstawaniu korzeni bocznych, sygnalizacji hormonalnej, regulacji czasu kwitnienia czy przejściu z fazy juwenilnej do wegetatywnej (Mallory i Vaucheret, 2006). Szczególną cechą roślinnych miRNA jest ich udział w odpowiedzi na czynniki stresowe, takie jak susza, niska temperatura czy niedobór azotu (Sunkar i in., 2007). Zwierzęce miRNA również regulują cały szereg procesów komórkowych (Siomi i Siomi, 2010), a w szczególności powiązano je z chorobami, takimi jak nowotwory czy reumatoidalne zapalenie stawów (Jiang i in., 2009). Dodatkowo, zarówno u roślin jak i zwierząt, ekspresji ulegają wirusowe miRNA (Nair i Zavolan, 2006). Nie są one homologami miRNA gospodarza, ale używają jego enzymów w procesie biogenezy. Regulują one zarówno cykl życiowy wirusa, jak i interakcje między wirusem a gospodarzem.

Ponieważ miRNA pełnią tak ważne i różnorodne funkcje w komórce, odkrywanie nowych miRNA i dalsze zgłębianie ich biologii może być kluczowe dla zrozumienia wielu procesów molekularnych, a także pozwolić na wykorzystanie tych cząsteczek w biologii molekularnej, biotechnologii czy medycynie. Z tego powodu powstał szereg metod bioinformatycznych służących do identyfikacji miRNA. Można je podzielić na dwie podstawowe kategorie: metody oparte na homologii, służące do szukania sekwencji podobnych do znanych miRNA oraz metody *de novo*, pozwalające na szukanie miRNA należących do wcześniej nieznanych rodzin. Opisany niżej *algorytm miRNEST* należy do pierwszej kategorii, zaś HuntMi jest narzędziem służącym do identyfikacji miRNA *de novo*. Obecnie oba podejścia coraz częściej łączy się z metodami eksperymentalnymi, zwłaszcza sekwencjonowaniem małych RNA z wykorzystaniem technologii NGS (ang. *Next-Generation Sequencing*). Rozwiązanie to zostanie wykorzystane podczas najbliższej aktualizacji bazy danych miRNEST.

2. HuntMi: an efficient and taxon-specific approach in pre-miRNA identification (Gudyś i in., 2013)

Obecnie dostępne narzędzia służące do identyfikacji miRNA *de novo* są obarczone istotnymi niedoskonałościami metodologicznymi oraz ograniczeniami w ich używaniu. Na przykład niektóre narzędzia są tworzone wyłącznie z myślą o gatunkach modelowych; do testowania narzędzi wykorzystuje się zbiór treningowy (zbiór testowy i treningowy powinny być rozłączne); rozpatruje się tylko jedną, wybraną metodę nauczania maszynowego; wykorzystuje się niskiej jakości sekwencje w zbiorach pozytywnych i/lub negatywnych; nie uwzględnia się problemu niezbalansowanie rozmiaru zestawów sekwencji pozytywnych i negatywnych, co skutkuje niewłaściwym oszacowaniem wydajności klasyfikatora.

Opracowując HuntMi, nowe narzędzie do identyfikacji miRNA, podjęliśmy próbę rozwiązania tych problemów, jednocześnie mając na celu maksymalizację czułości i specyficzności. W pierwszej kolejności przygotowaliśmy wysokiej jakości dane wejściowe. Zestaw sekwencji pozytywnych składał się z potwierdzonych eksperymentalnie pre-miRNA, zaś sekwencje negatywne były pobrane losowo z genomów i transkryptomów odpowiednich gatunków, po czym usunięte zostały wszystkie sekwencje, które wykazywały nawet niewielkie podobieństwo do znanych pre-miRNA. W eksperymentach kroswalidacyjnych (ang. *cross-validation experiments*) przetestowaliśmy cztery metody nauczania maszynowego (naiwny klasyfikator bayesowski, perceptron wielowarstwowy, maszynę wektorów nośnych i lasy losowe). Każda z tych metod była testowana na różnych zestawach parametrów wejściowych. Ostatecznie wybraliśmy metodę lasów losowych, jako że otrzymaliśmy dla niej najlepszą czułość i specyficzność. W dalszej kolejności wprowadziliśmy siedem nowych cech do klasyfikacji, oprócz 21 cech bazowych z programu microPred (Batuwita i Palade, 2009), co pozwoliło na poprawienie wydajności metody. Podjęliśmy także problem niezbalansowania zbiorów treningowych, tzn. różnicy między wielkością zbioru pozytywnego i negatywnego. W tym celu zaimplementowaliśmy nową technikę, nazwaną przez nas ROC-select, która okazała się być lepsza od innych znanych metod rozwiązywania problemu niezbalansowania, przynajmniej w dziedzinie identyfikacji miRNA.

