Rozprawa doktorska 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

Składam serdeczne podziękowania:

Pani prof. UAM dr hab. Izabeli Makałowskiej
za owocną współpracę w Pracowni Bioinformatyki, przekazaną wiedzę i doświadczenie, dobre rady, cierpliwość i życzliwość

Panu prof. dr hab. Włodzimierzowi J. Krzyżosiakowi
oraz Panu prof. dr hab. Piotrowi Zielenkiewiczowi
za wysiłek włożony w recenzowanie niniejszej pracy

Panu prof. dr hab. Bogdanowi Jackowiakowi, Dziekanowi Wydziału Biologii

Dyrekcji, Pracownikom i Doktorantom Instytutu Biologii Molekularnej i Biotechnologii

Koleżankom i Kolegom z Pracowni Bioinformatyki
w szczególności mgr Joannie Ciomborowskiej, mgr Michałowi Kabzie, mgr Elżbiecie Kaja i mgr Wojciechowi Rosikiewiczowi

a także wszystkim innym osobom, które w jakikolwiek sposób wspierały mnie na dotychczasowych etapach rozwoju naukowego.

Szczególne wyrazy wdzięczności kieruję
do moich Rodziców oraz Narzeczonej.

SPIS TREŚCI

NOWE METODY IDENTYFIKACJI mikroRNA

I. STRESZCZENIE

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.

II. SUMMARY

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.

OŚWIADCZENIE DOKTORANTA DOTYCZĄCE JEGO UDZIAŁU W POWSTANIU PRAC NAUKOWYCH STANOWIĄCYCH ROZPRAWĘ DOKTORSKĄ

III. OŚWIADCZENIE DOKTORANTA

Tytuł publikacji: *miRNEST database: an integrative approach in microRNA search and annotation*
Rola doktoranta: pierwszy autor

Wykonane przez doktoranta prace to przede wszystkim:

i) opracowanie szczegółowej koncepcji analiz;
ii) przygotowanie danych wejściowych do analiz;
iii) stworzenie potoku analitycznego służącego do identyfikacji zakonserwowanych mikroRNA w sekwencjach EST;
iv) wykorzystanie potoku analitycznego do wielkoskalowej analizy sekwencji EST u kilkuset gatunków roślin i zwierząt;
v) implementacja w języku Python algorytmu służącego do identyfikacji miejsc docelowych miRNA;
vi) zmapowanie sekwencji małych RNA do sekwencji pre-miRNA oraz obróbka wyników mapowania;
vii) zaprojektowanie bazy danych i jej implementacja w systemie mySQL; umieszczenie danych w bazie danych;
viii) opracowanie interfejsu dla bazy danych;
ix) główna rola w przygotowaniu manuskryptu;
x) aktualizacja i rozbudowa bazy danych.

Tytuł artykułu: *ERISdb: a database of plant splice sites and splicing signals*
Rola doktoranta: pierwszy autor

Zadania wykonane przez doktoranta (wymieniono tylko te, które dotyczą przewidywania struktury genów miRNA):

i) zaprojektowanie i zbudowanie potoku analitycznego służącego do przewidywania miejsc splicingowych z wykorzystaniem sekwecji genomowych i EST;
ii) przygotowanie danych wejściowych;
iii) użycie potoku analitycznego w wielkoskalowej analizie;
iv) umieszczenie wyników analizy w bazie danych ERISdb;
v) pobranie danych o strukturze genów miRNA z zewnętrznych źródeł, ich obróbka oraz umieszczenie w ERISdb;
vi) zaprojektowanie strony internetowej w bazie danych ERISdb służącej do wizualizacji struktur genów miRNA.

Tytuł artykułu: *HuntMi: an efficient and taxon-specific approach in pre-miRNA identification*
Rola doktoranta: dzielone pierwsze autorstwo z mgr inż. Adamem Gudysiem

Wykonane przez doktoranta prace:

i) współudział w opracowaniu szczegółowej koncepcji analiz;
ii) przygotowanie danych wejściowych do analiz;
iii) opracowanie siedmiu nowych cech wykorzystywanych przez klasyfikator mikroRNA; stworzenie narzędzi służących do obliczania tych cech;
iv) konsultacje w sprawie eksperymentów kroswalidacyjnych;
v) udział w tworzeniu HuntMi jako spójnego narzędzia służącego do identyfikacji mikroRNA, w tym napisanie skryptów pozwalających na integrację części składowych programu;
vi) uczestnictwo w analizach mających na celu porównanie HuntMi z wybranymi narzędziami do identyfikacji miRNA *de novo*;
vii) udział w pisaniu manuskryptu, a w szczególności rozdziałów *Background*, *Methods* (podrozdziały *Datasets* i *Features*), *Results and discussion* (podrozdział *Comparison with other tools*).

Tytuł artykułu: *Bazy danych mikroRNA*
Rola doktoranta: pierwszy autor

Zadania wykonane przez doktoranta:

i) szukanie dostępnych, opublikowanych baz danych miRNA;
ii) wybór baz danych do bardziej szczegółowego opisu;
iii) napisanie rozdziałów: *Wprowadzenie, Źródła informacji w bazach danych miRNA*, *Zautomatyzowane przeszukiwanie i pobieranie danych*;
iv) pomoc przy pisaniu i redagowaniu pozostałych części manuskryptu.

....

data podpis

OŚWIADCZENIA WSPÓŁAUTORÓW DOTYCZĄCE ICH UDZIAŁU W POWSTANIU PRAC NAUKOWYCH STANOWIĄCYCH ROZPRAWĘ DOKTORSKĄ

IV. OŚWIADCZENIA WSPÓŁAUTORÓW

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Biologii
Instytut Biologii Molekularnej i Biotechnologii

Poznań, 27.03.2012

W związku z wykorzystaniem przez mgr Michała Szcześniaka poniżej wymienionych publikacji jako rozprawy doktorskiej oświadczam, iż udział mój, jako promotora, polegał przede wszystkim na wspólnym opracowywaniu koncepcji badań oraz przygotowywaniu manuskryptów. Jednocześnie stwierdzam, iż we wszystkich wymienionych publikacjach wkład pracy mgr Michał Szcześniak był niezwykle duży, pełnił w nich wiodącą rolę i bezsprzecznie znacząco przyczynił się on do ich powstania.

Gudyś A, Szcześniak MW, Sikora M, Makalowska I. HuntMi: an efficient and taxon-specific approach in pre-miRNA identification (2013) BMC Bioinformatics. 2013 Mar 5;14(1):83.

Szcześniak MW, Kabza M, Pokrzywa R, Gudyś A, Makałowska I. ERISdb: a database of plant splice sites and splicing signals (2013) Plant Cell Physiol. 2013 Feb;54(2):e10. doi: 10.1093/pcp/pct001.

Szcześniak MW., Owczarkowska E., Gapski J., Makałowska I; Bazy danych microRNA (2012) Postępy Biochemii 58(1); 91-98

Szcześniak, MW., Deorowicz, S., Gapski, J., Kaczyński, Ł., Makałowska, I.; miRNEST database: an integrative approach in microRNA search and annotation (2012) Nucleic Acids Research 40:D198-204

Z poważaniem,

Izabela Makałowska

ul. Umultowska 89, Collegium Biologicum, 61-614 Poznań
NIP 777 00 06 350, REGON 000001293
tel. +48 61 829 59 50, fax. +48 61 829 59 49
ibmib@amu.edu.pl

www.ibmib.amu.edu.pl

Gliwice, 06.03.2013

mgr inż. Adam Gudyś
Instytut Informatyki
Politechnika Śląska

Oświadczam, że mój udział w powstaniu artykułu "HuntMi: an efficient and taxon-specific approach in pre-miRNA identification" obejmował opracowanie metody klasyfikacji danych niezrównoważonych, przeprowadzenie eksperymentów kroswalidacyjnych oraz analizę statystyczną ich wyników, udział w eksperymentach porównawczych z innymi narzędziami, a także udział w pisaniu manuskryptu.

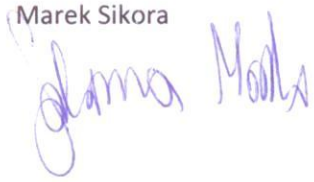
Adam Gudyś

Gliwice, 06.03.2013

dr Marek Sikora
Instytut Informatyki
Politechnika Śląska

Mój udział w pracach nad artykułem "HuntMi: an efficient and taxon-specific approach in pre-miRNA identification" obejmował nadzorowanie implementacji klasyfikatora do danych niezrównoważonych, udział w analizie statystycznej wyników eksperymentalnych oraz sprawdzenie manuskryptu.

Marek Sikora



Dr hab. inż. Sebastian Deorowicz, prof. nzw. Pol. Śl.
sebastian.deorowicz@polsl.pl

Gliwice, 15.03.2013 r.

OŚWIADCZENIE

Oświadczam, że jestem współautorem pracy pt. „miRNEST database: an integrative approach in microRNA search and annotation" opublikowanej w czasopiśmie „Nucleic Acids Research". Mój udział w powstaniu niniejszej pracy polegał na:
- implementacja w języku C++ algorytmu do szukania miejsc docelowych dla roślinnych miRNA.

Sebastian Deorowicz

Czytelny podpis

.....

Łukasz Kaczyński
kaczynskiluk@gmail.com
tel. +48 795 738 001

Poznań, 15.03.2013 r.

OŚWIADCZENIE

Oświadczam że jestem współautorem pracy pt. „miRNEST database: an integrative approach in microRNA search and annotation." opublikowanej w czasopiśmie „Nucleic Acids Research". Mój udział w powstaniu niniejszej pracy polegał na:
- wykonaniu pilotażowych analiz w małej skali skupionych na wstępnej ocenie efektywności potoku analitycznego.

Łukasz Kaczyński

Czytelny podpis

.....

Łukasz Kaczyński

Jakub Gapski
Poznań, 06.03.2013 r.
Jakub.gapski@gmail.com
tel. +48 61 829 5836

OŚWIADCZENIE

Oświadczam że jestem współautorem pracy pt. „miRNEST database: an integrative approach in microRNA search and annotation" opublikowanej w czasopiśmie „Nucleic Acids Research". Mój udział w powstaniu niniejszej pracy polegał na stworzeniu modułu pozwalającego na automatyczną aktualizację zawartości bazy danych miRNEST, wykorzystując dane z bazy miRBase.

Jakub Gapski

Czytelny podpis

.....

Gliwice, 06.03.2013

mgr inż. Adam Gudyś
Instytut Informatyki
Politechnika Śląska

Oświadczam, że mój udział w pracach nad artykułem "ERISdb: A Database of Plant Splice Sites and Splicing Signals" obejmował zaimplementowanie klasyfikatora do identyfikacji intronów U12 oraz ocenę istotności statystycznej nadreprezentacji potencjalnych sekwencji regulatorowych w splicingu.

Adam Gudyś

Gudyś

dr inż. Rafał Pokrzywa
rafal.pokrzywa@polsl.pl
tel. +48 602 363 999

Gliwice, 07.03.2013 r.

OŚWIADCZENIE

Oświadczam że jestem współautorem pracy pt. „ERISdb: A Database of Plant Splice Sites and Splicing Signals" opublikowanej w czasopiśmie „Plant and Cell Physiology" nr 54(2), 2013. Mój udział w powstaniu niniejszej pracy polegał na:

- zaprojektowaniu oraz napisaniu programu komputerowego w języku Java służącego do wyszukiwania nadreprezentowanych motywów w sekwencjach intronowych,
- konsultacjach w zakresie wykorzystania opracowanego programu komputerowego do wyszukiwania elementów regulatorowych splicingu,
- napisaniu fragmentu manuskryptu dotyczącego opisu metodologii obliczeniowej programu komputerowego do wyznaczania nadreprezentowanych motywów.

Rafał Pokrzywa

Rafał Pokrzywa

.....

Czytelny podpis

mgr Michał Kabza
Poznań, 06.03.2013 r.
mkabza@amu.edu.pl
tel. +48 61 829 5836

OŚWIADCZENIE

Oświadczam że jestem współautorem pracy pt. "ERISdb: A Database of Plant Splice Sites and Splicing Signals" opublikowanej w czasopiśmie „Plant and Cell Physiology". Mój udział w powstaniu niniejszej pracy polegał na potwierdzeniu istnienia zaadnotowanych intronów w genomach roślinnych za pomocą technik RNA-Seq.

Michał Kabza

Czytelny podpis

.....Michał Kabza.....

mgr Elzbieta Kaja (Owczarkowska)
eo@amu.edu.pl
tel. +48 61 829 5836

Poznań, 05.03.2013 r.

OŚWIADCZENIE

Oświadczam że jestem współautorem pracy pt. „Bazy danych miRNA." opublikowanej w czasopiśmie *„Postępy Biochemii", Tom 58, Nr 1/2012.* Mój udział w powstaniu niniejszej pracy polegał na:

- przeglądzie i analizie dostępnych baz danych mikro RNA,
- konsultacjach naukowych,
- przygotowaniu tabeli porównującej bazy danych mikro RNA

Elżbieta Kaja

Czytelny podpis

.....

Elżbieta Kaja

Jakub Gapski
Poznań, 06.03.2013 r.
Jakub.gapski@gmail.com
tel. +48 61 829 5836

OŚWIADCZENIE

Oświadczam że jestem współautorem pracy pt. „Bazy danych mikroRNA" opublikowanej w czasopiśmie „Postępy Biochemii". Mój udział w powstaniu niniejszej pracy polegał na współudziale w przygotowaniu opisów baz danych miRBase, miRNEST, miRecords, miR2Disease oraz PhenomiR.

Jakub Gapski

Czytelny podpis

.....

V. ROZPRAWA DOKTORSKA

BMC Bioinformatics

BioMed Central
The Open Access Publisher

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

BMC Bioinformatics 2013, **14**:83 doi:10.1186/1471-2105-14-83

Adam Gudy¿ (adam.gudys@polsl.pl)
Micha¿ Wojciech Szcze¿niak (miszcz@amu.edu.pl)
Marek Sikora (marek.sikora@polsl.pl)
Izabela Makalowska (izabel@amu.edu.pl)

ISSN 1471-2105

Article type Methodology article

Submission date 2 July 2012

Acceptance date 21 February 2013

Publication date 5 March 2013

Article URL http://www.biomedcentral.com/1471-2105/14/83

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

Adam Gudyś [1*]
*Corresponding author
Email: adam.gudys@polsl.pl

Michał Wojciech Szcześniak[2*]
*Corresponding author
Email: miszcz@amu.edu.pl

Marek Sikora[1,3]
Email: marek.sikora@polsl.pl

Izabela Makałowska[2]
Email: izabel@amu.edu.pl

[1]Institute of Informatics, Faculty Of Automatic Control, Electronics And Computer Science, Silesian University of Technology, Gliwice, Poland

[2]Laboratory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland

[3]Institute of Innovative Technologies EMAG, Katowice, Poland

Abstract

Background

Machine learning techniques are known to be a powerful way of distinguishing microRNA hairpins from pseudo hairpins and have been applied in a number of recognised miRNA search tools. However, many current methods based on machine learning suffer from some drawbacks, including not addressing the class imbalance problem properly. It may lead to overlearning the majority class and/or incorrect assessment of classification performance. Moreover, those tools are effective for a narrow range of species, usually the model ones. This study aims at improving performance of miRNA classification procedure, extending its usability and reducing computational time.

Results

We present HuntMi, a stand-alone machine learning miRNA classification tool. We developed a novel method of dealing with the class imbalance problem called ROC-select, which is based on thresholding score function produced by traditional classifiers. We also introduced new features to the data representation. Several classification algorithms in combination with ROC-select were tested and random forest was selected for the best balance between sensitivity and specificity. Reliable assessment of classification performance is guaranteed by using large, strongly imbalanced, and taxon-specific datasets in 10-fold cross-validation procedure. As a result, HuntMi achieves a considerably better performance than any other miRNA classification tool and can be applied in miRNA search experiments in a wide range of species.

Conclusions

Our results indicate that HuntMi represents an effective and flexible tool for identification of new microRNAs in animals, plants and viruses. ROC-select strategy proves to be superior to other methods of dealing with class imbalance problem and can possibly be used in other machine learning classification tasks. The HuntMi software as well as datasets used in the research are freely available at http://lemur.amu.edu.pl/share/HuntMi/.

Keywords

MicroRNA, Random forest, Imbalanced learning, Genome analysis

Background

MicroRNAs (miRNAs) are ∼21 bases long RNAs that post-transcriptionally control multiple biological processes, such as development, hematopoiesis, apoptosis and cell proliferation [1]. Mature miRNAs are derived from longer precursors called pre-miRNAs that fold into hairpin structures containing one or more mature miRNAs in one or both arms [2]. Their biogenesis is highly regulated at both transcriptional and post-transcriptional levels [3], and disregulation of miRNAs is linked to various human diseases, including cancer [4].

Identification of miRNA is a challenging task that allows us to better understand post-transcriptional regulation of gene expression. In last ten years a number of experimental and computational approaches were proposed to deal with the problem. However, experimental approaches, including direct cloning and Northern blot, are usually able to detect only abundant miRNAs. MicroRNAs that are expressed at very low levels or in a tissue- or stage-specific manner, often remain undetected. These problems are partially addressed by applying the deep-sequencing techniques that nevertheless require extensive computational analyses to distinguish miRNAs from other non-coding RNAs or products of RNA degradation [5].

Computational approaches in miRNA search can be homology-based, take advantage of machine learning methods, or use both of these. Homology-based approaches rely on conservation of sequences, secondary structures or miRNA target sites (e.g. RNAmicro [6], MIRcheck [7]). As a result, these methods are not suitable for detection of lineage- or species-specific miRNAs and miRNAs that evolve rapidly. Moreover, they are strongly limited by the current data and performance of available computational methods, including alignment algorithms [8]. Another problem is that there are as many as ∼11 million sequences that can fold into miRNA-like hairpins in the human genome [9], some of which originate from functional, non-miRNA loci. It is therefore no surprise that a large number of hairpins that are conserved between species could be mistakenly classified as miRNAs. Nevertheless, homology search has been successfully applied in many miRNA gene predictions, in both animals and plants [10, 11].

In some approaches, e.g. PalGrade [12] or miRDeep [5], experimental and computational procedures are combined. However, as mentioned above, experimental methods can not easily detect low-expression or tissue-specific miRNAs and/or they have to meet computational challenges, as in the case of deep sequencing technology. miRDeep, for instance, aligns deep sequencing reads to the genome and selects the regions that can form a hairpin structure. Then, using a probabilistic model, the hairpins are scored based on the compatibility of the position and frequency of sequenced reads with the secondary structure of the pre-miRNA. This method achieves high specificity at the cost of relatively low sensitivity.

Machine learning methods are amongst the most popular ways of miRNA identification nowadays. They share the same overall strategy. First, the features of primary sequence and secondary structure are extracted from known miRNAs (positive set) and non-miRNA sequences (negative set). Then, the features are used to construct a model which serves to classify candidate sequences as real pre-miRNAs or pseudo pre-miRNAs. There are several machine learning methods that have been applied in the field of miRNA identification. These include hidden Markov models (HMM) [13], random forest [14] and naïve Bayes classifier [15]. Support vector machine, however, seems to be the most popular framework nowadays and has been used in a number of well recognised tools. For instance, Triplet-SVM [16] classifies real human pre-miRNAs and pseudo pre-miRNAs using 32 structure- and sequence-derived features that refer to the dot-bracket representation of the secondary structure i.e. it considers the frequencies of triplets, such as "A(((" and "U.(.", consisting of the secondary structure of three adjacent nucleotides and the nucleotide in the middle. miPred [8] classifies human pre-miRNAs from pseudo hairpins represented by twenty nine folding features, using SVM-based approach. The features were evaluated with the F scores F1 and F2 on the class-conditional distributions to assess their discriminative power. Strongly correlated attributes were rejected. microPred [17] presents nineteen new features along with twenty nine taken from miPred. After feature selection, twenty one attributes were used to train the classifier. The improved feature selection approach and addressing the class imbalance problem resulted in high sensitivity and specificity of the method.

However, the existing machine learning approaches suffer from some drawbacks. First of all, they often make structural assumptions concerning stem length, loop size and numbers as well as a minimum free energy (MFE). Secondly, most of existing miRNA classifiers work well on data from model species and closely related ones; the classifiers trained on human data best fit the miRNA identification problem in human and other primates but perform unsatisfactorily when applied to, for example, invertebrates. Finally, the imbalance problem between the positive and negative classes is usually not addressed properly, while this is a crucial issue, as the number of microRNAs throughout a genome is much lower than the number of non-microRNAs (e.g. $\sim$1 400 miRNAs vs. $\sim$11 million pseudo hairpins in *H. sapiens*). The resulting difference in misclassification costs of positive and negative classes requires special techniques of learning from imbalanced data as well as a proper assessment metrics. Moreover, in order to accurately judge classifier performance in real-life applications, the problem of imbalance should be reflected in the testing datasets.

In this study we addressed all these issues. We made no preliminary assumptions about miRNA structure and carefully took into account class imbalance problem. We implemented a procedure of thresholding score function produced by traditional classifiers and called it ROC-select. This strategy turned out to be superior to other imbalance-suited techniques in miRNA classification. From all classifiers for which ROC-select procedure was applied we chose random forest as it yields the best balance between sensitivity and specificity. Regarding the data representation, we introduced seven new features and show that they further improve the classification performance. In the experiments we considered large and strongly imbalanced up-to-date sets of positive and negative examples, paying much attention to the data quality. The tests were performed using stratified 10-fold cross-validation (CV) giving reliable estimates of classification performance. Finally, we show that the method outperforms the existing miRNA classification tools, including microPred, without compromising the computational time.

Our miRNA classification method is freely available as a framework called HuntMi. HuntMi comes with trained models for animals, plants, viruses and separately for *H. sapiens* and *A. thaliana*. As a result, the tool can be used in miRNA classification experiments in a wide range of species. The user can use built-in models in the experiments or train new models using custom datasets prior to classification.

Methods

Datasets

In order to create positive sets, we retrieved all pre-miRNAs from miRBase release 17 [18] and filtered out the sequences lacking experimental confirmation. By using evidence-supported miRNAs only, we minimize the chance of introducing false positives into the set. The sequences were divided into five groups: *H. sapiens*, *A. thaliana*, animals, plants, and viruses.

Negative sets were extracted from genomes and mRNAs of ten animal and seven plant species as well as twenty nine viruses (Additional file 1: Table S1). Additional sets were prepared for *H. sapiens* and *A. thaliana*. Start positions were randomly selected, whereas end positions were calculated so that the sequence length distribution in the resulting negative dataset is the same as in the corresponding positive one. With this approach, the classifier achieves better performance when applied in real-life experiments, where miRNA candidates tend to have lengths similar to those of known miRNAs. Finally, in order to remove known miRNAs together with similar sequences that possibly represent unknown homologs of miRNAs, we ran BLASTN search against miRBase hairpins and filtered out sequences that produced E-value of 10^{-2} or lower. 96.17% of negative sequences prepared in this way possess structural features of real pre-microRNAs, including the minimum free energy below -0.05 (normalised to the sequence length) and number of pairings in the stem above 0.15 (also normalised to the length). At the same time these criteria are met by 97.61% of hairpins stored in miRBase.

Positive and negative sequences from the analysed species were gathered to form complete datasets that correspond to miRNA classification problem in the taxa. They will be referred to as *human*, *arabidopsis*, *animal*, *plant* and *virus* (Table 1). In addition, we used the dataset from microPred. It contains 691 non-redundant human pre-miRNAs from miRBase release 12, 754 non-miRNA ncRNA, 8 494 pseudo hairpins and is denoted as *microPred*.

Table 1 Datasets characteristics

Name	#Positives	#Negatives	Imbalance
human	1 406	81 228	57.8
arabidopsis	231	28 359	122.8
animal	7 053	218 154	30.9
plant	2 172	114 929	52.9
virus	237	839	3.5
microPred	691	9 248	13.4

Characteristics of biological datasets used in the experiments. Imbalance is defined as a ratio of #Negatives to #Positives. We limited dataset imbalance to several tens for practical reasons even though proportions of miRNAs to non-miRNAs in genomes are more extreme. In the case of virus dataset the imbalance is exceptionally low as we wanted to know how methods perform on moderately imbalanced problems. In addition, it is difficult to create representative dataset for viruses as their genomes differ significantly in sizes and most of them do not contain miRNAs.

Features

The twenty one features selected by [17] were used as a base representation in the experiments. Thus, we employed microPred scripts for extracting necessary attributes. In the case of *microPred* dataset we took precalculated features from webpage to make our results comparable with the existing research (some of the features are calculated using randomly generated sequences).

Beside twenty one microPred features, we calculated seven additional sequence- and structure-related attributes. First, we considered the frequencies of secondary structure triplets composed of three adja-

cent nucleotides and the middle nucleotide. We chose four of them that were shown to have the highest information gain [19]: "A(((", "U(((", "G(((", and "C(((", referred to as *tri_A*, *tri_U*, *tri_G*, and *tri_C*, respectively. The remaining features are: the maximal length of the amino acid string without stop codons found in three reading frames: *orf*; the cumulative size of internal loops found in the secondary structure: *loops*; a percentage of low complexity regions detected in the sequence using Dustmasker: *dm* (all Dustmasker settings were set to default except for score threshold for subwindows set from 20 to 15).

Imbalanced learning

Extensive research on imbalanced data classification has proven that standard machine learning techniques often overlearn a majority class sacrificing minority examples [20]. Therefore, special approaches for imbalanced problems have been developed. They can be divided into sampling methods, cost-sensitive learning, kernel methods, active learning and others [21]. microPred authors carried out exhaustive study of how several classification strategies from above perform in a microRNA prediction task [17]. They used standard support vector machine as a base classifier and combined it with random over/under-sampling, SMOTE (which is also a representative of sampling methods) and multi-classifier system. They additionally tested cost-sensitive SVM modifications like zSVM and DEC (different error costs), finding SMOTE to be the best strategy. In the research, geometric mean (G_m) of classification sensitivity (SE) and specificity (SP) was used as an assessment metric. G_m is common in imbalanced learning problems, including miRNA identification, as it takes into account unequal misclassification costs. Therefore, we also decided to use G_m in HuntMi study.

Our approach to microRNA prediction relies on the fact that classification with unequal costs is equivalent to thresholding conditional class probabilities at arbitrary quantiles [22]. Many classifiers provide continuous score function $s(x)$ describing degree of a membership of instance x to particular class. Ideally, such a function estimates perfectly a class conditional probability $P(c|x)$ and is denoted as well-calibrated score function [23]. In reality, classifiers produce scores which are often not calibrated [22] thus a lot of algorithms for calibrating them have been developed [23]. In addition, many meta learning techniques like bagging or classifier ensembles can be employed to produce score function on the basis of class labels alone [24]. As long as scoring function ranks instances properly, that is $s(x) < s(y) \Leftrightarrow P(c|x) < P(c|y)$, one can successfully use $s(x)$ directly to classify instances with unequal costs.

Our method combines the idea of thresholding classifier score function with receiver operating characteristics (ROC) [25]. For each threshold value T established at $s(x)$ function, a point in a ROC space can be generated. Varying T from $-\infty$ to $+\infty$ produces entire ROC curve. One can select a point on it with highest evaluation metric (G_m in the case) and read corresponding T value. In real applications ROC curves are generated by simply sorting elements of dataset by $s(x)$ values and updating true positive (*TP*) and false positive (*FP*) statistics for consecutive points. In order to prevent threshold selection procedure from overfitting towards training data, a separate set should be used for constructing ROC curve. Hence, an internal cross-validation with k_1 folds is employed for this purpose. As we are not interested in variance, ROC curves are averaged in a straightforward way - instances from all tuning folds together with assigned $s(x)$ values are gathered in a single set on which ROC generation procedure is applied [25]. Threshold leading to the highest value of evaluation metric is stored and used for classification of unknown instances. The threshold selection procedure described above will be referred to as ROC-select.

In the research we apply ROC-select only on classifiers directly providing scoring function, no meta learning techniques were examined. These classifiers are naïve Bayes [26], multilayer perceptron [27],

support vector machine [28] and random forest [29]. We used radial basis function as an SVM kernel as it is known to produce best classification results in wide range of applications [30]. In order to compare proposed strategy with other methods, we additionally tested SMOTE filter [31] combined with SVM as it gave best results in microPred experiments and a novel method of asymmetric partial least squares classification (APLSC), which came out to be superior to other strategies on several strongly imbalanced datasets [32].

Parameter selection and complexity analysis

In many studies including microRNA prediction, classifier parameters are selected in order to obtain best possible results for a particular domain. Hence, we decided to place parameter tuning phase in our pipeline as a preceding step for threshold selection. Parameter selection is also done with an internal cross-validation with a number of folds equal to k_2 and is straightforward. At first, a search space is defined by specifying a number of discrete values for each parameter to be tuned. Then, full cross-validation procedure is performed for each point in that space. Combination of parameter values leading to the highest average evaluation metric (G_m) is stored and used in threshold selection and, finally, for classification of unknown instances.

Let us denote number of points in the parameter space to be examined as λ. In addition, let $L(n)$ and $T(n)$ indicate time complexities of training and testing procedures for given classifier with respect to the dataset size n. ROC-select and parameter tuning are performed in $O(k_1(L(n(k_1-1)/k_1)+T(n/k_1))+n\log n)$ and $O(\lambda k_2(L(n(k_2-1)/k_2)+T(n/k_2)))$ time, respectively. As $(k-1)/k<1$ entire procedure is bounded by expression $O((k_1+\lambda k_2)L(n)+k_1T(n/k_1)+\lambda k_2T(n/k_2)+n\log n)$.

Experimental setting

All classification experiments were carried out using stratified 10-fold CV, hence distributions of testing samples are exactly the same as for the entire datasets. Taking into account strong imbalance of examined sets, obtained results approximate well the expected performance of a classifier in practical applications. Additionally, 10-fold CV was proven to be the best method of model evaluation in terms of bias and variance [33].

The detailed configuration of examined classifiers together with parameter values tested in a tuning phase are listed below (number of points in a parameter space for tuning phase given in parentheses). Parameters not mentioned here remained default.

- naïve Bayes: kernel estimation turned on,
- multilayer perceptron: validation set size $V=20\%$, validation threshold $E=50$, learning rate $\eta=0.1,0.2,...,0.5$, momentum $\mu=0.1,0.2,...,0.5$ ($\lambda=25$),
- SVM: feature normalization turned on, cost $C=10^{-2},10^{-1},...,10^{2}$, exponent in radial basis kernel $\gamma=2^{-2},2^{-1},...,2^{2}$ ($\lambda=25$),
- random forest: number of trees $i=10,21,...,219$ ($\lambda=20$),
- APLSC: number of dimensions $d=5,10,15,20$ ($\lambda=4$).

Preliminary experiments on naïve Bayes classifier confirmed that kernel estimation improves classification results, so this feature was turned on. Validation threshold parameter in a multilayer perceptron in-

dicates how many times in a row the validation set error can increase before training is terminated. Early tests showed that introducing validation with this stop condition does not influence classification results but significantly reduces training time, therefore we decided to use it in our research. SMOTE filter was configured to balance positive and negative sets perfectly. SVM parameters in SMOTE + SVM combination were tuned with a wider range of values, that is $C = 10^{-2}, 10^{-1}, ..., 10^{3}$, and $\gamma = 2^{-2}, 2^{-1}, ..., 2^{4}$ ($\lambda = 42$). Authors of microPred used a more exhaustive scanning strategy, however it is inapplicable for larger problems because of computational overhead. Hence, we limited search space to cover parameter values selected most commonly in preliminary experiments. Geometric mean (G_m) was chosen as an evaluation metric to be maximised. Numbers of folds, k_1 and k_2, were set to 10 and 5, respectively. We decided to use 5-fold CV in the parameter tuning because it allowed us to reduce times of analyses with respect to 10-fold CV almost by half (parameter tuning dominates over other stages in terms of computation time), rendering slightly inferior results [33]. This approach follows microPred, which also used 5-fold CV for parameter tuning.

ROC-Select strategy described in the paper was prepared as a plug-in to Weka [34] package which had been chosen as the basic environment for all classification experiments. It provided us with implementations of naïve Bayes, multilayer perceptron, random forest and SMOTE filter. Weka interface for LibSVM was used for support vector machine experiments. The original APLSC code written in MATLAB was wrapped in Java class and also attached to Weka as a plug-in.

Results and discussion

Threshold selection

The first step of the experiments was to check how the threshold selection strategy influences classification results. For each classifier undergoing ROC-select procedure four tests were carried out: no selection (I), threshold selection only (II), parameter selection only (III), both parameter and threshold selection (IV). Relative G_m changes of variants II, III and IV with respect to the variant I were calculated and averaged over all datasets beside *microPred* (Table 2). As one can see, applying threshold selection procedure leads to significant improvement in G_m values. The exception is naïve Bayes for which the gain is moderate. This can be explained by intrinsic resistance of naïve Bayes to the class imbalance problem - it performed well without applying ROC-select. In the case of naïve Bayes no parameters were tuned, thus variants III and IV are the same as I and II, respectively. In other cases the best results were obtained with combination of parameter and threshold tuning. It is important to note that variant II overtakes relevantly variant III. This confirms that standard machine learning techniques are not suited for imbalanced datasets and adjusting classifier parameters can reduce the problem of overlearning majority class only by a small marigin. To achieve best possible performance, classifiers suited for imbalanced problems (SMOTE + SVM and APLSC) were always tested with parameter tuning turned on (variant III). For computational reasons we decided to limit parameter space from 42 points to 25 while running SMOTE + SVM on *animal* set (same points as in SVM and ROC-select combination were used).

Absolute values of sensitivity, specificity and G_m for particular classifiers and datasets are given in Table 3. As applying ROC-select procedure improved performance much more relevantly than parameter tuning, only results for variants III and IV are presented. The general observation is that traditional classification algorithms at default threshold (variant III) clearly overlearn majority class and lose with SMOTE + SVM and APLSC in terms of G_m. The greater class imbalance, the more visible is this regularity. For instance in the case of *virus* dataset, which is only slightly imbalanced, traditional algorithms perform almost as good as imbalance-suited methods. The opposite is *human* set, in which methods

Table 2 Relative gains in classification results

Classifier	threshold selection (II)	parameter selection (III)	parameter + threshold selection (IV)
Naïve Bayes	1.11	0.00	1.11
Perceptron	7.70	0.26	7.76
SVM	10.11	1.89	10.29
Random forest	6.95	1.55	9.30

Relative percentage gains in G_m obtained by applying parameter and/or threshold selection on different classifiers averaged over all datasets.

are strongly biased towards negative class giving low sensitivity (less than 70%) and high specificity (almost 100%) which results in unsatisfactory values of G_m. The only exception is naïve Bayes which produces results similar to SMOTE + SVM or APLSC.

Applying ROC-select procedure to traditional classifiers (variant IV) balances their sensitivity and specificity significantly improving G_m values (except for naïve Bayes in which gains are moderate). The best results were on average obtained for random forest which beats SMOTE + SVM and APLSC in all datasets. However, multilayer perceptron and SVM also overperformed imbalance-suited methods in the majority of cases. The conclusion is twofold: (1) score function returned by examined classifiers properly ranks instances with respect to the conditional class probability, (2) ROC-select procedure successfully applies this knowledge to solve imbalanced classification problem.

Another interesting observation comes from comparison of imbalance-suited strategies, that is SMOTE + SVM and APLSC. Our experiments confirm previous findings that APLSC is superior to SMOTE [32]. It is especially visible in large and highly imbalanced sets like *human* or *plant*. We explain this by the fact that SMOTE is able to produce only a limited number of informative examples. Above some threshold value, synthetically generated instances introduce only noise. An important observation is that APLSC seems to be the only classifier which is biased towards minority class (sensitivity is always higher than specificity) which may be a useful feature in some applications.

If one analyses absolute results for particular datasets, it becomes clear that animal sets (*human* and *animal*) are more resilient to classification than plant sets (*arabidopsis* and *plant*), even though they are more balanced. This is probably caused by the fact that plant miRNAs are better separated from non-miRNAs in the attribute space, hence they are easier to distinguish. The worst absolute results in terms of G_m were observed for *microPred* dataset. We explain this by the low quality of this set (miRBase 12 was known to contain some false positives removed in later releases [18]) and lack of experimental evidence-based filtering.

Statistical analysis

In order to statistically evaluate differences between classifiers, Friedman rank test [35] at significance level $\alpha = 0.05$ was carried out with G_m being chosen as a performance metric. All the datasets beside *microPred* were used in the procedure. We tested imbalance-suited methods (SVM + SMOTE, APLSC) together with naïve Bayes, perceptron, SVM and random forest in variant IV . The resulting critical difference (CD) diagram for post-hoc Nemenyi tests [35] is shown in Figure 1. As one can see, random forest, SVM and perceptron (which are gathered near rank 2.) outperform APLSC, naïve Bayes and SVM + SMOTE (clustered near rank 5.). Random forest and SVM + SMOTE were confirmed to be the most and least accurate classifiers, respectively. The difference between them as well as the difference between SVM + SMOTE and the second best classifier (SVM) are statistically significant.

Table 3 Detailed classification results

Classifier	parameter selection (III)			parameter + threshold selection (IV)		
	SE	**SP**	**G_m**	**SE**	**SP**	**G_m**
			human			
Naïve Bayes	87.98	96.33	92.06	91.97	93.93	92.94
Perceptron	69.56	99.84	83.34	**94.17**	**94.99**	**94.58**
SVM	69.56	99.85	83.34	92.53	95.69	94.10
Random forest	68.21	99.85	82.53	91.53	96.34	93.90
SMOTE + SVM	77.67	99.02	87.69			
APLSC	94.88	92.14	93.50			
			arabidopsis			
Naïve Bayes	86.99	98.91	92.76	91.30	97.77	94.48
Perceptron	80.09	99.95	89.47	93.04	97.47	95.23
SVM	80.07	99.96	89.47	93.04	98.95	95.95
Random forest	83.55	99.94	91.38	**95.22**	**99.04**	**97.11**
SMOTE + SVM	88.71	99.64	94.02			
APLSC	96.09	90.42	93.21			
			animal			
Naïve Bayes	85.54	95.53	90.40	88.83	92.81	90.79
Perceptron	74.03	99.65	85.89	91.78	95.13	93.44
SVM	72.04	99.74	84.77	90.67	96.09	93.34
Random forest	72.52	99.72	85.04	**92.00**	**95.21**	**93.59**
SMOTE + SVM	84.56	98.68	91.35			
APLSC	91.93	91.13	91.53			
			plant			
Naïve Bayes	83.56	97.56	90.29	87.48	95.84	91.57
Perceptron	77.30	99.80	87.83	89.64	97.38	93.43
SVM	73.07	99.85	85.42	89.46	97.93	93.60
Random forest	78.41	99.81	88.47	**90.65**	**97.96**	**94.24**
SMOTE + SVM	81.31	99.32	89.86			
APLSC	92.77	89.39	91.07			
			virus			
Naïve Bayes	93.21	93.21	93.21	95.74	92.37	94.04
Perceptron	87.77	98.10	92.79	94.08	95.71	94.89
SVM	90.31	98.10	94.12	**95.38**	**95.35**	**95.37**
Random forest	88.59	98.45	93.39	93.26	96.31	94.77
SMOTE + SVM	91.99	97.14	94.53			
APLSC	96.61	92.97	94.77			
			microPred			
Naïve Bayes	80.32	94.27	87.02	89.43	87.91	88.67
Perceptron	82.35	99.37	90.46	90.74	94.65	92.67
SVM	79.31	99.72	88.93	89.29	97.01	93.07
Random forest	75.83	99.66	86.94	**91.89**	**96.36**	**94.10**
SMOTE + SVM	87.70	98.83	93.10			
APLSC	91.45	90.96	91.21			

Absolute results with parameter selection alone and parameter selection combined with threshold selection obtained through 10-fold CV. Results of best classifier for each dataset typed in bold.

Figure 1 Statistical significance diagram. Critical difference diagram for Nemenyi tests performed on *human, animal, arabidopsis, plant, virus* datasets. Average ranks of examined methods are presented. Bold lines indicate groups of classifiers which are not significantly different (their average ranks differ by less than CD value)

Running time

Time of analysis is an important issue determining applicability of presented methods for real-life problems. As all investigated algorithms are eager learning strategies, testing time was always irrelevant with respect to the training time and is not considered here. In Table 4 medians of training times of all CV runs are given. We show results for the *microPred* set as it was used in other studies, together with *arabidopsis* (the most imbalanced set), *plant* and *animal* (two largest sets). Execution times of most time consuming algorithm variants (IV for naïve Bayes, perceptron, SVM, random forest and III for SMOTE + SVM and APLSC) are given. As all the algorithms were implemented in a serial manner, single analysis utilised just one core of quad-core Intel Xeon W3550 3.06 GHz CPU used for the experiment.