Naszą metodę porównaliśmy z wiodącymi narzędziami do identyfikacji miRNA *de novo*: microPred (Batuwita i Palade, 2009), PlantMiRNAPred (Xuan i in., 2011) i MiRenSVM (Ding i in., 2010). Okazało się, że pod względem wydajności nasz algorytm je przewyższa. W dalszej kolejności, w oparciu o opracowaną metodę, zbudowaliśmy narzędzie HuntMi. Oprócz wyżej wspomnianych cech, niewątpliwą zaletą HuntMi jest jego elastyczność, gdyż na przykład

pozwala użytkownikowi w łatwy sposób tworzyć własne klasyfikatory, w oparciu o dane z dowolnego gatunku, a następnie wykorzystać je podczas identyfikacji miRNA.

3. miRNEST database: an integrative approach in microRNA search and annotation (Szcześniak i in., 2012)

Sekwencje EST, czyli znaczniki sekwencji ulegających ekspresji, są dostępne dla setek gatunków roślin i zwierząt (Boguski i in., 1993). Ponieważ pośród sekwencji EST można znaleźć sekwencje pre-miRNA, postanowiliśmy wykorzystać te dane do identyfikacji nowych miRNA. W tym celu zbudowaliśmy potok analityczny, nazwany *algorytmem miRNEST*, pozwalający na szukanie nowych miRNA na zasadzie podobieństwa do znanych dojrzałych miRNA. Podstawowe etapy analizy w potoku analitycznym to: i) szukanie sekwencji EST wykazujących podobieństwo do znanych dojrzałych miRNA; ii) składanie EST w tzw. kontigi; iii) usunięcie sekwencji tRNA i rRNA; iv) usunięcie sekwencji zajętych w ponad 60% przez regiony o niskiej złożoności (ang. *low-complexity regions*); v) sprawdzenie struktury drugorzędowej; vi) usunięcie kandydatów wykazujących podobieństwo do znanych białek; vii) usunięcie kandydatów o zbyt długiej sekwencji pre-miRNA (w przypadku zwierząt). Zidentyfikowaliśmy 10 004 miRNA u 221 gatunków zwierząt i 199 gatunków roślin. Uzyskane wyniki uzupełniliśmy danymi z innych źródeł: miRBase (Kozomara i Griffiths-Jones, 2009), PMRD (Zhang i in., 2010), microPC (Mhuantong i Wichadakul, 2009) oraz dwóch publikacji (Huang i in., 2009; Hao i in., 2010). W celu znalezienie podobieństw między sekwencjami zastosowaliśmy program BLAST (Altschul i in., 1990). Następnie zmapowaliśmy sekwencje ze 192 bibliotek małych RNA z bazy GEO (Barrett i in., 2011) do sekwencji pre-miRNA z wykorzystaniem narzędzia Bowtie (Langmead i in., 2009). Dodatkowo pobraliśmy dane z 13 baz danych miRNA, w tym miRTarBase (Hsu i in., 2010), Phenomir (Ruepp i in., 2010), dPORE-miRNA (Schmeier i in., 2011) czy Patrocles (Hiard i in., 2010).

Jako że roślinne miRNA cechują się wysokim stopniem komplementarności z docelowym mRNA, poszukiwanie roślinnych sekwencji docelowych metodami bioinformatycznymi jest z reguły stosunkowo prostym zadaniem. Wykorzystując nasz program, zidentyfikowaliśmy 6 963 sekwencje docelowe u 187 gatunków. W przypadku zwierząt często wykorzystuje się informację o zakonserwowaniu sekwencji docelowej między gatunkami, by otrzymać wiarygodne wyniki. Takie dane nie są dostępne dla zdecydowanej większości analizowanych gatunków zwierząt i dlatego informacje na temat miejsc docelowych zwierzęcych

miRNA pobraliśmy z odpowiednich baz danych.

Dane uzyskane na wyżej wymienionych etapach zdeponowaliśmy w nowej internetowej bazie danych, którą nazwaliśmy miRNEST. Interfejs użytkownika podzieliliśmy na pięć sekcji, pozwalających na dostęp do danych i opcji przeszukiwania różnego rodzaju. *Browse* umożliwia użytkownikowi dostęp do sekwencji miRNA przechowywanych w bazie danych, zarówno tych przewidzianych *algorytmem miRNEST*, jak również sekwencji z zewnętrznych źródeł. W *Search* zaimplementowano metody służące do przeszukiwania bazy i filtrowania prezentowanych użytkownikowi wyników, na przykład na podstawie sekwencji dojrzałego miRNA czy źródła sekwencji. Sekcja *Unclassified* gromadzi sekwencje przewidziane przez *algorytm miRNEST*, które jednak naruszają jeden z kryteriów: E-value uzyskane w trakcie przeszukiwania bazy UniProt > 1e-20 lub długość pre-miRNA ≤ 215 nt (tylko w przypadku zwierząt). *RNA-Seq* przedstawia wyniki mapowania małych RNA do sekwencji pre-miRNA. Ostatecznie, *Taxonomy* pozwala przeszukiwać na drzewie filogenetycznym gatunki, dla których w bazie miRNEST zdeponowano wyniki identyfikacji miRNA.