Table 4 Training times

Classifier	***microPred***	***arabidopsis***	***plant***	***animal***
Naïve Bayes	00:00:13	00:01:03	00:06:38	00:11:56
Perceptron	00:28:02	01:15:53	05:15:04	10:21:05
SVM	00:23:00	00:25:49	20:22:57	170:47:13
Random forests	00:17:27	00:59:15	07:58:10	23:07:23
SMOTE + SVM	01:26:00	04:05:17	252:02:10	281:11:12
APLSC	00:00:34	00:01:46	00:08:52	00:29:52

Classifier training times for selected datasets (medians over all cross-validation folds). Times are given in format *hh:mm:ss*.

One should remember that training times are influenced not only by the classification method itself, but also by the number of points in the parameter space to be analysed in a tuning stage. In the case of naïve Bayes classifier no parameters were tuned, thus it was the fastest classifier in the comparison (training times from seconds to minutes). For other classifiers undergoing ROC-select procedure, 20-25 points were evaluated. For smaller sets, training times obtained by multilayer perceptron, random forest and SVM were similar (tens of minutes). For larger sets support vector machines scaled worse than competitors (a few dozen of hours vs. hours). In the case of SMOTE + SVM strategy, 42 points were checked (except *animal* set in which only 25 points were examined). It is important to keep in mind that original microPred included more exhaustive, thus more time-consuming parameter tuning strategy. Limitation of search space did not prevent SMOTE + SVM from being the slowest strategy in our experiments though. In the case of *plant* and *animal* datasets single training took more than ten days which makes microPred strategy inapplicable for larger problems. In contrast, APLSC classifier (4 points in the parameter space) was very fast.

Eventually, we decided to use random forest combined with ROC-select as a basic strategy in HuntMi package due to its superior classification results and reasonable computation time.

Additional features

The next part of the experiments was to check how introducing additional features influences classification results. These experiments were carried out for random forest + ROC-select combination, selected

earlier as a basic strategy in HuntMi. As Table 5 shows, new features introduced additional information into classification procedure and improved final results. The absolute gain in G_m varied from 0.49 to 2.34. Wilcoxon test [35] performed on all datasets beside *microPred* confirmed predominance of the extended representation with p-value equal to 0.0952. For this reason we decided to use seven new features together with twenty one previously introduced to represent sequences in HuntMi package.

Table 5 Feature selection results

Dataset	SE	SP	G_m
human	95.31	97.18	96.24
arabidopsis	96.11	99.31	97.70
animal	94.92	96.60	95.76
plant	92.36	98.38	95.32
virus	96.18	95.95	96.06
microPred	92.76	96.46	94.59

Classification results obtained by ROC-select + random forest combination for extended representation including seven new features. These are also the final results for HuntMi software.

Comparison with other tools

The majority of miRNA classification studies focus on *H. sapiens*. As microPred was proven to be the best software in this field at the time of its publication, we decided not to consider its predecessors such as Triplet-SVM, MiPred or miPred in the comparison. The results produced by SMOTE + SVM combination on *microPred* dataset were very similar to those obtained by [17] ($G_m = 93.53$), which confirms that our experiments accurately estimate microPred performance. The small discrepancy is probably caused by different splits in cross-validation procedure (microPred used 5-fold CV for testing). HuntMi software gave $G_m = 94.59$ (see Table 5), which is a noticeable improvement over microPred. The predominance of HuntMi method over SMOTE + SVM combination employed by microPred holds also for all other sets and is statistically significant. To further test the performance of HuntMi, we prepared a set of animal microRNAs newly introduced in miRBase issues 18-19 and examined it on a classification model trained on the entire *animal* dataset (built upon miRBase 17). The obtained results clearly demonstrate that HuntMi is able to efficiently identify novel microRNAs in animals, achieving the sensitivity of over 90% in 8 out of 11 analysed species (Table 6). At the same time the sensitivity achieved by microPred is considerably lower, exceeding 90% only for *O. latipes*.

Table 6 Comparison with other tools: animal species

Species	#Sequences	microPred	HuntMi
Bombyx mori	4	75.00	100.00
Caenorhabditis elegans	16	87.50	93.75
Ciona intestinalis	19	89.47	73.68
Homo sapiens	175	85.14	93.14
Macaca mulatta	16	-	81.25
Mus musculus	139	64.03	94.96
Oryzias latipes	152	94.08	96.05
Pongo pygmaeus	54	83.33	94.44
Rattus norvegicus	38	76.32	97.37
Taeniopygia guttata	23	82.61	91.30
Tribolium castaneum	14	64.29	78.57

Classification sensitivity of microPred and HuntMi on animal miRNAs added in miRBase issues 18-19.

Several studies on improving microPred have been carried out. They exploited techniques like sample selection [36] or genetic algorithm-based feature selection [37, 38] resulting in very high values of G_m (up to 99). All these methods were, however, evaluated on balanced subsets of *microPred* dataset and

some of them suffered from important methodological incoherences like lack of random split of data into training and testing set and, more importantly, inclusion of training sequences in a testing set. Therefore, reported results do not accurately estimate the performance of presented strategies in real miRNA identification problems. In addition, these methods are not available as a ready to use packages.

Another strategy, MiRenSVM [39], employed SVM ensembles for miRNA classification. It was tested on moderately imbalanced dataset (697 human miRNAs, 5 428 pseudo harpins) with 3-fold CV resulting in $G_m = 94.76$. This value is very similar to the one obtained by HuntMi on *microPred* dataset which consisted of same positive examples and 50% more negatives. MiRenSVM was also tested on a set of 5 238 animal miRNAs successfully identifying 92.84% of them. As no negative sequences were included, specificity of the method is unknown. In our experiments, HuntMi was examined on a set consisting of 7 053 animal miRNAs and 218 154 pseudo hairpins. It outperformed MiRenSVM giving sensitivity of 94.92% and specificity of 96.60%. As MiRenSVM is not available as a tool, we were not able to compare its performance with HuntMi on miRNAs introduced in latest builds of miRBase.

Separate group of methods specialising in plant microRNA identification has been developed, of which the most recent is PlantMiRNAPred [19]. It joins feature and sample selection strategies to improve SVM classification results. The main dataset used in the research consisted of 1 906 real pre-miRNAs from miRBase 14 and 2 122 non-miRNAs generated by authors. 980 positive and 980 negative examples were selected using proposed sample selection method to train the classifier. Majority of the remaining sequences and 309 new miRNAs from miRBase 15-16 constituted the testing set. Surprisingly, as many as 634 training positives were also added to this set. This, together with lack of random split of data into training and testing sets results in overestimation of classification performance. Despite these incoherences, HuntMi performed smililarly to PlantMiRNAPred. After summing up results from PlantMiRNAPred study we obtained $G_m = 96.91$, while HuntMi gave 95.32 and 97.70 on *plant* and *arabidopsis* datasets respectively. To further evaluate performance of HuntMi package in plant microRNA classification, we tested it on miRNAs introduced in 18-19 builds of miRBase. Classification model was trained on the full *plant* dataset (constructed upon miRBase 17). As PlantMiRNAPred permits only for manual submissions of single sequences (service for processing FASTA files malfunctioned at the time of this study) we examined it on species with at most 200 newly introduced miRNAs. The results are presented in Table 7.

Table 7 Comparison with other tools: plant species

Species	#Sequences	PlantMiRNAPred	HuntMi
Arabidopsis thaliana	68	80.88	91.18
Cucumis melo	120	90.00	95.00
Glycine max	302	-	88.41
Hordeum vulgare	45	55.56	35.56
Malus domestica	206	88.83	99.51
Medicago truncatula	300	-	72.67
Nicotiana tabacum	163	84.66	93.25
Oryza sativa	169	60.95	69.82
Populus trichocarpa	89	89.89	97.75
Sorghum bicolor	58	94.83	94.83

Classification sensitivity of PlantMiRNAPred and HuntMi on plant miRNAs added in miRBase issues 18-19. PlantMiRNAPred failed to process some *Arabidopsis thaliana* miRNAs successfully. However, these sequences were treated as properly identified.

Based on obtained results, all the plant species examined by HuntMi can be divided into two groups. In the first group (*A. thaliana, C. melo, G. max, M. domestica, N. tabacum, P. trichocarpa, S. bicolor*) the classification sensitivity varied from 88.41% to 99.51% and is clearly superior to the performance of PlantMiRNAPred. The second group (*H. vulgare, M. truncatula* and *O. sativa*) was characterised

by much lower sensitivity (35.56% to 72.67%). Two of the latter species belong to monocotyledons, which could suggest that our tool is inefficient when analysing sequences from this plant group. However, we obtained satisfactory sensitivity for *S. bicolor* (94.64%). This encouraged us to look closer at microRNAs from low-sensitivity group and we discovered that a large fraction of miRNAs in these species do not meet commonly recognised criteria for annotation of plant miRNAs e.g. in the case of osa-MIR5489, osa-MIR5484, hvu-MIR6177, hvu-MIR6182, mtr-MIR5741d and some other miRNAs the mature microRNA lies outside the stem part of the hairpin. Additionally, most of new miRNAs were discovered using deep sequencing approach only, where it is sometimes only one or several reads that support the miRNA (e.g. osa-MIR5527). This data is insufficient to confirm that the miRNA is precisely excised from the stem. Similarly to HuntMi, PlantMiRNAPred produces unsatisfactory results when applied to *H. vulgare* or *O. sativa* miRNAs (sensitivities of 56% and 61%).

To sum up, in majority of cases HuntMi was able to obtain better results than competitors even though it was evaluated on larger and more imbalanced datasets. Experiments on animal and plant miRNAs introduced in releases 18-19 of miRBase confirmed that HuntMi outperforms other tools like microPred and PlantMiRNAPred. There are methods reporting higher G_m values than HuntMi. However, they were all tested on balanced datasets, often with important methodological flaws, which obstructs proper judgement of their performance in real-life tasks. Moreover, none of these methods is available as a ready to use package.

Conclusions

In this study we present a new machine learning-based miRNA identification package called HuntMi. It exploits ROC-select, a special strategy of thresholding score function output by classifiers, combined with random forest, which we find to produce best classification results. Twenty one features employed by microPred software together with seven new attributes are used as a data representation. The method was tested on large and strongly imbalanced datasets using stratified 10-fold cross-validation procedure. Classifiction performance was further verified on miRNAs newly introduced in latest builds of miRBase. As a result, HuntMi clearly outperforms state-of-the-art miRNA hairpin classification tools like microPred and PlantMiRNAPred without compromising the training time.

HuntMi comes with G_m-optimised models for *H. sapiens*, *A. thaliana*, animals, plants and viruses. There is a possibility to train a model on any dataset and subsequently use it in classification analysis. This feature may be useful if one is interested in predicting miRNAs in particular species or in applying different optimization criterion than G_m in ROC-select procedure. Therefore, HuntMi offers the highest flexibility of all existing microRNA classification packages.

Abbreviations

APLSC: asymmetric partial least squares classification; CV: cross-validation; FP: false positive; HMM: hidden Markov model; MFE: minimum free energy; ROC: receiver operating characteristic; SE: sensitivity; SMOTE: synthetic minority over-sampling technique; SP: specificity; SVM: support vector machine; TP: true positive.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AG and MWS contributed to the manuscript equally. AG prepared implementation of ROC-select method and performed experiments. MWS designed features used in classification and prepared datasets. Both AG and MWS analysed experimental results and drafted the manuscript. MS and IM revised the manuscript and supported the research from statistical and machine learning (MS) as well as biological (IM) side. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the European Social Fund grant UDA-POKL.04.01.01-00-106/09 to AG; National Science Centre grant 2011/01/N/NZ2/01653 to MWS; National Science Centre grant 2011/01/B/ST6/06868 to AG, MWS, IM; National Science Centre grant DEC-2011/01/D/ST6/07007 to MS; Faculty of Biology at AMU grant PBWB-08/2011 to MWS. We wish to thank Adam Adamarek for proofreading the manuscript.

References

1. Laganá A, Forte S, Giudice A, Arena MR, Puglisi PL, Giugno R, Pulvirenti A, Shasha D, Ferro A: **MiRó: A MiRNA Knowledge Base**. Database (Oxford) 2009, **2009:**bap008.
2. Cai X, Hagedorn CH, Cullen BR: **Human MicroRNAs are processed from capped, polyadenylated transcripts that can also function as MRNAs.** *RNA* 2004, **10:**1957–1966.
3. Davis-Dusenbery BN, Hata A: **Mechanisms of control of MicroRNA Biogenesis.** *J Biochem* 2010, **148:**381–392.
4. Brabletz S, Bajdak K, Meidhof S, Burk U, Niedermann G, Firat E, Wellner U, Dimmler A, Faller G, Schubert J, Brabletz T: **The ZEB1/miR-200 feedback loop controls notch signalling in cancer Cells.** *EMBO J* 2011, **30:**770–782.
5. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, N R: **Discovering MicroRNAs from deep sequencing data using MiRDeep.** *Nat Biotechnol* 2008, **26:**407–415.
6. Hertel J, Stadler PF: **Hairpins in a haystack: recognizing MicroRNA precursors in comparative genomics data.** *Bioinformatics* 2006, **22:**197–202.
7. Jones-Rhoades MW, Bartel DP: **Computational identification of plant MicroRNAs and their targets, including a stress-induced MiRNA.** *Mol Cell* 2004, **14:**787–799.
8. Ng KL, Mishra SK: **De Novo SVM Classification of precursor MicroRNAs from genomic pseudo hairpins using global and intrinsic folding measures.** *Bioinformatics* 2007, **23:**1321–1330.
9. Bentwich I: **Prediction and validation of MicroRNAs and their targets.** *FEBS Lett* 2005, **579:**5904–5910.
10. Mhuantong W, Wichadakul D: **MicroPC (microPC): A Comprehensive resource for predicting and comparing plant MicroRNAs.** *BMC Genomics* 2009, **10:**366.
11. Szczesniak M, Deorowicz S, Gapski J, Kaczynski L, Makalowska I: **MiRNEST database: an integrative approach in MicroRNA search and annotation.** *Nucleic Acids Res Database Issue* 2012, D198–D204.

12. Doran J, Strauss WM: **Bio-informatic trends for the determination of MiRNA-target interactions in mammals.** *DNA Cell Biol* 2007, **26:**353–360.

13. Kadri S, Hinman V, Benos PV: **HHMMiR: Efficient De Novo prediction of MicroRNAs using hierarchical hidden Markov models.** *BMC Bioinformatics* 2009, **10**(Suppl 1):S35.

14. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z: **MiPred: classification of real and pseudo MicroRNA precursors using random forest prediction model with combined features.** *Nucleic Acids Res* 2007, **35:**W339–W344.

15. Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK: **Combining multi-species genomic data for MicroRNA identification using a Naïve Bayes classifier**. *Bioinformatics* 2006, **22:**1325–1334.

16. Xue C, Li F, He T, Liu GP, Li Y, Zhang X: **Classification of real and pseudo MicroRNA precursors using local structure-sequence features and support vector machine.** *BMC Bioinformatics* 2005, **6:**310.

17. Batuwita R, Palade V: **MicroPred: effective classification of pre-miRNAs for human MiRNA gene prediction.** *Bioinformatics* 2009, **25:**989–995.

18. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data.** *Nucleic Acids Res* 2011, **39**(Database Issue): D152–D157.

19. 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.

20. Chawla NV, Japkowicz N, Kotcz A: **Editorial: special issue on learning from imbalanced data sets.** *SIGKDD Expl* 2004, **6:**1–6.

21. He H, Garcia EA: **Learning from imabalanced data.** *IEEE Trans Know and Data Eng* 2009, **21:**1263–1284.

22. Mease D, Wyner AJ, Buja A: **Boosted classification trees and class probability/quantile estimation.** *J Mach Learn Res* 2007, **8:**409–439.

23. Zadrozny B, Elkan C: **Transforming classifier scores into accurate multiclass probability estimates.** In *Proceedings of KDD 2002*. New York: ACM; 2002:694–699.

24. Domingos P: **MetaCost: A general method for making classifiers cost-sensitive.** In *Proceedings of KDD 1999*. New York: ACM; 1999:155–164.

25. Fawcett T: **An introduction to ROC analysis.** *Pattern Recogn Lett* 2006, **27:**861–874.

26. Duda RO, Hart PE: *Pattern Classification and Scene Analysis.* New York: Wiley; 1973.

27. Rosenblatt F: *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Washington: Spartan Books; 1962.

28. Boser BE, Guyon IM, Vapnik VN: **A training algorithm for optimal margin classifiers.** In *Proceedings of COLT 1996*. ACM Press; 1992:144–152.

29. Brieman L: **Random forests.** *Mach Learn* 2001, **45:**5–32.

30. Keerthi S, Lin CJ: **Asymptotic behaviours of support vector machines with gaussian kernel.** *Neural Comput* 2003, **15:**1667–1689.

31. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: **SMOTE: Synthetic Minority Over-sampling Technique.** *J Artif Intell Res* 2002, **16:**321–357.

32. Qu HN, Li GZ, Xu WS: **An asymmetric classifier based on partial least squares.** *Pattern Recogn* 2010, **43:**3448–3457.

33. Kohavi R: **A study of cross-validation and bootstrap for accuracy estimation and model selection.** In *Proceedings of IJCAI 1995, Vol. 2*. San Mateo: Morgan Kaufmann; 1995:1137–1143.

34. Hall M, Eibe F, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA data mining software: an update.** *SIGKDD Expl* 2009, **11:**10–18.

35. Demsar J: **Statistical comparisons of classifiers over multiple data sets.** *J Mach Learn Res* 2006, **7:**1–30.

36. Han K: **Effective sample selection for classification of Pre-miRNAs.** *Genet Mol Res* 2011, **10:**506–518.

37. Wang Y, Chen X, Jiang W, Li L, Li W, Yang L, Liao M, Lian B, Lv Y, Wang S, Wang S, Li X: **Predicting human MicroRNA precursors based on an optimized feature subset generated by GA-SVM.** *Genomics* 2011, **98:**73–78.

38. Xuan P, Guo M, Wang J, Wang CY, Liu XY, Liu Y: **Genetic algorithm-based efficient feature selection for classification of Pre-miRNAs.** *Genet Mol Res* 2011, **10:**588–603.

39. 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.

Additional file

Additional_file_1 as PDF
Additional file 1: A file with supplementary tables. Table S1 summarises animal and plant species and viruses from which non-miRNA sequences were extracted.

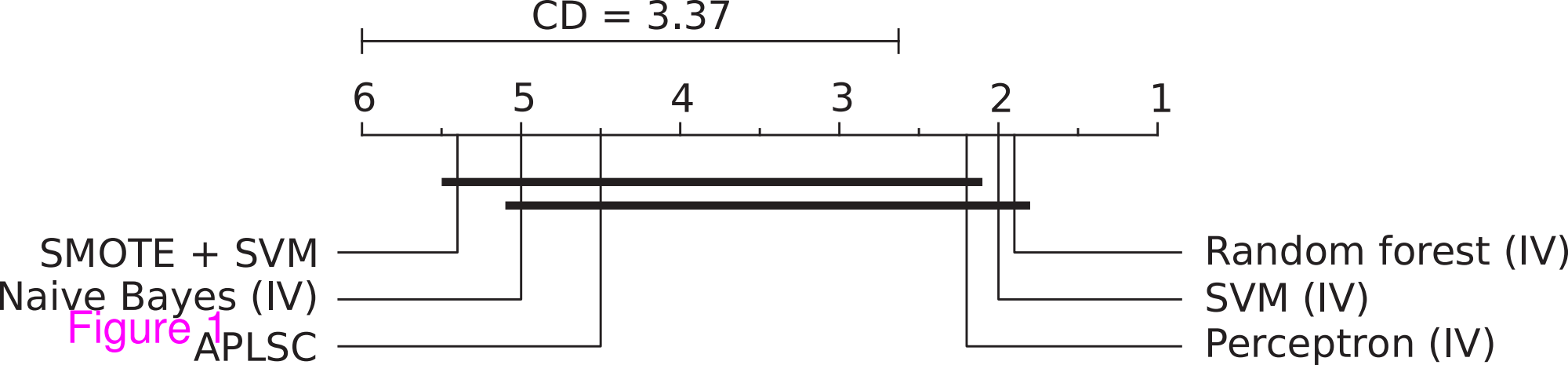

CD = 3.37
6
5
4
3
2
1
SMOTE + SVM
Naive Bayes (IV)
APLSC
Random forest (IV)
SVM (IV)
Perceptron (IV)

Figure 1

Additional files provided with this submission:

Additional file 1: 1556810936757324_add1.pdf, 35K
http://www.biomedcentral.com/imedia/1675913460936058/supp1.pdf

D198–D204 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 1 December 2011*
doi:10.1093/nar/gkr1159

miRNEST database: an integrative approach in microRNA search and annotation

Michał Wojciech Szcześniak[1,*], Sebastian Deorowicz[2], Jakub Gapski[1], Łukasz Kaczyński[1] and Izabela Makałowska[1,*]

[1]Laboratory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University, Umultowska 89, 61-614 Poznan and [2]Institute of Informatics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

Received August 12, 2011; Revised and Accepted November 10, 2011

ABSTRACT

Despite accumulating data on animal and plant microRNAs and their functions, existing public miRNA resources usually collect miRNAs from a very limited number of species. A lot of microRNAs, including those from model organisms, remain undiscovered. As a result there is a continuous need to search for new microRNAs. We present miRNEST (http://mirnest.amu.edu.pl), a comprehensive database of animal, plant and virus microRNAs. The core part of the database is built from our miRNA predictions conducted on Expressed Sequence Tags of 225 animal and 202 plant species. The miRNA search was performed based on sequence similarity and as many as 10 004 miRNA candidates in 221 animal and 199 plant species were discovered. Out of them only 299 have already been deposited in miRBase. Additionally, miRNEST has been integrated with external miRNA data from literature and 13 databases, which includes miRNA sequences, small RNA sequencing data, expression, polymorphisms and targets data as well as links to external miRNA resources, whenever applicable. All this makes miRNEST a considerable miRNA resource in a sense of number of species (544) that integrates a scattered miRNA data into a uniform format with a user-friendly web interface.