Obecnie baza danych miRNEST przechodzi aktualizację i rozbudowę. Wykonywane prace to m.in.:

i) Identyfikacja miRNA *de novo* z wykorzystaniem sekwencji genomów i bibliotek małych RNA pochodzących z sekwencjonowania w technologii NGS. W tym celu zbudowaliśmy potok analityczny i wykonaliśmy wielkoskalowe obliczenia, które pozwoliły na znalezienie setek nowych miRNA u 21 gatunków roślin i zwierząt.

ii) Przystosowaliśmy wyżej wspomniany potok analityczny do szukania miRNA, których prekursor obejmuje całą sekwencje intronu (tzw. mirtronów); znaleźliśmy 128 kandydatów u dwunastu gatunków zwierząt.

iii) Przeanalizowaliśmy degradomy dziesięciu gatunków roślin z wykorzystaniem programu PAREsnip (Folkes i in., 2012), aby znaleźć sekwencje docelowe miRNA potwierdzone eksperymentalnie; zidentyfikowaliśmy 1931 par miRNA-sekwencja docelowa.

iv) Sekwencje pre-miRNA przechowywane w bazie miRNEST przeanalizowaliśmy programem HuntMi.

v) Pobraliśmy zdeponowane w bazie ERISdb dane nt. struktury genów miRNA i dodatkowo wykonaliśmy analogiczne analizy dla pięciu gatunków roślin: *Brachypodium distachyon, Malus domestica, Medicago truncatula, Populus trichocarpa* i *Solanum lycopersicum.*

4. ERISdb: a database of plant splice sites and splicing signals (Szcześniak i in., 2013)

Badacze coraz bardziej są świadomi tego, że poznanie struktury genów mikroRNA, w tym alternatywnych form splicingowych, może być kluczowe w zrozumieniu niektórych aspektów ich biologii. Niestety większość poszukiwań miRNA skoncentrowanych jest na sekwencjach pre-miRNA i dojrzałych miRNA, przez co słabo poznaliśmy budowę genów miRNA. Niemniej jednak pojawiły się już pojedyncze prace dotyczące roślinnych miRNA, m.in. u *Arabidopsis thaliana* (Szarzynska i in., 2009), *Vitis vinifera* (Mica i in., 2010) czy ostatnio u *Hordeum vulgare* (Kruszka i in., 2013). Zwierzęce miRNA prawdopodobnie nie posiadają intronów lub introny występują w nich bardzo rzadko. Niewielka wiedza w tej dziedzinie zmotywowała nas do przeprowadzenia analiz bioinformatycznych skoncentrowanych na przewidywaniu miejsc splicingowych z wykorzystaniem sekwencji EST u siedmiu gatunków roślin: *Arabidopsis thaliana, Chlamydomonas reinhardtii, Glycine max, Oryza sativa, Physcomitrella patens, Selaginella moellendorffii* i *Zea mays*.

Pierwszym etapem obliczeń w naszym potoku analitycznym było szukanie sekwencji EST z bazy dbEST (Boguski i in., 1993), które odpowiadają znanym pre-miRNA przechowywanym w bazie danych miRBase (Kozomara i Griffiths-Jones, 2011). Szukanie wykonaliśmy programem Megablast (Altschul i in., 1990), wymagając by sekwencja EST wykazywała min. 97% identyczności z sekwencją pre-miRNA na min. 90% jej długości. Wyselekcjonowane sekwencje EST zostały zmapowane do genomów odpowiednich gatunków roślin z wykorzystaniem programu Splign (Kapustin i in., 2008). Otrzymane dane poddaliśmy dodatkowej obróbce, a następnie umieściliśmy je w utworzonej przez nas bazie danych ERISdb. Udało nam się zidentyfikować introny w 45 genach miRNA u pięciu gatunków roślin. Niektóre z tych genów posiadają więcej niż jeden intron (maksymalnie sześć) i czasami przechodzą one alternatywny splicing. W bazie danych ERISdb przewidziane miejsca splicingowe są przedstawione w postaci przyrównania trzech sekwencji: pre-miRNA, genomu oraz EST, co pozwala użytkownikowi zrozumieć kontekst w jakim pojawia się intron. W przypadku ośmiu miRNA u *A. thaliana* użyliśmy adnotacji z bazy danych Ensembl (Kersey i in., 2010). Dodatkowo, wykorzystaliśmy sekwencje pri-miRNA u *A. thaliana* uzyskane w eksperymentach RACE (Szarzynska i in., 2009), zaś w przypadku *V. vinifera* zdeponowaliśmy w bazie danych trzy miRNA z potwierdzeniem miejsc splicingowych w postaci RNA-Seq (Mica i in., 2010).