INTRODUCTION

Animal and plant miRNA genes are transcribed by RNA polymerase II or III into a primary transcript, called pri-miRNA (1). During initial steps of miRNA biogenesis, pri-miRNA is cut and a hairpin-shaped intermediate, called pre-miRNA, is produced. This process is catalyzed by Drosha in animals (2) and DCL1 (DICER-LIKE 1) in plants (3). Subsequently, a pre-miRNA is specifically cut at stem part of the hairpin and a miRNA/miRNA* duplex with 2-nt overhangs at 3′-ends is released. In animals this process is run by Dicer (2) and in plants it is controlled by DCL1 (3). In cytoplasm, one of duplex components, referred to as mature miRNA, gets incorporated into a riboprotein complex, named RISC (RNA-induced silencing complex) (4). RISC contains a functional unit, which allows regulation of the gene expression based on complementarity of the miRNA and the transcript of targeted gene. There are two modes of the regulation: by cleaving transcripts (5) and by inhibiting translation (6). The first one requires high complementarity between miRNA and targeted transcript and is ubiquitously observed in plants, while in animals it is translation inhibition that constitutes a major mechanism of miRNA action.

miRNAs regulate the expression of thousands of genes in plants and animals and are key players in developmental (7), stress-related (8) and signalling processes (9). A number of miRNAs have been associated with diseases in human, e.g. Alzheimer's disease (10), pancreatic cancer (11), or leukemia (12). Hence, identification of miRNAs and subsequent elucidation of their functions, both in plants and animals, became a critical issue not only in molecular biology but also in medical research and agriculture.

Recently, a number of investigations aimed at the identification of miRNAs have been published. Reported miRNAs were discovered either based on computational (13) or experimental approaches (14). Consequently, the growing number of miRNA studies led to accumulation of miRNA databases. However, many of these databases, like miRO (15) or miROrtho (16), are limited to species of high interest. Other resources are focused on selected taxa, e.g. microPC (13) and PMRD (17) contain only

*To whom correspondence should be addressed. Tel: +48 61 829 5836; Email: miszcz@amu.edu.pl
Correspondence may also be addressed to Izabela Makałowska. Tel: +48 61 829 5835; Email: izabel@amu.edu.pl

plant miRNAs, CoGemiR (18) consists of miRNAs from selected animal species and Vir-Mir db (19) is dedicated to viruses only. The main miRNA repository database, miRBase (20), although accommodates data from a wide range of species, contains only already published results. Similarly, databases like miRecords (21) or miRNAMap (22) consist exclusively of experimentally verified miRNAs. Contrary, microPC dataset is based solely on computational methods. Therefore, despite of the ever-growing number of miRNA-related resources, there is a lack of a single universal repository and users need to browse through a number of dispersed data sets to collect information related to specific species or miRNA type.

To overcome these limitations, we developed miRNEST as a comprehensive online resource for plant, animal and virus miRNAs. We applied a comparative approach to search for new microRNAs in animal and plant EST sequences from dbEST (23). The approach applied by us made it possible to identify 10 004 miRNA candidates in 221 animal and 199 plant species. To the best of our knowledge, for 236 species no miRNAs have been known before. Besides miRNA identification, we also performed miRNA target search in plant candidates and for 29 species we collected small RNA reads from Gene Expression Omnibus (24). Since our goal was not only to identify new miRNAs but also to develop a resource that would integrate miRNA data scattered across literature and databases, we complemented data resulting from our computational analysis with miRNA sequences from three other databases and two publications. In addition, based on availability, we incorporated the data from twelve resources providing further annotations of miRNAs from selected species. This gives the possibility to access, search and browse data from different resources simultaneously. Altogether the miRNEST database contains 39 122 miRNAs from 544 species. All data are presented in the same format via miRNEST interface and are available at http://mirnest.amu.edu.pl.

Data acquisition

We imported 16 961 known mature miRNA sequences from miRBase (20) and 9212 sequences from PMRD (17), which were used for comparative analysis and identification of conserved miRNAs. EST sequences of 225 animal and 202 plant species were downloaded from dbEST (23). We took into consideration only the species that had at least 10 000 EST sequences. For EST annotation, the UniProtKB/Swiss-Prot protein data set was obtained from UniProt (25), and to remove tRNAs and rRNAs from ESTs and for further miRNA annotation, we obtained ncRNA sequences from RFAM release 9.1 (26). Also, 192 small RNA deep sequencing libraries were downloaded from GEO, Gene Expression Omnibus (24).

In addition to mature miRNA and pre-miRNA data from miRBase and PMRD, we downloaded sequences from microPC (18) and two publications (27,28). The choice of databases and papers was made based on the overlap between resources and the availability of the data. For example, the sets of miRNAs deposited in miRNAMap (22) and RNAdb (29) fully overlapped with miRBase and therefore were not considered for the download. On the other hand, data from resources like miROrtho (16) or GrapeMiRNA (30) could not be downloaded and consequently, were not integrated with miRNEST. Targets data were collected from the largest and most recognized resources as miRDB (31), miRTarBase (32), miRecords (21), PMRD and ASRP (33). We also obtained miRNA expression data from PhenomiR 2.0 (34) and dbDEMC (35), miRNA regulation data and promoters from dPORE-miRNA (36) and PMRD, genomics information from CoGemiR v1.2b (18), miRNA polymorphisms from Patrocles (37) and imprinting data from ncRNAimprint (38). In particular, we took advantage of data available in miRBase, which among others provided us with literature references, links to external databases, genomics data, and served as a source of uniform, non-redundant miRNA nomenclature, which was of great importance when integrating the miRNEST external miRNA data. The list of all databases used for the data assembly is provided in Supplementary Table S1 and the range of miRNA overlap between resources in Supplementary Table S2.

Prediction of miRNAs

The prediction of animal and plant microRNA candidates was performed according to modified algorithm applied by Zhang *et al.* (39) and Mhuantong *et al.* (13) (Figure 1). The modifications include removal of sequences with high percentage of low-complexity regions, assembling ESTs into contigs using CAP3 (40) and position-based allowance for substitutions in mature miRNAs. Schema of all data processing steps is presented in Figure 1.

In the first step, mature miRNA sequences from miRBase and PMRD were searched against EST sequences using BLASTN (41). We filtered the BLASTN search results, based on pre-computed distribution of mismatches in mature miRNAs within miRNA homologs in animals and plants (Supplementary Figure S1). Then, EST sequences, clustered based on the similarity to the same miRNA, were assembled into contigs using CAP3. This step is important in plant microRNA prediction where pre-miRNA length occasionally exceeds 600 nucleotides (Supplementary Figure S2). Next, we ran BLASTN search against RFAM non-coding RNA sequences to remove rRNAs and tRNAs (with *E*-value $<$ 1e-20). After that we searched for low-complexity regions using Dustmasker (42) and removed sequences containing more than 60% of low complexity regions. Then, RNA secondary structures were predicted using RNAfold (43). We filtered the secondary structures and left only the candidates with mature miRNA located in a stem part of the hairpin structure and with no more than five mismatches and two bulges between mature miRNA and the opposite hairpin arm (Supplementary Figure S3). At the same time, we extracted the hairpin sequences out of longer contig or singleton. After that, we performed a BLASTX search against UniProt to remove sequences with high similarity to proteins (*E*-value $<$ 1e-20). Finally, we removed animal pre-miRNAs with length exceeding 215 nt, which is the

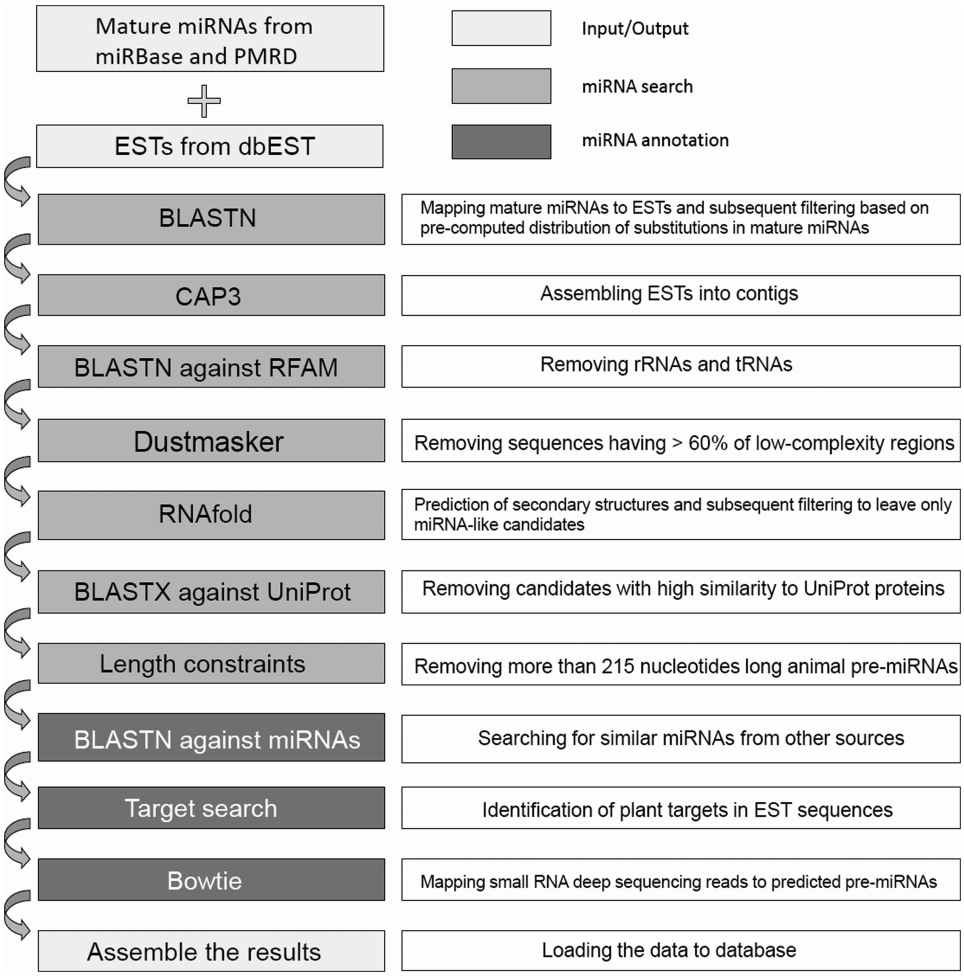


Figure 1. A computational pipeline applied for prediction of microRNAs in EST sequences and their annotation. The seven steps in miRNA search part are designed to minimize the false positives rate and provide a high quality set of candidates. In some of them, plant- and animal-specific parameters were applied, as described in the main text. The annotation part serves to provide more data on predicted miRNAs and no candidates are discarded there.

maximum length for animal pre-miRNA in miRBase. As a result, we obtained 10 004 miRNA candidates in 221 animal and 199 plant species.

The ratio of sequences removed from the data set at various filtering steps differed between species; examples for *Arabidopsis thaliana* and *Bos taurus* are presented in Supplementary Table S3.

miRNA prediction accuracy

Specificity. To assess the specificity of miRNEST algorithm we generated one million random sequences that were subsequently subject to the algorithm. The initial number of BLAST hits, i.e. mature miRNAs against generated sequences, was 231 260. After filtering steps, Dustmasker and BLASTN against RFAM, the number of candidates has decreased substantially to 1120 and only three candidates left after secondary structure check point. The last step, BLAST against UniProt, reduced the number of candidates to two, which produces the false positives ratio of 0.0002% (two per million). This result is comparable with the one calculated in a very similar way by authors of microPC database (13). Applying a similar approach they obtained the ratio of 0.00064%. We have also checked how many of our human miRNA predictions are classified as true miRNAs by microPred (44), a tool for miRNA/non-miRNA classification. This tool was specifically designed for human pre-miRNAs and it classified as miRNAs 75.9% of miRNEST predicted human miRNAs. At the same time, 90.16% of miRBase human pre-miRNAs are considered as true miRNAs by this program.

Sensitivity. To estimate the sensitivity of applied algorithm known pre-miRNAs of *B. taurus* and *A. thaliana*, from miRBase database, were subjected to miRNA search. Out of 229 Arabidopsis hairpins, 209 (91.3%) have been recovered. In case of *B. taurus*, the result was lower and only 61.1% (392 out of 662) of hairpins have been recognized as miRNA by miRNEST algorithm. This is mostly because of the architecture of animal hairpins: most of the miRNAs that have not been recovered bear more than five mismatches and/or two bulges between mature miRNA and the opposite hairpin arm.

Nonetheless, we decided to keep this filtering threshold in order to maintain the low false positives ratio. Authors of microPC database, where miRNAs are also predicted based on EST sequences analysis, estimated the sensitivity of their method for all analyzed plant species as 81.3%. We did calculation separately for individual species but on average, the rate of specificity is similar to this obtained in microPC.

Further annotation of predicted miRNAs

Plant mature miRNAs generally show a high level of complementarity with targeted mRNA sequences, thus target identification usually is not a very complex problem here. On the other hand, effective methods for target search in animals heavily rely on the evolutionary conservation of miRNA target site (45). Since such data are unavailable for a great majority of analyzed by us animal sequences, we decided to focus on target search for plant candidates and we used external target data for animal species whenever possible.

The targets were searched for in EST sequences from corresponding species applying standard rules of plant miRNA–mRNA interactions (46). In the final scoring a mismatch was given a score of 1, a wobble (G:U) was given a score of 0.5 and a bulge was given a score of 2.0. All matches with score above 3.5 were discarded. It was also important that positions 10 and 11 of miRNA perfectly matched to its target and that there was no more than 1 mismatch at positions 2 to 9. As a result, we identified targets for 6963 mature miRNAs in 187 plant species. Sequences of all potentially targeted ESTs were checked against UniProt data for functional annotations.

For 29 species, we downloaded small RNA deep sequencing libraries from GEO. The reads were mapped to predicted pre-miRNAs using Bowtie (47). The mapping was performed against both pre-miRNA strands and only one mismatch was allowed. The choice of species for which deep sequencing data were downloaded was solely based on the availability.

Processing of external data

The goal of our project was not only to predict miRNAs in EST sequences but also to build a comprehensive miRNA database. Therefore, we supplemented data set from our computational analyses with the sequence data from three external resources: miRBase, PMRD, microPC and two publications (27,28) reporting miRNAs not deposited in miRBase. All imported sequences were run through our pipeline so that all data are deposited and displayed in identical format. However, filtering steps were turned off to ensure that all external miRNAs would be incorporated into the miRNEST database even if they did not match criteria applied in miRNEST algorithm. Finally, we run reciprocal BLASTN search: each miRNEST pre-miRNA against each other to identify similar miRNAs across data sets.

Web interface

The miRNEST web interface has been divided into five sections that correspond to distinct types of data.

Browse. By clicking Browse in main menu, the user gets access to all miRNA sequences: miRNEST predictions and miRNAs from external sources. For each miRNA record we provide miRNA* sequence, coordinates of miRNA and miRNA* in pre-miRNA, guide strand, number of mismatches and bulges between mature miRNA and the opposite hairpin arm, secondary structures of pre-miRNAs, family assignment, experimental evidence and identical miRNEST, miRBase, PMRD or microPC miRNAs, whenever found (Figure 2). There are also links to miRNEST internal resources providing access to results of BLAST search, target predictions in EST sequences as well as target information from external sources, most similar mature miRNAs, source sequences for miRNAs predicted from ESTs, graphical display of deep sequencing reads mapping as well as data collected from external resources: miRNA genomics data, SNPs, promoters, TFBSs, miRNA-disease association, miRNA polymorphism, expression data, literature references, links to external resources and information on imprinted miRNAs. From here users have also opportunity to run BLASTN searches against selected sources of miRNA sequences and to build ClustalW (48) alignments based on BLASTN results. If species is selected, its full NCBI taxonomy is provided (49). By clicking a taxon, all miRNEST species belonging to the taxon are displayed as active links to miRNA data in corresponding species. Users can also browse through species in a taxonomic tree view.

Search. Using a search option users have the possibility to filter miRNA data by a number of parametres: species, sequence source (miRNEST prediction and/or external sources), mature miRNA sequence or its part, hairpin length, free folding energy, number of allowed mismatches and bulges between mature miRNA and the opposite arm in pre-miRNA hairpin structure, and *E*-values for BLAST search that was pre-run against UniProt, RFAM, miRBase, PMRD, and microPC. We also incorporated the target search option, where user can search for specific targets by typing mature miRNA sequence and selecting a species of interest. Moreover, users can limit their search to the records that have additional data, downloaded from outside resources, like experimental evidence or target sequences.

Unclassified. Unclassified section provides users with a list of miRNEST predictions that were not classified as potential miRNAs as 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 $\leq$215 nucleotides. There are 66 predictions that fall into this section due to the first criterion, 465—the latter and 8 predictions that violence both criteria. *Unclassified* and *Search* possess a very similar interface, however some options are unavailable in *Unclassified* category.

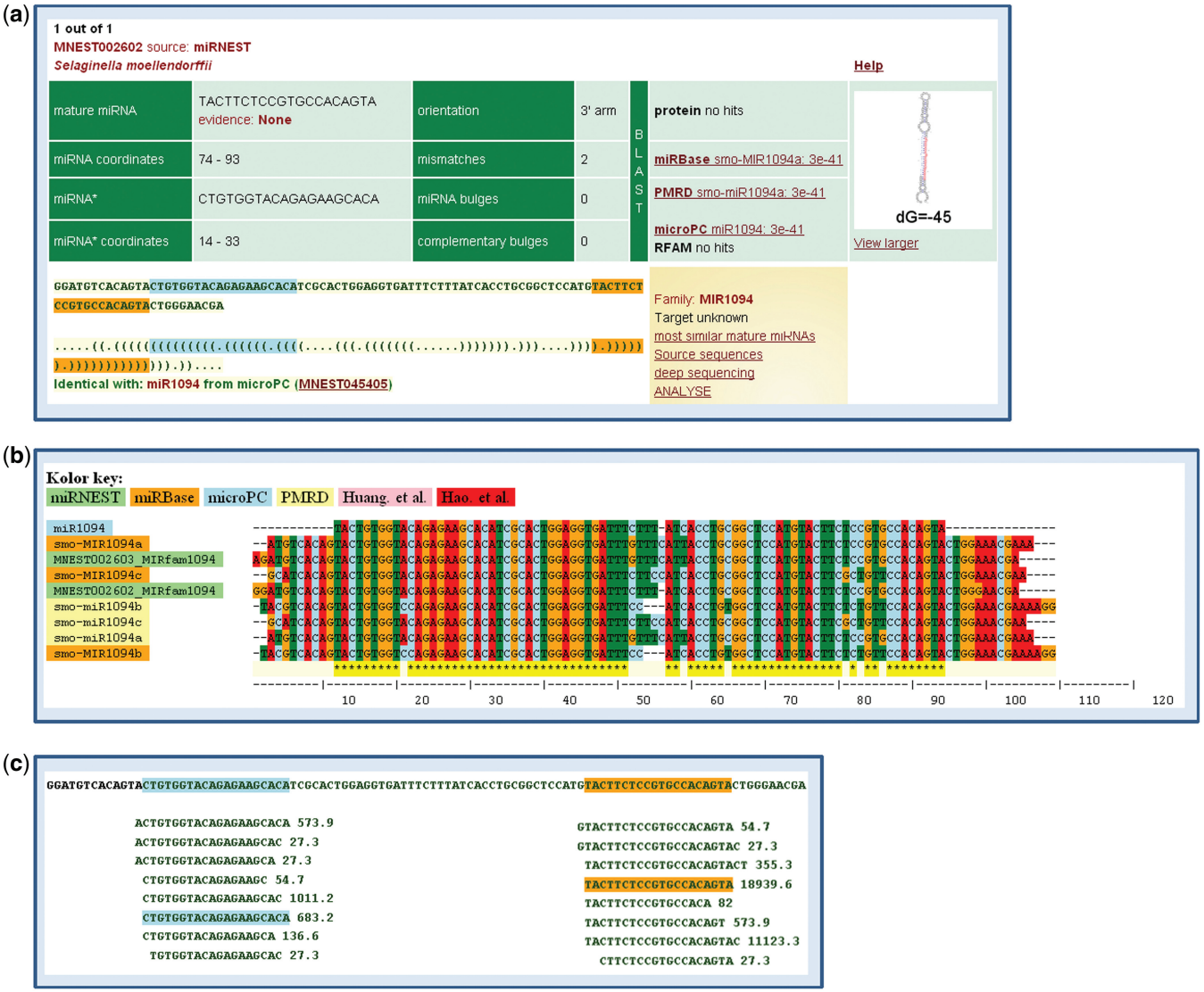


Figure 2. (**a**) An example of miRNEST record (MNEST002602) (**b**) by clicking *analyse* link it is possible to perform BLASTN search and ClustalW analysis. (**c**) *deep sequencing* link leads to mapping pattern of small RNA deep sequencing reads.