5. Bazy danych mikroRNA (Szcześniak i in., 2012)

Szybki postęp w opracowywaniu obliczeniowych i eksperymentalnych metod szukania nowych miRNA i ich analizy poskutkował znaczącym przyrostem danych i koniecznością tworzenia dedykowanych baz danych. Jedną z pierwszych baz danych miRNA był miRBase. Dziś baza ta gromadzi dane o miRNA u 67 gatunków roślin, 97 gatunków zwierząt oraz 26 wirusów i jest uznawana ze referencyjną bazę danych w dziedzinie mikroRNA. Innymi kolekcjami sekwencji miRNA są PMRD (*Plant MicroRNA Database*), microPC i miRNEST. Ponadto istnieje szereg baz danych poświęconych różnym aspektom biologii miRNA, jak profile ekspresji miRNA, ich sekwencje docelowe czy polimorfizm sekwencji. W sumie można naliczyć około 60 repozytoriów poświęconych miRNA. Skutkuje to tym, że coraz trudniej jest znaleźć użytkownikowi odpowiednią bazę danych i dotrzeć do interesujących go danych. W związku z tym postanowiliśmy napisać pracę przeglądową o bazach danych miRNA, w której krótko charakteryzujemy 51 baz danych opublikowanych do listopada 2011 roku. Dodatkowo omawiamy podstawowe źródła informacji w tych bazach oraz sugerujemy jak powinna wyglądać dobra baza danych miRNA.

6. Bibliografia

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990; 215:403-410.
2. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res. 2011; 39:D1005–D1010.
3. Batuwita R, Palade V. MicroPred: effective classification of pre-miRNAs for human MiRNA gene prediction. Bioinformatics 2009, 25:989–995.
4. Boguski MS, Lowe TM, Tolstoshev CM. dbEST—database for "expressed sequence tags" Nat Genet. 1993;4:332–333.
5. Ding J, Zhou S, Guan J: MiRenSVM. towards better prediction of MicroRNA precursors using an ensemble SVM classifier with multi-loop features. BMC Bioinformatics 2010, 11 (Suppl 11):S35.
6. Folkes L, Moxon S, Woolfenden HC, Stocks MB, Szittya G, Dalmay T, Moulton V. PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. Nucleic Acids Res. 2012; 40(13):e103.
7. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet. 2009;10(2):94-108.
8. Gudyś A, Szcześniak MW, Sikora M, Makalowska I. HuntMi: an efficient and taxon-specific approach in pre-miRNA identification BMC Bioinformatics 2013, 14:83.
9. Hao L, Cai P, Jiang N, Wang H, Chen Q. Identification and characterization of microRNAs and endogenous siRNAs in Schistosoma japonicum. BMC Genomics 2010; 11:55.
10. Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. Nucleic Acids Res. 2010; 38:D640–D651.
11. Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM, Chien CH, Wu MC, Huang CY, Tsou AP, Huang HD. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. 2010;39:D163–D169.
12. Huang J, Hao P, Chen H, Hu W, Yan Q, Liu F, Han ZG. Genome-wide identification of Schistosoma japonicum microRNAs using a deep-sequencing approach. PLoS One 2009;

4:e8206.

13. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 2009; 37: D98-104.
14. Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol. Direct 2008; 3:20.
15. Kersey PJ, Staines DM, Lawson D, Kulesha E, Derwent P, Humphrey JC, Hughes DS, Keenan S, Kerhornou A, Koscielny G, Langridge N, McDowall MD, Megy K, Maheswari U, Nuhn M, Paulini M, Pedro H, Toneva I, Wilson D, Yates A, Birney E. Ensembl Genomes: extending Ensembl across the taxonomic space. Nucleic Acids Res. 2010; 38:D563-D569.
16. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011; 39:D152–D157.
17. Kruszka K, Pacak A, Swida-Barteczka A, Stefaniak AK, Kaja E, Sierocka I, Karlowski W, Jarmolowski A, Szweykowska-Kulinska Z. Developmentally regulated expression and complex processing of barley pri-microRNAs. BMC Genomics 2013; 14:34.
18. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25.
19. Mallory AC, Vaucheret H. Functions of microRNAs and related small RNAs in plants. Nat Genet. 2006;38 Suppl:S31-6. Erratum in: Nat Genet. 2006 Jul;38(7):850.
20. Mhuantong W, Wichadakul D. MicroPC (microPC): A Comprehensive resource for predicting and comparing plant MicroRNAs. BMC Genomics 2009, 10: 366.
21. Mica E, Piccolo V, Delledonne M, Ferrarini A, Pezzotti M, Casati C, Del Fabbro C, Valle G, Policriti A, Morgante M, Pesole G, Pè ME, Horner DS. High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera. BMC Genomics 2009; 10:58.
22. Nair V, Zavolan M. Virus-encoded microRNAs: novel regulators of gene expression. Trends Microbiol. 2006; 14(4):169-75.
23. Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ. PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. Genome Biol. 2010;11:R6.
24. Schmeier S, Schaefer U, MacPherson CR, Bajic VB. dPORE-miRNA: polymorphic regulation of microRNA genes. PLoS One. 2011; 6:e16657.

25. Siomi H, Siomi MC. Posttranscriptional regulation of microRNA biogenesis in animals. Mol Cell. 2010; 38(3):323-32.
26. Sunkar R, Chinnusamy V, Zhu J, Zhu JK. Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. Trends Plant Sci. 2007;12(7):301-9.
27. Szarzynska B, Sobkowiak L, Pant BD, Balazadeh S, Scheible WR, Mueller-Roeber B, Jarmolowski A, Szweykowska-Kulinska Z. Gene structures and processing of Arabidopsis thaliana HYL1-dependent pri-miRNAs. Nucleic Acids Res. 2009; 37:3083-3093.
28. Szcześniak MW, Deorowicz S, Gapski J, Kaczyński Ł, Makalowska I. miRNEST database: an integrative approach in microRNA search and annotation. Nucleic Acids Res. 2012; 40:D198-204.
29. Szcześniak MW, Kabza M, Pokrzywa R, Gudyś A, Makałowska I. ERISdb: a database of plant splice sites and splicing signals. Plant Cell Physiol. 2013; 54(2):e10.
30. Szcześniak MW, Owczarkowska E, Gapski J, Makałowska I. Bazy danych mikroRNA. Postepy Bioch. 2012; 58(1).
31. Vaucheret H, Voinnet O, Watanabe Y, Weigel D, Zhu JK. Criteria for annotation of plant MicroRNAs. Plant Cell 2008; 20(12):3186-90.
32. Xuan P, Guo M, Liu X, Huang Y, Li W, Huang Y. PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. Bioinformatics 2011, 27: 1368–1376.
33. Zhang Z, Yu J, Li D, Zhang Z, Liu F, Zhou X, Wang T, Ling Y, Su Z. PMRD: plant microRNA database. Nucleic Acids Res. 2010; 38:D806–813.

1. Introduction

Discovery of small regulatory RNAs has immensely changed our understanding of gene expression regulation. These RNAs can be divided into several major classes, like miRNA, siRNA, or piRNA that perform a wide range of molecular functions (Ghildiyal and Zamore, 2009). Among them, miRNAs (microRNAs) are the class that probably gained most attention. Countless experiments and analyses greatly increased our knowledge about their biogenesis and functions. Also the number of known miRNAs rose dynamically; so far, miRNAs have been discovered mostly in plants, animals and viruses but recently it was shown that also in fungi and protists these small RNAs can be expressed.

In plants miRNAs participate in different aspects of plant growth and developmental processes, including lateral root formation, hormone signaling, flowering time, or transition from juvenile to adult vegetative phase (Mallory and Vaucheret, 2006). In particular, plant miRNAs are known for their roles in response to stress conditions, like drought, low temperatures or nitrogen deficiency (Sunkar *et al*., 2007). In animals miRNAs are believed to regulate up to 60% of protein-coding genes and, like in plants, are implicated in a number of biological processes (Siomi and Siomi, 2010). Notably, miRNAs have been associated with diseases, like cancers or rheumatoid arthritis (Jiang *et al*., 2009). In animals and plants also virus miRNAs can be expressed (Nair and Zavolan, 2006). They show no resemblance to miRNAs encoded in plant or animal genomes but use host machinery during biogenesis. They regulate both viral life cycle and the interaction between viruses and their hosts.

The fact that miRNAs participate in all major molecular processes in a cell and that they could find multiple applications in biotechnology, molecular biology or medicine, motivated extensive development of miRNA search methods. The methods can be divided into two groups: homology-based, allowing us to search for sequences similar to already known miRNAs, and *de novo* methods that make it possible to identify miRNAs belonging to novel miRNA families. The *miRNEST algorithm*, described below, belongs to the first group, while HuntMi is a *de novo* miRNA search tool. Currently both approaches are frequently used together with experimental data, especially small RNA libraries from Next-Generation Sequencing (NGS) technologies.

2. HuntMi: an efficient and taxon-specific approach in pre-miRNA identification

(Gudyś *et al*., 2013)

Currently available *de novo* miRNA search methods suffer from some methodological drawbacks and serious limitations in usage. For instance, the tools perform satisfactorily on data from model species only; training dataset is used in testing procedure; a single machine learning method is tested; low-quality positive and negative datasets are used; finally, the imbalance problem between the size of positive and negative sets usually is not addressed properly, or is ignored, which results in overlearning a majority class and misjudging the classifier performance.