RNA-Seq. This component of the miRNEST database presents small RNA sequencing data that correspond to miRNEST predictions from 29 species (Supplementary Table S4). The user can select species, miRNA of interest, small RNA deep sequencing library, strand and number of allowed mismatches between reads and pre-miRNA. The results are displayed in a graphical form as an alignment of reads to the pre-miRNA. If no library is specified, then unique reads from all libraries from a given species are displayed. Otherwise, the read counts, as transcripts per million, are supplied. *RNA-Seq* data are also accessible from Browse section through deep sequencing internal link that corresponds to a given miRNA record.

Taxonomy. Taxonomy option provides a user with a phylogenetic tree of species that are included in miRNEST predictions. It presents clickable taxon names (kingdom, phylum, subclass, order, family, genus, and species) along with a number of taxon-specific miRNA families. By clicking on the taxon, the user gains access to more detailed information on the taxon-specific miRNA families and links to corresponding miRNEST records.

Comparison with other miRNA databases

miRNEST is a large collection of miRNA sequences and associated miRNA data (Supplementary Table 1). It encompasses 522 animal and plant species as well as 22 viruses (Supplementary Table S5) and as many as 9429 miRNA predictions can be found uniquely in miRNEST, though a fraction of the predictions overlap with known miRNA data (Supplementary Table S5). For comparison, miRBase (Release 17), the second largest miRNA database collects data from only 153 distinct

animals, plants and viruses. As many as 236 miRNEST species were not taken into account in any other miRNA database. These include a number of model species as well as the ones of high interest in agriculture and medical research like *Salmo salar*, *Takifugu rubripes*, *Actinidia chinensis*, *Anolis carolinensis*, *Bos indicus* or *Trichinella spiralis*.

Another distinguishing feature of miRNEST is the extent of usage of small RNA deep sequencing reads. We incorporated data from 192 libraries from 29 species retrieved from GEO and mapped the reads to predicted pre-miRNAs. We carefully selected the libraries in order to make sure that they encompass a wide array of tissues, developmental stages and conditions. For comparison, deepBase (50), which is a platform for annotating and discovering small and long ncRNAs from next-generation sequencing, collects data from 185 small RNA libraries from seven species and in miRBase, Release 17, deep sequencing data has been added for seven species, yet this functionality is now being expanded to more species.

miRNEST also gives access to large-scale target search predictions for 187 plant species that were generated using standard rules for plant miRNA target prediction. For more than half of the species this is the first and only miRNA targets prediction.

Finally, miRNEST predictions are complemented with a wide range of external data retrieved from 13 databases. Our goal was to create a resource that would integrate the miRNA data that are currently scattered across multiple resources and to limit existing necessity of searching multiple databases to investigate a single miRNA or miRNAs from a given species.

AVAILABILITY AND REQUIREMENTS

miRNEST is freely available at http://mirnest.amu.edu.pl. The database was constructed using Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), PHP 5.2.11 (http://www.php.net/), and MySQL 4.0.31 (http://www.mysql.com/). pre-miRNA secondary structures are drawn using Java lightweight applet VARNA (51) which requires the installation of Java plugin.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–3, Supplementary Tables 1–5.

ACKNOWLEDGEMENTS

The authors thank Professor Wojciech Makałowski for proofreading the article and Joanna Ciomborowska, Michał Kabza and Elżbieta Owczarkowska for helpful suggestions on the database web interface.

FUNDING

Faculty of Biology at Adam Mickiewicz University in Poznan, Poland (PBWB-08/2011); the European Social Fund within the project—'Fellowships for PhD Students in Wielkopolska Region conducting research in strategic areas supporting regional development'; the Ministry of Science and Higher Education (N N301 160935 to I.M. and N N516 441938 to S. D.). Funding for open access charge: Faculty of Biology at Adam Mickiewicz University in Poznan.

Conflict of interest statement. None declared.

REFERENCES

1. Cai,X., Hagedorn,C.H. and Cullen,B.R. (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, **10**, 1957–1966.
2. Lee,Y., Ahn,C., Han,J., Choi,H., Kim,J., Yim,J., Lee,J., Provost,P., Rådmark,O., Kim,S. *et al.* (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.
3. Kurihara,Y., Takashi,Y. and Watanabe,Y. (2006) The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. *RNA*, **12**, 206–212.
4. Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.
5. Reinhart,B.J., Weinstein,E.G., Rhoades,M.W., Bartel,B. and Bartel,D.P. (2002) MicroRNAs in plants. *Genes Dev.*, **16**, 1616–1626.
6. Lai,E.C. (2002) Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.*, **30**, 363–364.
7. Kedde,M. and Agami,R. (2008) Interplay between microRNAs and RNA-binding proteins determines developmental processes. *Cell Cycle*, **7**, 899–903.
8. Leung,A.K. and Sharp,P.A. (2010) MicroRNA functions in stress responses. *Mol. Cell*, **40**, 205–215.
9. O'Neill,L.A., Sheedy,F.J. and McCoy,C.E. (2011) MicroRNAs: the fine-tuners of Toll-like receptor signalling. *Nat. Rev. Immunol.*, **11**, 163–175.
10. Yao,J., Hennessey,T., Flynt,A., Lai,E., Beal,M.F. and Lin,M.T. (2010) MicroRNA-related cofilin abnormality in Alzheimer's disease. *PLoS One*, **5**, e15546.
11. Brabletz,S., Bajdak,K., Meidhof,S., Burk,U., Niedermann,G., Firat,E., Wellner,U., Dimmler,A., Faller,G., Schubert,J. *et al.* (2011) The ZEB1/miR-200 feedback loop controls Notch signalling in cancer cells. *EMBO J.*, **30**, 770–782.
12. Schotte,D., Pieters,R. and Den Boer,M.L. (2011) MicroRNAs in acute leukemia: from biological players to clinical contributors. *Leukemia*, June 24 (doi:10.1038/leu.2011.151; epub ahead of print).
13. Mhuantong,W. and Wichadakul,D. (2009) MicroPC (microPC): a comprehensive resource for predicting and comparing plant microRNAs. *BMC Genomics*, **10**, 366.
14. Zhang,L., Chia,J.M., Kumari,S., Stein,J.C., Liu,Z., Narechania,A., Maher,C.A., Guill,K., McMullen,M.D. and Ware,D. (2009) A genome-wide characterization of microRNA genes in maize. *PLoS Genet.*, **5**, e1000716.
15. Laganà,A., Forte,S., Giudice,A., Arena,M.R., Puglisi,P.L., Giugno,R., Pulvirenti,A., Shasha,D. and Ferro,A. (2009) miRò: a miRNA knowledge base. *Database*, **2009**, bap008.
16. Gerlach,D., Kriventseva,E.V., Rahman,N., Vejnar,C.E. and Zdobnov,E.M. (2009) miROrtho: computational survey of microRNA genes. *Nucleic Acids Res.*, **37**, D111–7.
17. Zhang,Z., Yu,J., Li,D., Zhang,Z., Liu,F., Zhou,X., Wang,T., Ling,Y. and Su,Z. (2010) PMRD: plant microRNA database. *Nucleic Acids Res.*, **38**, D806–813.
18. Maselli,V., Di Bernardo,D. and Banfi,S. (2008) CoGemiR: a comparative genomics microRNA database. *BMC Genomics*, **9**, 457.
19. Li,S.C., Shiau,C.K. and Lin,W.C. (2008) Vir-Mir db: prediction of viral microRNA candidate hairpins. *Nucleic Acids Res.*, **36**, D184–D189.

20. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
21. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
22. Hsu,P.W., Huang,H.D., Hsu,S.D., Lin,L.Z., Tsou,A.P., Tseng,C.P., Stadler,P.F., Washietl,S. and Hofacker,I.L. (2006) miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. *Nucleic Acids Res.*, **34**, D135–D139.
23. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for "expressed sequence tags". *Nat Genet.*, **4**, 332–333.
24. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
25. The UniProt Consortium. (2011) Ongoing and future developments at the universal protein resource. *Nucleic Acids Res.*, **39**, D214–D219.
26. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
27. Huang,J., Hao,P., Chen,H., Hu,W., Yan,Q., Liu,F. and Han,Z.G. (2009) Genome-wide identification of *Schistosoma japonicum* microRNAs using a deep-sequencing approach. *PLoS One*, **4**, e8206.
28. Hao,L., Cai,P., Jiang,N., Wang,H. and Chen,Q. (2010) Identification and characterization of microRNAs and endogenous siRNAs in *Schistosoma japonicum*. *BMC Genomics*, **11**, 55.
29. Pang,K.C., Stephen,S., Dinger,M.E., Engström,P.G., Lenhard,B. and Mattick,J.S. (2007) RNAdb 2.0—an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.*, **35**, D178–D182.
30. Lazzari,B., Caprera,A., Cestaro,A., Merelli,I., Del Corvo,M., Fontana,P., Milanesi,L., Velasco,R. and Stella,A. (2009) Ontology-oriented retrieval of putative microRNAs in *Vitis vinifera* via GrapeMiRNA: a web database of de novo predicted grape microRNAs. *BMC Plant Biol.*, **9**, 82.
31. Wang,X. (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA*, **14**, 1012–1017.
32. Hsu,S.D., Lin,F.M., Wu,W.Y., Liang,C., Huang,W.C., Chan,W.L., Tsai,W.T., Chen,G.Z., Lee,C.J., Chiu,C.M. *et al.* (2010) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D169.
33. Gustafson,A.M., Allen,E., Givan,S., Smith,D., Carrington,J.C. and Kasschau,K.D. (2005) ASRP: the Arabidopsis small RNA project database. *Nucleic Acids Res.*, **33**, D637–D640.
34. Ruepp,A., Kowarsch,A., Schmidl,D., Buggenthin,F., Brauner,B., Dunger,I., Fobo,G., Frishman,G., Montrone,C. and Theis,F.J. (2010) PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol.*, **11**, R6.
35. Yang,Z., Ren,F., Liu,C., He,S., Sun,G., Gao,Q., Yao,L., Zhang,Y., Miao,R., Cao,Y. *et al.* (2010) dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics*, **11(Suppl. 4)**, S5.
36. Schmeier,S., Schaefer,U., MacPherson,C.R. and Bajic,V.B. (2011) dPORE-miRNA: polymorphic regulation of microRNA genes. *PLoS One*, **6**, e16657.
37. Hiard,S., Charlier,C., Coppieters,W., Georges,M. and Baurain,D. (2010) Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic Acids Res.*, **38**, D640–D651.
38. Zhang,Y., Guan,D.G., Yang,J.H., Shao,P., Zhou,H. and Qu,L.H. (2010) ncRNAimprint: a comprehensive database of mammalian imprinted non-coding RNAs. *RNA*, **16**, 1889–1901.
39. Zhang,B.H., Pan,X.P., Wang,Q.L., Cobb,G.P. and Anderson,T.A. (2005) Identification and characterization of new plant microRNAs using EST analysis. *Cell Res.*, **15**, 336–360.
40. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
41. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
42. Morgulis,A., Gertz,E.M., Schäffer,A.A. and Agarwala,R. (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.*, **13**, 1028–1240.
43. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chem.*, **125**, 167–188.
44. Batuwita,R. and Palade,V. (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, **25**, 989–995.
45. Alexiou,P., Maragkakis,M., Papadopoulos,G.L., Reczko,M. and Hatzigeorgiou,A.G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**, 3049–3055.
46. Schwab,R., Palatnik,J.F., Riester,M., Schommer,C., Schmid,M. and Weigel,D. (2005) Specific effects of microRNAs on the plant transcriptome. *Dev. Cell*, **8**, 517–527.
47. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
48. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X Version 2.0. *Bioinformatics*, **23**, 2947–2948.
49. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
50. Yang,J.H., Shao,P., Zhou,H., Chen,Y.Q. and Qu,L.H. (2010) deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res.*, **38**, D123–D130.
51. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

PCP
PLANT & CELL PHYSIOLOGY

ERISdb: A Database of Plant Splice Sites and Splicing Signals

Michał Wojciech Szcześniak[1,*], Michał Kabza[1], Rafał Pokrzywa[2], Adam Gudyś[2] and Izabela Makałowska[1,*]

[1]Laboratory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University, 61-614 Poznan, Poland
[2]Institute of Informatics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland
*Corresponding authors: Michał Wojciech Szcześniak, E-mail, miszcz@amu.edu.pl; Izabela Makałowska: E-mail, izabel@amu.edu.pl.
(Received November 7, 2012; Accepted January 2, 2013)

Splicing is one of the major contributors to observed spatio-temporal diversification of transcripts and proteins in metazoans. There are numerous factors that affect the process, but splice sites themselves along with the adjacent splicing signals are critical here. Unfortunately, there is still little known about splicing in plants and, consequently, further research in some fields of plant molecular biology will encounter difficulties. Keeping this in mind, we performed a large-scale analysis of splice sites in eight plant species, using novel algorithms and tools developed by us. The analyses included identification of orthologous splice sites, polypyrimidine tracts and branch sites. Additionally we identified putative intronic and exonic *cis*-regulatory motifs, U12 introns as well as splice sites in 45 microRNA genes in five plant species. We also provide experimental evidence for plant splice sites in the form of expressed sequence tag and RNA-Seq data. All the data are stored in a novel database called ERISdb and are freely available at http://lemur.amu.edu.pl/share/ERISdb/.

Keywords: MicroRNA • Splice sites • Splicing signals • U12 introns.

Abbreviations: EST, expressed sequence tag; miRNA, microRNA; PPT, polypyrimidine tract; PWM, position weight matrix; RACE, rapid amplification of cDNA ends; SRE, splicing regulatory element.

Introduction

One of the most prominent features of eukaryotic genes is that they possess quite a complex structure, which is attributed to the presence of spliceosomal introns. In photosynthetic eukaryotes, the vast majority of protein-coding genes (up to 90%) contain introns (Barbazuk and McGinnis 2008, Labadorf et al. 2010). The presence of introns has notable functional consequnces for the cell. For instance, the primary transcripts are longer, up to hundreds of thousands of bases, than the actual coding sequence, which requires a lot more substrates and energy in the process of transcription. It therefore does not seem surprising that introns play fundamental roles, being key elements in the process of alternative splicing, which is a critical contributor to the transcriptome and proteome complexity in most eukaryotes. In the process of alternative splicing, primary transcripts from intron-containing genes are spliced by differential selection of splice sites in a spatiotemporal manner, leading to production of multiple mature mRNAs from a single gene (Pan et al. 2008, Kalsotra and Cooper 2011). Protein isoforms produced in this way may possess altered functions (Stamm et al. 2005). Additionally, alternative splicing plays a key role in gene regulation through regulated production of splice variants with a premature termination codon that are degraded in nonsense-mediated decay (Palusa and Reddy 2010). It also may lead to production of alternative splice forms that contain or lack microRNA (miRNA) target sequences (Tan et al. 2007). As a result, post-transcriptional regulation through alternative splicing constitutes an elaborate mechanism to fine-tune gene expression and provide proteome diversity.

The (alternative) splicing is performed by a large ribonucleoprotein complex called the spliceosome. In many eukaryotes, including most plant and animal species, there are two types of spliceosomes, the major and minor ones. The major spliceosome is responsible for splicing of the vast majority of introns in both plants and animals, and interacts with so-called U2 introns. The latter participates in splicing of U12 introns and splices out ~0.3% of introns in human (Lavine and Durbin 2001) and ~0.15% in *Arabidopsis thaliana* (Zhu and Brendel 2003). U2 introns usually possess GT and AG terminal dinucleotides at their 5′ and 3′ termini, respectively. In U12 introns, the terminal dinucleotides are more divergent, with GT–AG and AT–AC being the most prevalent dinucleotides. A distinctive feature of U12 introns is that they possess a well-conserved donor site with a consensus sequence RTATCCTTT as well as a distinct branch site with a consensus sequence TTCCTT RAY (Dietrich et al. 1997). Relatively high evolutionary conservation of U12 introns suggests that they might play important roles in the cell, and indeed they have been implicated in several

Plant Cell Physiol. 54(2): e10(1–8) (2013) doi:10.1093/pcp/pct001, available online at www.pcp.oxfordjournals.org

molecular phenomena (Patel et al. 2002, Hirose et al. 2004, Hastings et al. 2005). So far, a few studies have been performed to search for U12 introns in eukaryotes, but there is still little known about them in plants (Alioto 2007).

Correct recognition of splice sites by the spliceosome is critical for proper excision of introns from a primary transcript. There are three canonical splicing signals that guide a spliceosome to splice sites. The first one constitutes splice sites themselves, one at the 5′ end of an intron (donor site) and the second at the 3′ end (acceptor site), with the most important roles being played by highly conserved intronic terminal dinucleotides. The two latter elements are the polypyrimidine tract (PPT), rich in C and U and usually 15–20 bases long, and the branch site containing a so-called branch point nucleotide, required to produce a lariat intermediate, a key step in the splicing process. In plants, there is another element, the UA-rich tract, which is required for effective splicing of U2 introns (Goodall and Filipowicz 1989) and improves the splicing of U12 introns (Lewandowska et al. 2004). Although these core sequence features are quite conserved across species, they alone are not sufficient to define exons and introns and recruit the splicing machinery. In fact, some plant introns lack PPTs or UA tracts. Additional intronic and exonic sequences, usually referred to as splicing regulatory elements (SREs) or *cis*-acting elements, are important for both constitutive and alternative splicing. The SREs function as either splicing enhancers or suppressors and affect splice site choice by interacting with proteins that are collectively called *trans*-acting factors. Depending on the location of SREs and their effect on splicing, they are grouped into four classes: intronic splicing enhancers (ISEs), intronic splicing silencers (ISSs), exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs) (Wang and Burge 2008, Wang et al. 2009).

In recent years, much attention was paid to a class of non-coding RNAs called miRNAs, which resulted in identification of thousands of miRNAs in hundreds of plant, animal and protist species (Lin et al. 2009, Szcześniak et al. 2012). These small RNA molecules post-transcriptionally regulate the expression of thousands of genes in plants and animals either by transcript cleavage (Reinhart et al. 2002) or by translational repression (Lai 2002), and are key players in stress-related, developmental and signaling processes (Kedde and Agami 2008, Leung and Sharp 2010, O'Neill 2011). As a result, much hope is placed in untangling mechanisms of miRNA function and harnessing them in a number of applications in biotechnology, medicine or molecular biology. This depends on a satisfactory understanding of their biogenesis and expression regulation, which in turn might require uncovering their exon–intron structures and the confines of a gene. This is true at least for plant miRNAs, as animal miRNAs supposedly lack introns. However, little has been done to determine miRNA gene structures in plants, except for single analyses in *A. thaliana* (Szarzynska et al. 2009) and *Vitis vinifera* (Mica et al. 2010).