When creating HuntMi, a novel tool for *de novo* miRNA search, we took measures to address these problems and achieve high sensitivity and specificity at the same time. First of all, we made sure that the input data for computations is of high quality. To achieve this, the positive datasets were composed of experimentally verified, up-to-date miRNAs, while negative ones were extracted randomly from genomes and transcriptomes of the corresponding species; sequences bearing even a slight similarity to known miRNAs were discarded. We carefully examined four machine learning methods (naïve Bayes, multilayer perceptron, support vector machine, and random forests); each method was tested with a combination of input parameters to find the settings that best fit miRNA classification problem. We selected random forests as an approach yielding best balance between specificity and sensitivity and this method was applied in the following computations. Next, seven new features for data representation were introduced, besides 21 features previously used in microPred (Batuwita and Palade, 2009). We show that the features improve the classification performance; some of them were never used before in miRNA classification task and they possibly represent biologically relevant features of miRNAs. We also took into account the class imbalance problem by implementing a procedure of thresholding score function that is returned by a classifier score function. This strategy, named ROC-select, turned out to be superior to other imbalance-suited techniques, at least in miRNA classification field.

We compared the performance of our method with leading *de novo* miRNA search tools: microPred (Batuwita and Palade, 2009), PlantMiRNAPred (Xuan *et al*., 2011), MiRenSVM (Ding *et al*., 2010), and it outperforms all of them. Further, we developed the method into a freely available tool named HuntMi. A distinctive feature of HuntMi is its flexibility, as it can be used for plants, animals and viruses. There is also possibility to easily train new classifiers on user provided datasets prior to classification analysis.

3. miRNEST database: an integrative approach in microRNA search and annotation (Szcześniak *et al*., 2012)

Encouraged by the fact that miRNAs are represented in ESTs and that there are hundreds of species with > 10 000 ESTs in dbEST database (Boguski *et al*., 1993), we decided to develop an efficient, homology-based miRNA search method and perform a large scale analysis in a wide array of species. Finally, in order to make the results available for the scientific community, we built an on-line database.

There are several major steps in developed by us miRNA search pipeline that we called *miRNEST algorithm*: i) looking for candidate ESTs by similarity search against known mature miRNAs; ii) assembling the ESTs into contigs; iii) removal of tRNAs and rRNAs; iv) removal of sequences bearing > 60% of low-complexity regions; v) secondary structure checkpoint; vi) removal of miRNA candidates that are similar to known proteins; vii) filtering by hairpin length (animal sequences). Using the pipeline we identified 10,004 miRNA candidates in 221 animal and 199 plant species. Predictions done with *miRNEST algorithm* were complemented with miRNA sequences from external resources: miRBase (Kozomara and Griffiths-Jones, 2009), PMRD (Zhang *et al*., 2010), microPC (Mhuantong and Wichadakul, 2009) and two publications (Huang *et al*., 2009; Hao *et al*., 2010). We run a BLAST search - each sequence against each other - in order to find similarities across stored datasets. In the next step we downloaded 192 small RNA libraries from 29 plant and animal species (based on data availability) from GEO database (Barrett *et al*., 2011) and aligned them to pre-miRNAs stored in miRNEST using Bowtie (Langmead *et al*., 2009). Additional miRNA-associated data was downloaded from 13 resources, including miRTarBase (Hsu *et al*., 2010), Phenomir (Ruepp *et al*., 2010), dPORE-miRNA (Schmeier *et al*., 2011), or Patrocles (Hiard *et al*., 2010).

A high level of complementarity with targeted mRNA sequences usually characterizes plant mature miRNAs and therefore target search in plants is a far less challenging task than in animals, where the evolutionary conservation of miRNA target sites is required to obtain plausible target candidates. Such data is unavailable for a majority of analysed animal species, thus the target search was only performed for plant miRNAs using in-house script, while in case of animals, external data was used. Altogether, we identified targets for 6 963 mature miRNAs in 187 plant species.

We incorporated the abovementioned data into a newly created miRNEST database. The web interface of the database is divided into five sections to help navigate through different data types and structures. *Browse* section gives direct access to all miRNA sequences stored in miRNEST, namely miRNEST predictions and miRNAs from external resources. In *Search*, a number of search options grant the possibility to filter data by user-provided parameters, like hairpin length, mature miRNA sequence or miRNA source. *Unclassified* section provides miRNEST predictions that were not classified as potential miRNAs because they violated at least one of the following criteria: E-value for BLASTX search against UniProt > 1e-20 or pre-miRNA length for animal candidate ≤ 215 nucleotides. *RNA-Seq* component contains small RNA deep sequencing results aligned to predicted pre-miRNAs. Finally, *Taxonomy* provides users with a phylogenetic tree of all species with predicted by us miRNAs. By clicking on the taxon, one can access more detailed data on taxon-specific miRNA families and links to corresponding miRNEST records.