Keeping in mind the above-mentioned insufficiency of plant data, we have performed large-scale analyses of plant introns and splice sites in eight species: *A. thaliana*, *Chlamydomonas reinhardtii*, *Glycine max*, *Oryza sativa*, *Physcomitrella patens*, *Selaginella moellendorffii*, *V. vinifera* and *Zea mays*. These included a search for PPTs, UA-rich tracts, and branch sites, determining expressed sequence tags (ESTs) and RNA-Seq reads that support annotated splice sites, identification of orthologous splice sites, finding novel U12 introns and uncovering miRNA gene structures. In order to accomplish these tasks, we developed novel tools and algorithms, including a highly accurate classifier for U12 intron search and a tool for identification of splicing *cis*-regulatory elements (both available for download). We also collected some external data from published papers and databases to complement our findings. All data are deposited in a newly created online database with a user-friendly interface. We called the resource ERISdb and made it available at http://lemur.amu.edu.pl/share/ERISdb/.

Methods

Data download

Genome, protein and transcript sequences as well as annotation data and orthologous gene relationships were downloaded from Ensembl Plants release 15 (Vilella et al. 2009, Kersey et al. 2010). EST sequences were downloaded from dbEST release 120701 (Boguski et al. 1993), and miRNA sequences from miRBase 19 (Kozomara and Griffiths-Jones 2011). RNA-Seq data were retrieved from NCBI's Sequence Read Archive (Leinonen et al. 2011) and included seven libraries: SRP002417 (whole plant, *P. patens*), SRP002417 (aerial tissue, *S. moellendorffii*), DRS000668 (tissue pool, *O. sativa*) and SRP011480 (*Z. mays*, four libraries: immature tassel, seedling root, seedling shoot and unpollinated ear tip). U12 splice site data were downloaded from U12DB (Alioto 2007). The data comprised U12 splice site and branch site sequences of 17 animal species and *A. thaliana*. RACE (rapid amplification of cDNA ends) sequencing products for *A. thaliana* miRNAs were retrieved from NCBI, based on data provided in the corresponding publication (Szarzynska et al. 2009). Additionally, from the study on *V. vinifera* miRNAs (Mica et al. 2010), we downloaded deep sequencing data that support introns in miRNA genes. Finally, when mapping SREs to splice sites, sequences of exon splicing enhancers identified by Pertea et al. (2007) were used along with putative SREs identified in this research.

Branch site search

For branch site identification, we used a Perl script developed by Schwartz et al. (2008). The script operates in three steps. First, it scans the 100 nucleotides (nt) upstream of the 3′ splice site and identifies the following heptamers: NNYTRAY, NNCTYAC, NNRTAAC and NNCTAAA, which were previously identified in hemiascomycetous yeast (Bon et al. 2003) and *Schizosaccharomyces pombe* (http://www.sanger.ac.uk/Projects/S_pombe/intron.shtml). Then, it scores each

heptamer according to the number of mismatches from the optimal consensus of TACTAAC and, finally, discards all introns in which the best-scoring hit is not the most downstream one. Although the last step discards a relatively large fraction of introns, it is believed to reduce the false-positive rate significantly (Schwartz et al. 2008).

Identification of PPTs and UA-rich tracts

PPTs and UA-rich tracts were searched in intronic regions of up to 50 bases upstream of the 3′ splice site. Additionally—in the case of PPTs—we required that they end within the last 10 bases of an intron. The same parametres for the PPT search were previously applied in a large-scale analysis of eukaryotic splice sites (Schwartz et al. 2008) as PPTs beyond these confines are unlikely to be functional (Coolidge et al. 1997, Kol et al. 2005). The algorithm implemented in Python searched for the longest string with the C + U (in the case of PPTs) or A + U (for UA tracts) composition exceeding 85%. Moreover, the tracts were required to be at least five bases long. If a PPT was found downstream of a branch site, it was considered as a 'putative PPT', otherwise we called it a 'CT tract'.

Search for splicing *cis*-regulatory elements

We wrote a Java program to search for splicing motifs in intronic and exonic sequences. In the first step, by sliding the window of size m (input parameter) across the input sequences, the program identifies all m-length substrings and counts the number of their exact occurrences in the input sequences. In the next step, all substrings are connected to each other by similarity. Two substrings k and h are connected if and only if the number of similar nucleotides between k and h is equal to or greater than the value of the input parameter s. Each distinct substring with all connected substrings forms a motif represented by the position weight matrix (PWM). To identify motifs that are over-represented in the input sequences, the program calculates the number of matches of corresponding PWMs in the input sequences and in the reference sequence. The reference sequence can be provided as an input parameter to the program or is randomly generated with the probability of each residue taken from the input sequences. To achieve satisfactory performance, we used the PWM matching algorithm proposed by Beckstette et al. (2004). Nevertheless, we decided to replace the originally used enhanced suffix arrays with the index based on the wavelet tree data structure (Grossi et al. 2003), which is more efficient than a suffix array and consumes fewer memory resources. Next, for each motif, the program computes the ratio of the number of matches of PWM in the input sequences to the number of matches in the reference sequence. Motifs with a ratio that is greater than the value of the input parameter v are qualified to the final step, where the program calculates the percentage of input sequences matching the corresponding PWM. If the calculated percentage is greater than the value of the input parameter d, the motif is reported as over-represented in the input sequences.

In this research, we focused on identification of 7- and 8-mers (parametres $m = 7$, $s = 6$ and $m = 8$, $s = 7$) in all intronic sequences, separately for each species and separately for 3′ and 5′ splice sites as well as independently for short (<120 bases) and long introns. Here, we required that the level of random occurrences of deviation (parameter v) was at least 3 and the fraction of sequences with at least one occurrence of the motif was 0.05 (parameter d).

Over-representation of the discovered motifs was further verified statistically. Let K denote the number of occurrences of a particular motif in all analyzed intronic regions (it is a sum of occurrences of all sequences from a corresponding cluster) and N be the number of all m-length windows in the data. Additionally, let p indicate the theoretical probability of finding a particular motif in a randomly generated sequence (the distribution of ACGT symbols is taken from the data). For each potential *cis*-regulatory element, we calculated the P-value as the probability of a motif occurring in the analyzed intronic regions by chance $i \geq K$ times:

$$p\text{-}value = P(i \geq K) = 1 - P(i < K) = 1 - \sum_{i=0}^{k-1} \binom{N}{i} p^i (1-p)^{N-i}$$

Obtained P-values were always <0.0001. Owing to the fact that SREs are supposed to be over-represented, this is strong statistical support for us having found functional SREs. However, we were unable to estimate the false-negative rate as there is no comprehensive and experimentally verified set of plant SREs.

Independently, we searched for *cis*-regulatory elements within 50 nt of exonic sequences surrounding introns that are retained in at least one splice form. In this case, due to a limited number of sequences, the calculations were performed simultaneously for all species and all introns of this type, and the v parameter was set to 4.

Additionally, we used exonic splicing enhancer predictions in *A. thaliana* by Pertea et al. (2007), which include 84 hexamers. Thirty-five of these motifs were shown experimentally to affect splicing. We mapped the hexamers to exonic sequences located 50 bases upstream and downstream of the splice sites in all analyzed plant species.

Providing experimental support for splice sites

In order to provide experimental evidence for the analyzed splice sites, we mapped EST sequences to all transcripts in the corresponding plant species and referred the mapping results to annotated splice site coordinates. We used Megablast here and required that the identity is at least 98% and that the alignment length constitutes at least 90% of the EST length. We made sure that the ESTs do not map to transcripts of other genes with higher BLAST scores. We kept only those alignments where ESTs overlapped with known exon–exon junction(s) and a part of the alignment spanned over at least 10 bases of both exons. Additionally, we used RNA-Seq reads from seven libraries in four plant species (**Supplementary Table S1**).

The reads were filtered using our in-house Python scripts, and only those with a minimum quality score of 20 over 95% of bases were kept. Reads spanning known introns were detected using TopHat2 (Trapnell et al. 2009), with the gene model annotations supplied by the Ensembl Plants release 15 (Kersey et al. 2010).

Identification of orthologous splice sites

We extracted 5′ and 3′ splice sites from orthologous genes; these were truncated to contain up to 100 bases of exonic sequence and up to 120 bases of intronic sequence, depending on the total exon and intron lengths. In order to find out which splice sites are orthologous, we performed a BLAST search and filtered out the cases where the minimum e-value for an alignment was 1 e-5, both query and subject were in the same orientation in BLAST HSPs (high-scoring segment pairs), and query and subject represented splice sites of orthologous genes. The orthologous splice sites were subsequently re-aligned using ClustalW (Thompson et al. 1994) to obtain alignments of the whole splice site sequences, later used for visualization purposes in the database.

U12 intron search

U12 intron data from U12DB served as a positive data set (true U12 introns). The negative data set was comprised of *C. elegans* splice sites as this nematode species is devoid of the U12 spliceosome and U12 introns (Burge et al. 1998). Using splice site and branch site sequences from the positive data set, we generated PWMs and applied them to calculate a set of features for both positive and negative data sets. The set of 50 features corresponded to 49 nucleotide positions in splice sites and branch sites that were considered, as well as the position of the branch site in relation to the 3′ splice site. The 49 nucleotide positions included 10 nt of exonic and 15 nt of intronic sequence at the 5′ splice site, 6 nt from exonic and 9 nt from the intronic sequence at the 3′ splice site, as well as 9 nt from the branch site. We used implementation of the random forest machine learning algorithm in Weka 3.6.8 (Frank et al. 2004). The settings were default except for the number of trees, set from 10 to 100. We also used an in-house plugin written in Java to balance between sensitivity and specificity by thresholding the class conditional probability function, and we set the plugin to higher specificity at the cost of sensitivity. Using this approach, we generated a classification model that was finally applied to discriminate between U2 and U12 introns in plants. We performed the analysis for all species except for *C. reinhardtii* which does not contain U12 introns (Bartschat and Samuelsson 2010) and *A. thaliana* that was used to train the classifier, and U12DB data were used instead. In the 10-fold cross-validation performed on the full data set, the procedure yielded an exceedingly high sensitivity and specificity, 99.517% and 99.997%, respectively. To assess the classification performance further, we ran the classifier for *C. reinhardtii* splice sites and obtained seven false positives from the set of 102,382 redundant introns.

Splice sites in microRNA genes

In the first step, we identified EST sequences that correspond to known pre-miRNAs from miRBase (Kozomara and Griffiths-Jones 2011). We searched the ESTs from dbEST (Boguski et al. 1993) using Megablast (Altschul et al. 1990) and required that the identity is $\geq$97% and that the EST sequence contains at least 90% of known pre-miRNA sequence. The selected ESTs were subsequently mapped to the corresponding genome using Splign (Kapustin et al. 2008) with default settings. The alignments were finally checked manually to remove cases where ESTs came from the antisense strand and to improve the alignment if the splice site was broken because of an imperfection of the EST alignment software. For *A. thaliana* we additionally downloaded the sequences from RACE experiments (Szarzynska et al. 2009) and used a similar approach to that in the case of ESTs. Here, however, it was not required that the RACE product contains 90% of pre-miRNA sequence as the sequences have already been assigned to corresponding pre-miRNAs and often contained $<$90% of pre-miRNA sequence.

WebLogos and PWMs

WebLogos were generated using the WebLogo (Crooks et al. 2004) command line client with default settings except for -c (for color logos) and -w (logo width, depending on the number of bases in a logo). All splice sites as well as branch sites of both U2 and U12 introns were used. PWMs were generated from the same data sets as WebLogos using in-house Python scripts.

ERISdb: Database Composition and Usage

Browse

The Browse page gives access to all 1,610,648 splice sites in the analyzed plant species. In order to make browsing through this vast amount of data more straightforward, we divided it into three steps. (i) Species selection. Here, one can choose from eight species: *A. thaliana*, *C. reinhardtii*, *G. max*, *O. sativa*, *P. patens*, *S. moellendorffii*, *V. vinifera* and *Z. mays*. (ii) Transcript selection. In this page, one can browse through all transcripts in the selected species. Along with transcript names, also the description, genomic coordinates and transcript biotype are shown. One can filter the data by selecting transcript biotype, chromosome or specifying the gene name or a keyword (this searches through the description data). Alternatively, one can input the transcript name and proceed directly to the third step. (iii) Splice site selection. In this view, two integrated parts help select the splice site of interest: the splice site selection panel with numbered 5′ and 3′ splice sites and a graphical representation of the transcript exon–intron structure. When hovering over the splice site selection panel, the green line marker moves towards the corresponding splice site in the gene structure, thus enabling the user to select the desired

splice site precisely. After clicking, the user is redirected to the splice site data page.

Splice site data page

This page displays the following data. (i) General information on the splice site, e.g. intron type (U2 or U12) or splice site sequence. (ii) Orthologs: an alignment of the selected splice site and orthologous sequences. Direct links to orthologous splice site data are provided. (iii) ESTs: ESTs that support the splice site are shown in alignment with the genomic sequence. The intronic sequences are truncated to 10 bases at both termini for display purposes and the alignment itself is truncated to 100 bases. One can also view full alignments, where the splice site of interest is marked in yellow as it often happens that the alignment spans over more than one splice site. (iv) RNA-Seq: detected by TopHat2, a block of exonic sequences that spans over a selected splice site and is supported by RNA-Seq data is marked with yellow. (v) PPT, BS, UA tract: these features are displayed together as alignment to the intronic sequence. In the case of a branch site (BS), the score value is provided which details the similarity of the identified branch site to the fungal consensus sequence of TACTAAC. (vi) *cis*-Regulatory elements: here, alignments of exonic and intronic SREs to the splice site sequence are provided. Putative intronic SREs, exonic SREs identified by Pertea et al. (2007) and exonic SREs identified for retained introns are marked with distinct colors. In the case of intronic SREs, upon clicking a selected sequence, the associated information is displayed: a WebLogo, PWM and sequence score based on the PWM.

U12 introns

This page grants direct access to all U12 introns identified in seven plant species. The data can be filtered by species or type of terminal dinucleotides (GT–AG, AT–AC or other). One can also select a gene of interest from a drop-down list. The displayed data include 5′ and 3′ splice site sequences as well as orthologous genes with a U12 intron. The highly conserved intronic sequence at the 5′ end and terminal dinucleotide at the 3′ end are marked with orange, while branch sites are marked with yellow.

MicroRNA genes

Here, one can choose from ERISdb predictions (45 miRNAs), Ensembl annotations (eight miRNAs) or miRNA introns supported by RNA-Seq (three miRNAs). Upon selection of an miRNA of interest from ERISdb predictions, experimentally supported gene structures are displayed. The view includes alignments of three sequences: EST, pre-miRNA and the corresponding genomic sequence. The intronic sequences, if longer than 100 bases, are truncated to 50 bases at both ends. In the case of Ensemble Plant miRNAs, the user is redirected to the splice site data page in ERISdb. Finally, for three *V. vinifera* miRNAs with RNA-Seq support for introns, we provide an alignment of reads to the predicted splice sites.

WebLogos

WebLogos for 3′ and 5′ splice sites as well as branch sites in corresponding species, separately for U2 and U12 introns, are presented here. Additionally, the corresponding PWMs for splice sites and branch sites are available.

Other pages

ERISdb supports download of various types of computed data as well as selected software used during the analyses. The data files along with short description are available from the Download page. The Help page guides the user through the database structure and usage and provides examples for clarity.

Discussion and Conclusions

There are three signals that play key roles in correct splicing of introns: (i) 5′ and 3′ splice sites; (ii) PPTs and/or UA-rich tracts in plants; and (iii) branch sites. De novo identification of splice sites was not a subject of this research as we relied on the gene structure annotation from Ensemble Plants. Instead, we provide experimental support for splice sites (see the Methods). As for PPTs, it still appears unclear whether they always play the same roles as in metazoan introns, but growing experimental evidence supports the idea that at least in some genes PPTs are essential for effective splicing and are located in the canonical position between a putative branch site and the 3′ splice site (Brown et al. 2002) though they are less pronounced (Schwartz et al. 2008) and in fact they are lacking in a number of introns. In this research, we classified the CU-rich regions which were found into two categories, keeping in mind that an upstream branch point is required for PPTs to be functional (Simpson et al. 2004): a putative PTT if there was branch site found upstream, or a CU tract if the tract overlapped with a branch site or a branch site was not found.

In plant introns, the picture is even more complex, as they often possess UA tracts, crucial for effective splicing in both U2 and U12 introns (Goodall and Filipowicz 1989, Lewandowska et al. 2004). Although they may be found anywhere throughout the intron, some of them are located between the putative branch site and the 3′ splice site, and it was demonstrated in the potato invertase gene that the same U-rich sequence can function as either a PPT or a UA-rich element, depending on the presence or absence of a functional branch point upstream (Simpson et al. 2004). Keeping this in mind, we present branch sites, PPTs and UA tracts together in an alignment with intronic sequence in the splice site data page to make the data more informative, but we refrain from considering UA-rich tracts as PPTs even if they fall into the canonical PPT location.

This study is the first attempt to search globally for *cis*-regulatory elements in plant introns. The only other study was focused on exonic splicing enhancers in *A. thaliana* (Pertea et al. 2007). We have identified putative intronic *cis*-acting elements in 797,423 plant splice sites, mostly at 3′ splice sites (**Table 1**). As more than half of the analyzed splice sites do

not possess discovered elements, we assume that this might be caused by the usage of exonic splicing elements instead. It is also possible that there are rare and less conserved intronic elements, which are difficult to discover. Additionally, we searched for putative exonic *cis*-acting elements that might account for intron retention events, and we uncovered two clusters of over-represented 8-mers that possess a consensus sequence of CGSCGCCG, where S stands for C or G. The sequence was found in exonic sequences associated with 2,393 (13.3%) retained introns. At the same time it could be found for as few as 15,404 (0.97%) remaining introns. We decided to focus on intron retention as it is common in plants and accounts for 56% of alternative splicing events in *A. thaliana* (Wang and Brendel 2006). Furthermore, the process might have considerable implications for post-transcriptional regulation of gene expression by affecting target sites for miRNAs, as predicted for humans (Tan et al. 2007), or through the process of nonsense-mediated mRNA decay (Lewis et al. 2003).

U12 introns in *A. thaliana* were first discovered over a decade ago (Shukla and Padgett 1999), yet since then little has been done to identify them in other plant species. Although they are relatively rare (detected in 0.22% of genes in this study), these introns are shown to be implicated in a number of phenomena, and the ability to discriminate between U2- and U12-type introns might be crucial to untangle splicing mechanisms and their regulation in some introns. To this end, using a machine learning approach (random forest algorithm), we generated a highly sensitive and specific classifier that was further used to search for U12 introns in six plant species. Our approach is superior to the widely used PWM-based approach as it is able to learn quite complex relationships between analyzed features based on positive (U12 introns) and negative (U2 introns) input data. In addition, we took advantage of a branch site position in a 3′ splice site, as it was suggested that this feature might possess discriminative power for a search for plant U12 introns (Lewandowska et al. 2004). Altogether we have identified U12 introns in 2,041 plant genes (**Table 1**; **Supplementary Table S2**), 82.4% of which were found to have orthologs with U12 introns as well. A fraction of the remaining U12 introns could possess U2 orthologs, as it is widely observed that U12 introns tend to convert to U2 introns (Lin et al. 2010).

Table 1 Summary of data stored in ERISdb

All introns	**1,610,648**
Supported by EST	704,744 (43.8%)
Supported by RNA-Seq	425,777 (26.4%)
CU tract found	856,226 (53.2%)
UA tract found	1,431,052 (88.8%)
Branch site found	920,403 (57.1%)
5′ splice site with intronic SREs	320,110 (19.9%)
3′ splice site with intronic SREs	477,313 (29.6%)
5′ splice site with identified ortholog	331,794 (20.6%)
3′ splice site with identified ortholog	331,498 (20.6%)
U12 introns	3,039 (0.22%)
SREs associated with retained introns[a]	15,404 (0.97% of non-retained introns)
	2,393 (13.30% of retained introns)
Hexamers from Pertea et al. (2007)[a]	1,422,984 (88.35%)
Intronic SREs	1,055,021 (65.50%)

[a]Calculated for exonic sequences associated with an intron.

In contrast to animal miRNAs, which are often derived from introns or untranslated regions of genes, plant miRNAs are usually transcribed from dedicated MIR genes. The transcription is usually guided by RNA polymerase II, and the resulting primary transcripts, called pri-miRNAs, are capped, polyadenylated and frequently contain introns (Szarzynska et al. 2009, Mica et al. 2010). Despite the fact that it might be important to know the miRNA gene structure for applications in biotechnology or molecular biology, very few studies have been performed to determine the miRNA gene structures in plants: in *A. thaliana* through direct cloning of pri-miRNAs (Szarzynska et al. 2009) and in *V. vinifera* using deep sequencing of the whole transcriptome (Mica et al. 2010). These studies generally show that in plant miRNAs there is a wide variation of transcript forms due to frequent events of alternative splicing and usage of alternative transcription starts sites and alternative polyadenylation sites, yet they uncovered only a fraction of miRNA splice sites as cloning is a laborious task and also because pri-miRNAs have relatively short physiological half-lives. However, previously we demonstrated that novel miRNAs can be successfully discovered from EST sequences, both in plants and in animals (Szczesniak et al. 2012). Encouraged by these results, in this study we performed analyses aimed at determination of miRNA gene structures based on alignment of EST sequences to the corresponding genome. We managed to find introns in 45 miRNA precursors 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 (**Fig. 1**). It is important to point out that when presented in ERISdb, exon–intron structures do

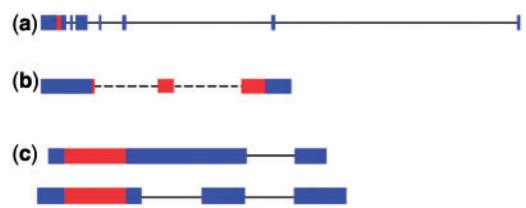


Fig. 1 Schematic representation of gene structures of three plant miRNAs predicted by us. Pre-miRNA is marked in red. (a) osa-miR156d is produced from a precursor containing six introns and is a record holder in this respect. (b) osa-MIR444c has exceedingly long introns (2,784 and 6,722 bases), and both of them reside in the pre-miRNA sequence. (c) ppt-MIR536c: an example of an alternatively spliced intron in an miRNA gene.

not correspond to the full gene architecture as this would require a search for promoters and/or transcription starts sites as well as polyadenylation sites, while we focused on identification of splice sites.

ERISdb is a comprehensive database of splice sites and splicing signals. In the current version, it provides data for eight plant species. The data stored in ERISdb include (i) general information on splice sites such as WebLogos, branch sites or PPTs; (ii) experimental evidence for annotated splice sites (ESTs and RNA-Seq data); (iii) orthologous splice sites; (iv) putative intronic and exonic *cis*-regulatory elements; (v) U12 introns; and (vi) introns in miRNA genes. This is the first database to cover such a diverse set of plant splice site-associated data, most of which are not available from any other resource. We also developed novel tools and made them available for download from the Download page. The novelty of database content along with the aesthetic, user-friendly design of the web interface and data download options should make ERISdb a very useful resource for further research in plant molecular biology.

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported by the National Science Centre [2011/01/N/NZ2/01653 to M.W.S. and I.M.]; the European Social Fund [UDA-POKL.04.01.01-00-106/09 to A.G. and R.P.].

References

Alioto, T.S. (2007) U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.* 35: D110–D115.

Altenhoff, A.M., Schneider, A., Gonnet, G.H. and Dessimoz, C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* 39: D289–D294.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.

Barbazuk, W.B., Fu, Y. and McGinnis, K.M. (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res.* 18: 1381–1392.

Bartschat, S. and Samuelsson, T. (2010) U12 type introns were lost at multiple occasions during evolution. *BMC Genomics* 11: 106.

Beckstette, M., Strothmann, D., Homann, R., Giegerich, R. and Kurtz, S. (2004) *PoSSuMsearch*: fast and sensitive matching of position specific scoring matrices using enhanced suffix arrays. *Proc. German Conf. Bioinformatics* 53: 53–64.

Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nat. Genet.* 4: 332–333.

Bon, E., Casaregola, S., Blandin, G., Llorente, B., Neuvéglise, C., Munsterkotter, M. et al. (2003) Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res.* 31: 1121–1135.

Brown, J.W., Simpson, C.G., Thow, G., Clark, G.P., Jennings, S.N., Medina-Escobar, N. et al. (2002) Splicing signals and factors in plant intron removal. *Biochem. Soc. Trans.* 30: 146–149.

Burge, C.B., Padgett, R.A. and Sharp, P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol. Cell* 2: 773–785.

Coolidge, C.J., Seely, R.J. and Patton, J.G. (1997) Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res.* 25: 888–896.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.* 14: 1188–1190.

Dietrich, R.C., Incorvaia, R. and Padgett, R.A. (1997) Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol. Cell* 1: 151–160.

Frank, E., Hall, M., Trigg, L., Holmes, G. and Witten, I.H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics* 20: 2479–2481.

Goodall, G.J. and Filipowicz, W. (1989) The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell* 58: 473–483.

Grossi, R., Gupta, A. and Vitter, J. (2003) High-order entropy-compressed text indexes. *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms* 841–850.

Hastings, M.L., Resta, N., Traum, D., Stella, A., Guanti, G. and Krainer, A.R. (2005) An LKB1 AT–AC intron mutation causes Peutz–Jeghers syndrome via splicing at noncanonical cryptic splice sites. *Nat. Struct. Mol. Biol.* 12: 54–59.

Hirose, T., Shu, M.D. and Steitz, J.A. (2004) Splicing of U12-type introns deposits an exon junction complex competent to induce nonsense-mediated mRNA decay. *Proc. Natl Acad. Sci. USA* 101: 17976–17981.

Kalsotra, A. and Cooper, T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* 12: 715–729.

Kapustin, Y., Souvorov, A., Tatusova, T. and Lipman, D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* 3: 20.

Kedde, M. and Agami, R. (2008) Interplay between microRNAs and RNA-binding proteins determines developmental processes. *Cell Cycle* 7: 899–903.

Kersey, P.J., Lawson, D., Birney, E., Derwent, P.S., Haimel, M., Herrero, J. et al. (2010) Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.* 38: D563–D569.

Kol, G., Lev-Maor, G. and Ast, G. (2005) Human–mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation. *Hum. Mol. Genet.* 14: 1559–1568.

Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39: D152–D157.

Labadorf, A., Link, A., Rogers, M.F., Thomas, J., Reddy, A.S. and Ben-Hur, A. (2010) Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii. *BMC Genomics* 11: 114.

Lai, E.C. (2002) Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* 30: 363–364.

Lavine, A. and Durbin, R. (2001) A computational scan for U-12 dependent introns in the human genome sequence. *Nucleic Acids Res.* 29: 4006–4013.

Leinonen, R., Sugawara, H. and Shumway, M. and International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. *Nucleic Acids Res.* 39: D19–D21.

Leung, A.K. and Sharp, P.A. (2010) MicroRNA functions in stress responses. *Mol. Cell* 40: 205–215.

Lewandowska, D., Simpson, C.G., Clark, G.P., Jennings, N.S., Barciszewska-Pacak, M., Lin, C.F. et al. (2004) Determinants of plant U12-dependent intron splicing efficiency. *Plant Cell* 16: 1340–1352.

Lewis, B.P., Green, R.E. and Brenner, S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA* 100: 189–192.

Lin, C.F., Mount, S.M., Jarmołowski, A. and Makałowski, W. (2010) Evolutionary dynamics of U12-type spliceosomal introns. *BMC Evol. Biol.* 10: 47.

Lin, W.C., Li, S.C., Lin, W.C., Shin, J.W., Hu, S.N., Yu, X.M. et al. (2009) Identification of microRNA in the protist Trichomonas vaginalis. *Genomics* 93: 487–493.

Matsushima, A., Kobayashi, N., Mochizuki, Y., Ishii, M., Kawaguchi, S., Endo, T.A. et al. (2009) OmicBrowse: a Flash-based high-performance graphics interface for genomic resources. *Nucleic Acids Res.* 37: W57–W62.

Mica, E., Piccolo, V., Delledonne, M., Ferrarini, A., Pezzotti, M., Casati, C. et al. (2009) High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera. *BMC Genomics* 10: 58.

O'Neill, L.A., Sheedy, F.J. and McCoy, C.E. (2011) MicroRNAs: the fine-tuners of Toll-like receptor signalling. *Nat. Rev. Immunol.* 11: 163–175.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S. et al. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38: D196–D203.

Palusa, S.G. and Reddy, A.S. (2010) Extensive coupling of alternative splicing of pre-mRNAs of serine/arginine (SR) genes with nonsense-mediated decay. *New Phytol.* 185: 83–89.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40: 1413–1415.

Patel, A.A., McCarthy, M. and Steitz, J.A. (2002) The splicing of U12-type introns can be a rate-limiting step in gene expression. *EMBO J.* 21: 3804–3815.

Pertea, M., Mount, S.M. and Salzberg, S.L. (2007) A computational survey of candidate exonic splicing enhancer motifs in the model plant Arabidopsis thaliana. *BMC Bioinformatics* 8: 159.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B. and Bartel, D.P. (2002) MicroRNAs in plants. *Genes Dev.* 16: 1616–1626.

Rouard, M., Guignon, V., Aluome, C., Laporte, M.A., Droc, G., Walde, C. et al. (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res.* 39: D1095–D1102.

Schwartz, S.H., Silva, J., Burstein, D., Pupko, T., Eyras, E. and Ast, G. (2008) Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.* 18: 88–103.

Shukla, G.C. and Padgett, R.A. (1999) Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. *RNA* 5: 525–538.

Simpson, C.G., Jennings, S.N., Clark, G.P., Thow, G. and Brown, J.W. (2004) Dual functionality of a plant U-rich intronic sequence element. *Plant J.* 37: 82–91.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D. et al. (2005) Function of alternative splicing. *Gene* 344: 1–20.

Szarzynska, B., Sobkowiak, L., Pant, B.D., Balazadeh, S., Scheible, W.R., Mueller-Roeber, B. et al. (2009) Gene structures and processing of Arabidopsis thaliana HYL1-dependent pri-miRNAs. *Nucleic Acids Res.* 37: 3083–3093.

Szcześniak, M.W., Deorowicz, S., Gapski, J., Kaczyński, Ł. and Makalowska, I. (2012) miRNEST database: an integrative approach in microRNA search and annotation. *Nucleic Acids Res.* 40: D198–D204.

Tan, S., Guo, J., Huang, Q., Chen, X., Li-Ling, J., Li, Q. et al. (2007) Retained introns increase putative microRNA targets within 3′ UTRs of human mRNA. *FEBS Lett.* 581: 1081–1086.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673–4680.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. and Birney, E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19: 327–335.

Walker, N.S., Stiffler, N. and Barkan, A. (2007) POGs/PlantRBP: a resource for comparative genomics in plants. *Nucleic Acids Res.* 35: D852–D856.

Wang, B.B. and Brendel, V. (2006) Genome wide comparative analysis of alternative splicing in plants. *Proc. Natl Acad. Sci. USA* 103: 7175–7180.

Wang, X., Wang, K., Radovich, M., Wang, Y., Wang, G., Feng, W. et al. (2009) Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing. *BMC Genomics* 10(Suppl. 1), S4.

Wang, Z. and Burge, C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 14: 802–813.

Zhu, W. and Brendel, V. (2003) Identification, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome. *Nucleic Acids Res.* 31: 1–12.

Bazy danych mikroRNA

STRESZCZENIE

Cząsteczki mikroRNA (miRNA) są małymi cząsteczkami RNA, pełniącymi kluczowe funkcje w regulacji wielu procesów komórkowych. Wiąże się z nimi nadzieje na rozwiązanie szeregu problemów współczesnej medycyny, biotechnologii i innych nauk biologicznych. Liczba projektów badawczych na ich temat, jak również publikacji, nieustannie rośnie, czemu towarzyszy przyrost danych oraz liczby baz danych. Aktualnie istnieje 51 baz danych miRNA, a ich liczba dynamicznie wzrasta, przez co coraz trudniej jest się po nich poruszać. Dodatkowo, niemałym problemem stały się takie zjawiska, jak niewystarczająca dokumentacja lub niska jakość danych czy interfejsu graficznego. Nadzieją na rozwiązanie tych problemów jest stale podnoszący się standard baz danych, tendencja do tworzenia zintegrowanych systemów bazodanowych, udostępniających dane zawarte w kilku tematycznych bazach danych w jednolitym formacie oraz systemów do automatycznego pozyskiwania informacji.

Michał Szcześniak

Elżbieta Owczarkowska

Jakub Gapski

Izabela Makałowska✉

Pracownia Bioinformatyki, Instytut Biologii Molekularnej i Biotechnologii, Wydział Biologii, Uniwersytet im. A. Mickiewicza, Poznań

✉Pracownia Bioinformatyki, Instytut Biologii Molekularnej i Biotechnologii, Wydział Biologii, Uniwersytet im. A. Mickiewicza w Poznaniu, ul. Umultowska 89, 61-614 Poznań; tel. (61) 829 5835; e-mail: izabel@amu.edu.pl

Artykuł otrzymano 12 grudnia 2011 r.
Artykuł zaakceptowano 21 stycznia 2012 r.

Słowa kluczowe: baza danych, mikroRNA, miRBase

Wykaz skrótów: EST – znaczniki sekwencji ulegających ekspresji; HMM – ukryte modele Markowa; NGS – sekwencjonowanie DNA nowej generacji; SVM – maszyna wektorów podpierających; UTR – region genu nieulegający translacji

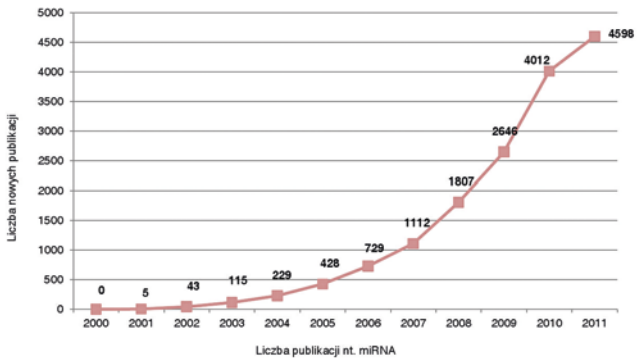
WPROWADZENIE

Cząsteczki miRNA są małymi, niekodującymi cząsteczkami RNA, pełniącymi liczne regulatorowe funkcje w komórkach zwierząt i roślin. miRNA regulują między innymi odpowiedź na stres środowiskowy [1], szlaki przekazywania sygnałów [2] czy procesy rozwojowe [3]. Liczne miRNA powiązano z chorobami u człowieka, takimi jak na przykład białaczka [4], rak trzustki [5] czy choroba Alzheimera [6]. Z tych powodów identyfikacja miRNA i poznawanie ich funkcji stało się niezwykle ważnym zagadnieniem nie tylko w biologii molekularnej, ale również w naukach medycznych i rolniczych.

Powstawanie dojrzałych cząsteczek miRNA przebiega w kilku etapach [7]. Najpierw gen miRNA ulega transkrypcji z udziałem polimerazy RNA II lub III. Powstały transkrypt, zwany pri-miRNA, podlega dalszej obróbce - cięciom katalitycznym, prowadzącym do otrzymania tzw. cząsteczki pre-miRNA, zwykle o długości 50-100 nukleotydów. Cząsteczka ta posiada charakterystyczną strukturę drugorzędową typu spinki do włosów (ang. *hairpin loop*, *stem-loop*), w której można wyróżnić część osiową (trzonek, ang. *stem*), zawierającą komplementarne do siebie fragmenty sekwencji oraz pętlę z niesparowanymi nukleotydami. Dojrzałe miRNA jest wycinane z części osiowej pre-miRNA, po czym wbudowane zostaje w kompleks wyciszający RISC (ang. *RNA-Induced Silencing Complex*), gdzie uczestniczy w procesach regulowania ekspresji genów na zasadzie cięcia docelowego mRNA bądź hamowania jego translacji.

W ciągu ostatniej dekady opracowano szereg algorytmów i programów komputerowych służących do identyfikacji i analizy funkcjonalnej miRNA *in silico* (metodami bioinformatycznymi). Jednocześnie pojawiły się innowacyjne techniki laboratoryjne, służące do odkrywania nowych miRNA, analizy ich poziomu ekspresji czy funkcji molekularnych. Skutkiem zwiększonego zainteresowania tematyką miRNA jest szybko narastająca ilość danych na ich temat. Znajduje to odzwierciedlenie w liczbie artykułów o miRNA, których w samym 2010 roku opublikowano 3995, wobec jedynie 5 w roku 2001 (Ryc. 1).

Głównym powodem tak dynamicznego przyrostu danych są analizy skoncentrowane na poszukiwaniach nowych miRNA, w oparciu o dane pochodzące z sekwencjonowania małych cząsteczek RNA technikami nowej generacji (NGS, ang. *Next Generation Sequencing*) oraz analizy *in silico* na poziomie genomów i transkryptomów. Znaczna ilość danych generowana jest również w trakcie komputerowych poszukiwań potencjalnych docelowych mRNA dla miRNA. Nie ulega zatem wątpliwości, że istnieje obecnie ogromne zapotrzebowanie na repozytoria, które pozwalałyby na przeglądanie, filtrowanie i analizę danych. W odpowiedzi na tę potrzebę powstało już 51 internetowych baz danych związanych z miRNA (PubMed, grudzień 2011), a ich liczba narasta coraz szybciej (Ryc. 2).



Rycina 1. Wzrost liczby publikacji na temat miRNA. Stan na grudzień 2011 r.

Niestety, przyrost liczby baz danych, choć ogólnie jest pozytywnym zjawiskiem, stwarza niemały kłopot użytkownikowi, chcącemu otrzymać potrzebne informacje. Wielokrotnie, aby uzyskać dostęp do istniejących danych na temat interesującej nas cząsteczki miRNA, trzeba przeszukać kilka a nawet kilkanaście baz danych. Brakuje także repozytorium baz danych miRNA, dzięki któremu użytkownik mógłby poznać wszystkie dostępne źródła oraz dowiedzieć się jakiego rodzaju dane są zdeponowane w konkretnej bazie. Niniejsze opracowanie, będące przeglądem istniejących, opublikowanych baz danych miRNA, wychodzi naprzeciw potrzebom użytkowników.

ŹRÓDŁA INFORMACJI W BAZACH DANYCH miRNA

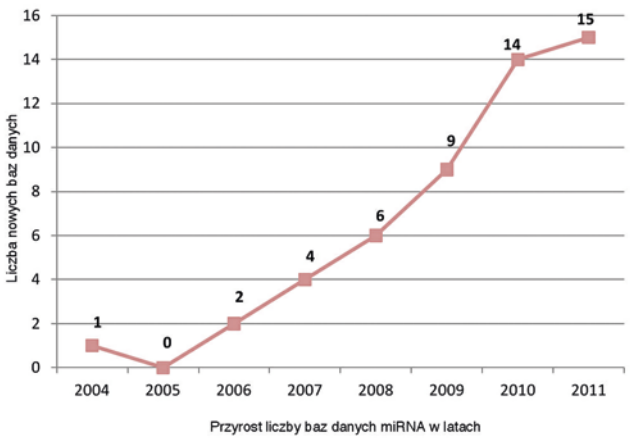
Ilość, jakość oraz charakter informacji gromadzonych w bazach danych miRNA ściśle zależy od metody, która posłużyła do ich otrzymania. Najogólniej, metody te można podzielić na *in silico* oraz eksperymentalne. Te pierwsze zwykle charakteryzują się wysoką czułością oraz niską specyficznością. Z tego powodu nieustannie rozwijane są nowe algorytmy pozwalające na obniżanie odsetka błędnych danych. Metody eksperymentalne z kolei, choć zwykle pozwalają na uzyskanie danych o dużo wyższej jakości, cechują się wysoką czaso- i pracochłonnością oraz wiążą się z wyższymi kosztami niż analizy bioinformatyczne. Dlatego jedynie znikoma część informacji zdeponowanych w bazach danych posiada potwierdzenie eksperymentalne.

IDENTYFIKACJA miRNA

W przypadku metod *in silico* służących do identyfikacji miRNA, możemy wyróżnić dwie główne grupy. Pierwsza skupia metody oparte na zachowaniu sekwencji i/lub struktury drugorzędowej miRNA pozwalają na identyfikację ortologów i paralogów znanych już miRNA, jednak nie znajdują zastosowania w przypadku poszukiwania miRNA należących do nowych rodzin. Druga grupa to algorytmy oparte na metodach nauczania maszynowego, takich jak ukryte modele Markowa (HMM, ang. *Hidden Markov Models*), maszyna wektorów podpierających (SVM, ang. *Supported Vector Machine*) czy sieci neuronowe [8, 9]. Ich zaletą jest zdolność do odkrywania nowych rodzin miRNA, jednakże muszą zostać odpowiednio wytrenowane na wysokiej jakości podzbiorze znanych miRNA – zarówno proces trenowania jak i przygotowanie odpowiedniego zbioru jest sporym wyzwaniem dla bioinformatyków.