Currently miRNEST undergoes a major update:

i) We developed a pipeline for miRNA discovery in a genomic scale using small RNA libraries. The algorithm performs multiple filtering steps to obtain high-quality candidates; in particular, much attention is paid at the profile of reads mapped to the hairpin. Using this approach, we predicted hundreds of novel miRNAs in 21 plant and animal species.

ii) We modified the abovementioned pipeline to search for mirtrons, i.e. miRNAs with their pre-miRNA sequence spanning the entire intron. We identified 128 mirtron candidates in twelve animal species.

iii) We analysed degradomes from ten plant species using PAREsnip (Folkes *et al*., 2012) to identify experimentally supported miRNA targets. Altogether, we found 1931 miRNA-target associations.

iv) We used HuntMi to analyze all hairpins stored in miRNEST. Each sequence was assigned "-1" (not a miRNA) or "1" (true miRNA).

v) miRNA gene structures will be added to miRNEST. Here, ERISdb predictions will be used (five species) and complemented with predictions for five more plant species: *Brachypodium distachyon, Malus domestica, Medicago truncatula, Populus trichocarpa,* and *Solanum lycopersicum.*

4. ERISdb: a database of plant splice sites and splicing signals (Szcześniak *et al*., 2013)

It becomes more and more clear that in order to understand miRNA biology and apply them in biotechnology and molecular biology it might be necessary to find out more about miRNA gene structures, including alternative splice forms. Unfortunately, almost all miRNA search studies are concentrated on pre-miRNA and/or mature miRNA prediction. As a result, little has been done to determine miRNA gene structures in plants, except for single analyses in *Arabidopsis thaliana* (Szarzynska *et al*. 2009), *Vitis vinifera* (Mica *et al*. 2010) and very recently in *Hordeum vulgare* (Kruszka *et al*., 2013). By contrast, animal miRNAs are generally thought to be devoid of introns. Keeping in mind the insufficiency of our knowledge about miRNA gene structures, we performed large-scale splice site prediction in miRNA genes using EST sequences in seven plant species: *Arabidopsis thaliana, Chlamydomonas reinhardtii, Glycine max, Oryza sativa, Physcomitrella patens, Selaginella moellendorffii,* and *Zea mays*.

In our pipeline, in the first step we searched for EST sequences that correspond to known pre-miRNAs from miRBase (Kozomara and Griffiths-Jones, 2011). We screened the ESTs from dbEST database (Boguski *et al*., 1993) using Megablast (Altschul *et al*., 1990) and required that the identity is 97% or more over at least 90% of annotated pre-miRNA sequence. The selected ESTs were subsequently mapped to the plant genome using Splign (Kapustin *et al*., 2008). For *Arabidopsis thaliana* we downloaded the sequences from RACE experiments (Szarzynska *et al*., 2009) and used a similar approach as in case of ESTs. Finally, for *Vitis vinifera* we downloaded three miRNAs with RNA-Seq support for introns (Mica *et al*., 2010).

Altogether, we identified introns in 45 miRNAs in five plant species. Some of the miRNAs contain multiple introns (up to six), there are also several cases of alternative splicing via intron retention. Additionally, 8 miRNAs with annotated introns from Ensembl (Kersey *et al*., 2010) and 3 miRNAs with RNA-Seq support were incorporated. In the *miRNA gene structures* section of ERISdb, a new database of plant splice sites and splice signals, one can see the splice site predictions as alignment of three sequences: genomic DNA, EST, and pre-miRNA sequence. In case of eight Ensembl miRNAs in *A. thaliana*, the user is redirected to *splice site data* page in ERISdb, while *V. vinifera* miRNAs with RNA-Seq support are presented as alignment of reads to the splice sites.

5. microRNA databases (original title: **Bazy danych mikroRNA**) (Szcześniak *et al*., 2012)

Development of miRNA search methods, both experimental and computational, resulted in rapid accumulation of miRNA data and need for dedicated databases. miRBase was one of the very first of them and today it is considered as a reference database that stores miRNAs from 67 plant and 97 animal species as well as 26 viruses. PMRD (Plant MicroRNA Database), microPC, and miRNEST are other resources of miRNA sequences. There are also databases that store other miRNA-associated data, like expression profiles, targets, polymorphisms, and many more, summing up to about sixty databases that are to our disposal nowadays.

As it becomes more and more difficult to find miRNA data of interest in the fast expanding realm of biological databases, we wrote a review about miRNA databases. In the review we described several representative databases and shortly characterized all available 51 miRNA databases (as for November 2011). Additionally, we considered the sources of miRNA data and concerns about data quality, database design and functionality. The conclusion was that there is need for large, integrative resources rather than small databases dedicated for a limited group of specialists.

6. References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990; 215:403-410.
2. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res. 2011; 39:D1005–D1010.
3. Batuwita R, Palade V. MicroPred: effective classification of pre-miRNAs for human MiRNA gene prediction. Bioinformatics 2009, 25:989–995.
4. Boguski MS, Lowe TM, Tolstoshev CM. dbEST—database for "expressed sequence tags" Nat Genet. 1993;4:332–333.
5. Ding J, Zhou S, Guan J: MiRenSVM. towards better prediction of MicroRNA precursors using an ensemble SVM classifier with multi-loop features. BMC Bioinformatics 2010, 11 (Suppl 11):S35.
6. Folkes L, Moxon S, Woolfenden HC, Stocks MB, Szittya G, Dalmay T, Moulton V. PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. Nucleic Acids Res. 2012; 40(13):e103.
7. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet. 2009;10(2):94-108.
8. Gudyś A, Szcześniak MW, Sikora M, Makalowska I. HuntMi: an efficient and taxon-specific approach in pre-miRNA identification BMC Bioinformatics 2013, 14:83.
9. Hao L, Cai P, Jiang N, Wang H, Chen Q. Identification and characterization of microRNAs and endogenous siRNAs in Schistosoma japonicum. BMC Genomics 2010; 11:55.
10. Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. Nucleic Acids Res. 2010; 38:D640–D651.
11. Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM, Chien CH, Wu MC, Huang CY, Tsou AP, Huang HD. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. 2010;39:D163–D169.
12. Huang J, Hao P, Chen H, Hu W, Yan Q, Liu F, Han ZG. Genome-wide identification of Schistosoma japonicum microRNAs using a deep-sequencing approach. PLoS One 2009;

4:e8206.

13. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 2009; 37: D98-104.
14. Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol. Direct 2008; 3:20.
15. Kersey PJ, Staines DM, Lawson D, Kulesha E, Derwent P, Humphrey JC, Hughes DS, Keenan S, Kerhornou A, Koscielny G, Langridge N, McDowall MD, Megy K, Maheswari U, Nuhn M, Paulini M, Pedro H, Toneva I, Wilson D, Yates A, Birney E. Ensembl Genomes: extending Ensembl across the taxonomic space. Nucleic Acids Res. 2010; 38:D563-D569.
16. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011; 39:D152–D157.
17. Kruszka K, Pacak A, Swida-Barteczka A, Stefaniak AK, Kaja E, Sierocka I, Karlowski W, Jarmolowski A, Szweykowska-Kulinska Z. Developmentally regulated expression and complex processing of barley pri-microRNAs. BMC Genomics 2013; 14:34.
18. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25.
19. Mallory AC, Vaucheret H. Functions of microRNAs and related small RNAs in plants. Nat Genet. 2006;38 Suppl:S31-6. Erratum in: Nat Genet. 2006 Jul;38(7):850.
20. Mhuantong W, Wichadakul D. MicroPC (microPC): A Comprehensive resource for predicting and comparing plant MicroRNAs. BMC Genomics 2009, 10: 366.
21. Mica E, Piccolo V, Delledonne M, Ferrarini A, Pezzotti M, Casati C, Del Fabbro C, Valle G, Policriti A, Morgante M, Pesole G, Pè ME, Horner DS. High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera. BMC Genomics 2009; 10:58.
22. Nair V, Zavolan M. Virus-encoded microRNAs: novel regulators of gene expression. Trends Microbiol. 2006; 14(4):169-75.
23. Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ. PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. Genome Biol. 2010;11:R6.
24. Schmeier S, Schaefer U, MacPherson CR, Bajic VB. dPORE-miRNA: polymorphic regulation of microRNA genes. PLoS One. 2011; 6:e16657.

25. Siomi H, Siomi MC. Posttranscriptional regulation of microRNA biogenesis in animals. Mol Cell. 2010; 38(3):323-32.
26. Sunkar R, Chinnusamy V, Zhu J, Zhu JK. Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. Trends Plant Sci. 2007;12(7):301-9.
27. Szarzynska B, Sobkowiak L, Pant BD, Balazadeh S, Scheible WR, Mueller-Roeber B, Jarmolowski A, Szweykowska-Kulinska Z. Gene structures and processing of Arabidopsis thaliana HYL1-dependent pri-miRNAs. Nucleic Acids Res. 2009; 37:3083-3093.
28. Szcześniak MW, Deorowicz S, Gapski J, Kaczyński Ł, Makalowska I. miRNEST database: an integrative approach in microRNA search and annotation. Nucleic Acids Res. 2012; 40:D198-204.
29. Szcześniak MW, Kabza M, Pokrzywa R, Gudyś A, Makałowska I. ERISdb: a database of plant splice sites and splicing signals. Plant Cell Physiol. 2013; 54(2):e10.
30. Szcześniak MW, Owczarkowska E, Gapski J, Makałowska I. Bazy danych mikroRNA. Postepy Bioch. 2012; 58(1).
31. Vaucheret H, Voinnet O, Watanabe Y, Weigel D, Zhu JK. Criteria for annotation of plant MicroRNAs. Plant Cell 2008; 20(12):3186-90.
32. Xuan P, Guo M, Liu X, Huang Y, Li W, Huang Y. PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. Bioinformatics 2011, 27: 1368–1376.
33. Zhang Z, Yu J, Li D, Zhang Z, Liu F, Zhou X, Wang T, Ling Y, Su Z. PMRD: plant microRNA database. Nucleic Acids Res. 2010; 38:D806–813.