Metody bioinformatyczne, zwłaszcza w przypadku analiz przeprowadzanych na poziomie genomów, zwykle prowadzą do otrzymania znacznego odsetka fałszywie pozytywnych wyników. W ostatnich latach sposobem na zmniejszenie tego problemu stało się wsparcie wyników wygenerowanych poprzez analizę sekwencji genomowych danymi pochodzącymi z eksperymentów NGS, które dostarczają informacji na temat ekspresji - a więc istnienia - dojrzałego miRNA. Tak działają m.in. miRDeep [10] oraz miRanalyzer [11]. Istnieją również algorytmy, które poszukują miRNA w znacznikach sekwencji ulegających ekspresji (EST, ang. *Expressed Sequence Tags*) [12, 13], bądź też wyłącznie w oparciu o dane z eksperymentów NGS [14]. W tym drugim przypadku zazwyczaj odkrywa się jedynie dojrzałe miRNA, jako że długość zsekwencjonowanych cząsteczek RNA jest mniejsza niż długość prekursorów miRNA.

Badania eksperymentalne, które coraz częściej są nierozerwalnie powiązane z analizami *in silico*, koncentrują się na dostarczeniu dowodu eksperymentalnego na istnienie miRNA, jak również służą do weryfikacji przewidzianych komputerowo funkcji miRNA. Eksperymentalne metody służące do wykazania obecności miRNA i poznania poziomu ich ekspresji muszą pokonać kilka trudności, takich jak mały rozmiar dojrzałych miRNA, brak ogonów poli(A) i znaczne podobieństwo sekwencji (a nawet identyczność) pomiędzy różnymi przedstawicielami tej samej rodziny miRNA. Wykorzystywane tutaj metody to qPCR (ang. *quantitative Polymerase Chain Reaction*), sekwencjonowanie, Northern blot oraz mikromacierze. Zostały one wykorzystane z powodzeniem w wielu badaniach, niemniej jednak posiadają liczne techniczne ograniczenia. Na przykład, niektóre z metod wymagają dużych ilości początkowego materiału (np. > 10 µg całkowitego RNA), podczas gdy inne - wzbogacenia RNA we frakcję małych RNA [15]. Poza tym niektórych metod nie można stosować w eksperymentach wielkoskalowych, jak Northern blot, który jest czasochłonny i dodatkowo charakteryzuje sie stosunkowo niską czułością.



Rycina 2. Wzrost liczby baz danych miRNA od 2004 r. Stan na grudzień 2011 r.

POZNAWANIE FUNKCJI miRNA

Poznanie docelowych mRNA dla miRNA ma kluczowe znaczenie podczas rozszyfrowywania ich funkcji regulatorowych. Stosowane tutaj metody bioinformatyczne można podzielić na dwie kategorie. Programy i metody należące do pierwszej z nich sprawdzają komplementarność pozycji 2-8 dojrzałego miRNA (tzw. regionu *seed*) z sekwencją 3′UTR regulowanego mRNA, energię swobodnę zwijania się kompleksu RNA-RNA oraz stopień zachowania między gatunkami sekwencji dojrzałego miRNA i jego miejsca wiązania na mRNA. W oparciu o te kryteria działają DIANA--microT [16], RNAhybrid [17] czy microInspector [18].

Druga kategoria metod oparta jest na nauczaniu maszynowym. Sztandarowym przykładem jest tutaj program PicTar [19], który skanuje przyrównanie wielu sekwencji 3′UTR w poszukiwaniu zachowanych w ewolucji fragmentów, komplementarnych do regionu *seed* miRNA, a następnie filtruje dupleksy mRNA-3′UTR na podstawie ich stabilności termodynamicznej. Ostatecznie, każdy kandydat otrzymuje punktację wyliczaną z wykorzystaniem ukrytych modeli Markowa (HMM).

Docelowe mRNA dla miRNA, które zostały przewidziane bioinformatycznie, powinny zostać potwierdzone metodami laboratoryjnymi. Najlepiej, jeśli uda się wykazać, że para miRNA-mRNA spełnia wszystkie cztery poniższe kryteria [20].

a) Fizyczna interakcja między miRNA a mRNA.
Podejście eksperymentalne polega tutaj najczęściej na wklonowaniu całej sekwencji 3′UTR potencjalnego genu docelowego do plazmidu z otwartą ramką odczytu dla lucyferazy lub GFP (białko zielonej fluorescencji, ang. *Green Fluorescent Protein*). Plazmid i miRNA są transfekowane do komórek gospodarza, a następnie mierzy się aktywność lucyferazy bądź luminescencję.

b) Koekspresja *in vivo* mRNA i miRNA.
Koekspresję można sprawdzać szeregiem metod służących do badania poziomu ekspresji mRNA, jak Northern blot czy qPCR. Z kolei by wykazać koekspresję tkankowospecyficzną lub nawet na poziomie pojedynczej komórki, stosuje się hybrydyzację *in situ*, wykorzystując m.in. znakowane digoksygeniną (DIG) antysensowne miRNA.

c) Wpływ miRNA na ilość produktu genu, będącego pod jego kontrolą.
Jeśli mRNA jest pod kontrolą określonego miRNA, ilość powstającego z niego białka powinna maleć w obecności miRNA. By to sprawdzić, komórki transfekuje się plazmidem zawierającym sekwencję, która udaje docelowe mRNA, 'podkradając' miRNA. Skutkiem tego, poziom prawdziwego docelowego mRNA oraz odpowiedniego białka powinien być wyższy niż w przypadku próby kontrolnej bez plazmidu. Ilość białka sprawdza się metodą Western blot. Alternatywnie, do wykazania różnic w ekspresji białka można wykorzystać test immunoenzymatyczny ELISA (ang. *Enzyme-Linked Immunosorbent Assay*).

d) Regulacja mRNA poprzez miRNA wiąże się z modyfikacją odpowiedniej funkcji biologicznej.
W zależności od regulowanego mRNA, często możliwe jest zaobserwowanie odpowiednich zmian fenotypowych. By je dostrzec, stosuje się tutaj szeroki wachlarz technik biologii molekularnej, jako że zmiany mogą dotyczyć na przykład szlaków przekazywania sygnałów, podziałów komórek, ich różnicowania, programowanej śmierci czy migracji komórek.

BAZY DANYCH miRNA

Dzięki analizom bioinformatycznym i molekularnym posiadamy coraz więcej informacji o miRNA i ich roli w szlakach metabolicznych i regulatorowych. Towarzyszy temu zapotrzebowanie na klasyfikowanie danych i stworzenie szybkich systemów służących do ich przechowywania i przeszukiwania. W rezultacie powstały liczne internetowe bazy danych miRNA, które kolekcjonują sekwencje miRNA, a także różnego rodzaju dane dotyczące ich biologii, włączając regulowane przez nie geny czy profile ekspresji w różnych tkankach. Poniżej omówionych zostało kilka baz danych miRNA, reprezentujących różne kierunki badań nad miRNA. Dodatkowo przedstawiono bazę miRNEST, która jest próbą integracji danych zawartych w różnych bazach danych w ramach jednolitego systemu bazodanowego. Krótka charakterystyka 51 opublikowanych do tej pory baz danych miRNA znajduje się w Tabeli 1.

Baza danych miRBase
Baza miRBase jest referencyjnym repozytorium sekwencji miRNA [21]. W wersji 17 obejmuje 16 772 sekwencje prekursorów miRNA (pre-miRNA) i 19 724 sekwencje dojrzałych miRNA ze 153 gatunków. Główne zadania spełniane przez tę bazę danych to utrzymywanie konsekwentnego systemu nazewnictwa nowych miRNA oraz pełnienie funkcji centralnego repozytorium opublikowanych sekwencji miRNA.

Każdy wpis w bazie, oprócz nazwy i sekwencji dojrzałego miRNA i pre-miRNA, zawiera numer dostępu, którego format jest stały i nie ulega zmianie pomiędzy wersjami bazy danych. W przypadku, gdy znane są sekwencje genomowe gatunku, udostępniane są współrzędne genomowe pre-miRNA. miRNA są dzielone na rodziny, w których obrębie znajdują się homologiczne geny miRNA. Użytkownik korzystający z miRBase może uzyskać dostęp do danych, poprzez i) przeglądanie wszystkich dostępnych wpisów w bazie, ii) przeszukiwanie na podstawie podobieństwa do zadanej sekwencji, iii) podanie przedziałów współrzędnych genomowych, iv) wyszukiwanie z użyciem słów kluczowych, v) masowe ściągnięcie wszystkich dostępnych danych. miRBase znajduje się pod adresem http://www.mirbase.org/.

Baza miRNEST
miRNEST [12] kolekcjonuje zwierzęce, roślinne i wirusowe miRNA. Centralną część tej bazy danych stanowią 10 004 miRNA ze 199 gatunków roślin oraz 221 gatunków zwierząt, zidentyfikowane metodą bioinformatyczną. Poszukiwanie nowych miRNA zostało przeprowadzone z wykorzystaniem sekwencji EST w oparciu o zachowanie

sekwencji dojrzałego miRNA (identyfikacja homologów znanych już miRNA). W przypadku 29 gatunków do pre-miRNA zmapowano odczyty pochodzące ze 192 bibliotek małych RNA pobranych z bazy GEO (ang. *Gene Expression Omnibus*) [22]. Dodatkowo, miRNEST został wyposażony w dane pochodzące z 13 zewnętrznych baz danych miRNA oraz dwu publikacji. Dane te dotyczą sekwencji miRNA (miRBase [21], microPC [13], PMRD [23]), ich ekspresji (phenomiR [24], dbDEMC [25]), polimorfizmów (Patrocles [26]), docelowych mRNA i funkcji miRNA (miRDB [27], miRTarBase [28], miRecords [29], PMRD [23], ASRP [30]), regulacji miRNA i ich promotorów (dPORE-miRNA [31], PMRD [23]), genomiki (CoGemiR [32]) oraz imprintingu (ncRNAimprint [33]). Wszystko to sprawia, że miRNEST jest obecnie największym repozytorium miRNA, obejmującym 544 gatunki, gromadzącym dane pochodzące z wielu źródeł i udostępniającym je w jednolitym formacie. Istnieje tutaj możliwość przeszukiwania i przeglądania danych, a także wykonywania podstawowych analiz, takich jak przeszukiwanie programem BLASTN [34] czy też przyrównanie wielu sekwencji programem ClustalW [35]. Baza jest dostępna pod adresem http://mirnest.amu.edu.pl.

Baza miRecords

Baza miRecords [29] jest zintegrowanym repozytorium informacji o interakcjach miRNA – gen docelowy u zwierząt. Dostępna pod adresem http://mirecords.biolead.org baza podzielona jest na dwie części, jedna jest poświęcona miejscom docelowym miRNA, które zostały potwierdzone eksperymentalnie, a druga – miejscom przewidzianym *in silico*. W części poświęconej potwierdzonym miejscom docelowym zdeponowane są informacje dotyczące 2 286 interakcji pomiędzy 548 miRNA a 1 579 genami docelowymi w 9 gatunkach zwierząt. Dane te pozyskano z literatury. Szczególny nacisk kładziony jest na systematyczną i dobrze zorganizowaną dokumentację eksperymentalnych dowodów na istnienie interakcji pomiędzy miRNA a danym genem. Druga część bazy miRecords poświęcona jest miejscom docelowym przewidzianym za pomocą aż 11 różnych programów bioinformatycznych (Tab. 1, pozycja 16). Dostęp do informacji o potwierdzonych i przewidzianych miejscach docelowych możliwy jest poprzez wyszukiwarki umieszczone na głównej stronie bazy. Interakcji miRNA-gen można szukać poprzez wprowadzenie nazwy gatunku, nazwy miRNA oraz opcjonalnie nazwy bądź numeru dostępu genu docelowego. Na stronie wyników wyszukiwania w każdym wierszu zawarta jest nazwa miRNA, nazwa i numer identyfikacyjny docelowego genu w bazie RefSeq [36], odnośnik do szczegółowych danych na temat interakcji miRNA z genem docelowym oraz informacje na temat interakcji miRNA-gen wygenerowane przez każdy z 11 programów. Główna strona bazy miRecords umożliwia dostęp do dokumentacji projektu, jak również pozwala ściągnąć zawartość bazy w postaci arkusza programu Excel.

Baza miR2Disease

Baza miR2Disease [37] jest repozytorium informacji na temat regulowania genów przez miRNA w różnych chorobach u człowieka. W tej adnotowanej przez człowieka bazie znajdują się 3 273 powiązania pomiędzy 349 sekwencjami miRNA a 163 chorobami, wprowadzone na podstawie przeanalizowania ponad 100 artykułów z serwisu PubMed. Każdy wpis zawiera szczegółowe informacje o związku miRNA-choroba, takie jak numer identyfikacyjny miRNA (ID), nazwa choroby, krótki opis występującego związku, wzór ekspresji miRNA i sposób w jaki analizowano ekspresję miRNA, eksperymentalnie potwierdzone docelowe mRNA dla miRNA oraz odnośniki do literatury. Wszystkie wpisy odnośnie terminologii chorób zostały zorganizowane według kontrolowanego słownictwa medycznego wykorzystującego Jednolity System Języka Medycznego (UMLS, ang. *Unified Medical Language System*) [38]. Oprócz łatwego w obsłudze systemu wyszukiwania za pomocą miRNA ID, nazwy choroby lub genów będących celem dla miRNA, prezentowane są użytkownikowi odnośniki do innych baz danych miRNA, zawierających dalsze informacje o wyszukiwanej frazie lub miRNA ID. Dodatkową zaletą systemu zaimplementowanego w miR2Disease jest funkcja przeszukiwania rozmytego (ang. *fuzzy search*), pozwalająca w połączeniu z kontrolowanym słownictwem medycznym na znalezienie w bazie informacji o związku miRNA-choroba nawet w przypadku, gdy użytkownik nie zna dokładnej nazwy choroby zapisanej w bazie danych. Użytkownik ma ponadto możliwość przesłania własnych informacji o powiązaniach miRNA-choroba, które po analizie przez kuratorów bazy mogą zostać dodane do miR2Disease. Baza miR2Disease jest dostępna pod adresem http://www.mir2disease.org/.

Baza PhenomiR

Baza PhenomiR (http://mips.helmholtz-muenchen.de/phenomir) jest źródłem informacji o ekspresji miRNA w chorobach i procesach biologicznych [24]. Zawarte w bazie dane pochodzą z 296 artykułów opisujących 542 przypadki deregulacji miRNA. Każdy przypadek zapisywany jest w bazie danych z takimi informacjami na temat miRNA i warunków eksperymentu, jak charakter zmiany ekspresji miRNA (wzrost lub spadek), metoda eksperymentalna (mikromacierze, RT-PCR, Northern blot), wskaźnik zmiany poziomu ekspresji miRNA czy pochodzenie próbki biologicznej. Każdemu wpisowi przyporządkowany jest numer PubMed ID oraz odnośnik do odpowiedniej publikacji w serwisie PubMed. Do adnotacji miRNA wykorzystane zostały dane z miRBase [21]. Adnotację chorób przeprowadzono w oparciu o OMIM Morbid Map (ang. *Online Mendelian Inheritance in Man Morbid Map*) [39], alfabetyczny spis chorób opisanych w OMIM. Przewagą OMIM Morbid Map nad takimi słownikami chorób, jak DO (ang. *Disease Ontology*) lub MeSH (ang. *Medical Subject Heading*) jest zawarcie dodatkowych informacji dotyczących choroby, wliczając cechy kliniczne, genetykę populacji i powiązane z nią geny. Adnotację procesów biologicznych przeprowadzono zgodnie z terminami zawartymi w Gene Ontology [40], natomiast w przypadku linii komórkowych i tkanek – wykorzystując BTO (ang. *Brenda Tissue Ontology*) [41].

INNE BAZY DANYCH miRNA

W Tabeli 1 wyszczególnionych zostało 51 baz danych poświęconych miRNA. Oprócz nich istnieją bazy danych o szerszym zakresie gromadzonych danych, które gromadzą dane na temat miRNA, jednak nie jest to podstawowe zadanie, jakie spełniają. Należy tutaj wspomnieć przede wszystkim przeglądarki genomowe (UCSC Genome Browser [42],

Tabela 1. Istniejące bazy danych miRNA.

	Nr	Nazwa bazy danych	Gatunki (*liczba*)	Rodzaj danych	Metody i źródła danych	PMID*
Sekwencje miRNA	1	**miRBase**	zwierzęta, rośliny, wirusy (153)	opublikowane miRNA, referencyjne źródło adnotacji miRNA	literatura, dane od użytkowników, program RNAfold	20205188
	2	**PMRD**	rośliny (123)	przewidziane *in silico* miRNA, ich ekspresja i mRNA docelowe	literatura, eksperymenty mikromacierzowe	19808935
	3	**microPC**	rośliny (125)	przewidziane *in silico* miRNA	algorytm do identyfikacji miRNA w sekwencjach EST	19660144
	4	**miROrtho**	zwierzęta (46)	przewidziane *in silico* miRNA	programy: R-COFFEE, RNAplfold, RNAalifold	18927110
	5	**Vir-Mir db**	wirusy (1491)	przewidziane *in silico* miRNA	program Srnaloop, baza danych NCBI	17702763
	6	**miRNAMap**	zwierzęta (13)	potwierdzone eksperymentalnie miRNA i ich mRNA docelowe	programy: miRanda, RNAhybrid, TargetScan, eksperymenty qPCR	16381831
	7	**GrapeMiRNA**	Winorośl	przewidziane *in silico* miRNA	program FindMiRNA	19563653
	8	**miRNEST**	zwierzęta, rośliny, grzyby (544)	miRNA przewidziane *in silico* i/lub potwierdzone eksperymentalnie, mRNA docelowe, polimorfizm i regulacja ekspresji miRNA	literatura, algorytm do identyfikacji miRNA w sekwencjach EST, 13 baz danych miRNA (patrz: podrozdział *miRNEST*), GEO, NCBI	22135287
Sekwencje docelowe	9	**miRWalk (dawniej: Argonaute)**	człowiek, mysz, szczur	przewidziane oraz potwierdzone mRNA docelowe	bazy danych: GenBank, Ensembl, miRBase, programy: DIANA-microT, miRanda, miRDB, PicTar, PITA, RNA22, TargetScan/ TargetScanS, miRWalk	21605702
	10	**HOCTAR**	Człowiek	mRNA docelowe	programy: miRanda, TargetScan, PicTar.	21435384
	11	**RepTar**	człowiek, mysz	przewidziane *in silico* mRNA docelowe	nowy algorytm oparty na założeniu, że miRNA może posiadać więcej niż jedno miejsce wiązania do pojedynczej sekwencji UTR	21149264
	12	**miRTarBase**	zwierzęta, rośliny, wirusy (14)	mRNA docelowe	literatura	21071411
	13	**miRGator**	człowiek, mysz	mRNA docelowe miRNA i ich ekspresja, powiązania miRNA z chorobami	bazy danych: PhenomiR, GEO, ArrayExpress, programy: targetScan, PITA, miRanda, miRbridge	21062822
	14	**starBase**	człowiek, mysz, C. elegans, rzodkiewnik pospolity, ryż, winorośl	mRNA docelowe	eksperymenty CLIP-Seq i Degradome-Seq	21037263
	15	**miRSel**	człowiek, mysz, szczur	mRNA docelowe	bazy danych: HGNC, MGD, Entrez Gene, Swiss-Prot Protein Database, miRGen, miRBase	20233441
	16	**miRecords**	zwierzęta (9)	mRNA docelowe	literatura, programy: DIANAmicroT, MicroInspector, miRanda, miTarget, MirTarget2, NbmirTar, PicTar, PITA, RNA 22, RNA Hybrid, TargetScan/TargetScanS	18996891
	17	**TarBase**	zwierzęta (6)	mRNA docelowe (tylko eksperymentalne)	literatura	18957447
	18	**miRDB**	człowiek, mysz, szczur, pies, kura	mRNA docelowe oraz adnotacja funkcjonalna miRNA	baza danych miRBase, nowy algorytm do szukania mRNA docelowych	18426918
	19	**MicroRNA.org**	*człowiek, mysz, szczur, muszka owocowa,* C. elegans	mRNA docelowe i ekspresja miRNA	literatura, program miRanda, bazy danych: miRBase, UCSC	18158296
	20	**MiRonTop**	człowiek, mysz, szczur	mRNA docelowe	bazy danych: miRBase, NCBI, programy: Targetscan, MicroCosm Targets, Miranda, PicTar	20959382

Ekspresja miRNA i sekwencji docelowych	21	**CIRCUITSdb**	człowiek, mysz	regulacja ekspresji miRNA przez czynniki transkrypcyjne	literatura, bazy danych: TransmiR, TarBase, Myc Target Gene	20731828
	22	**mESAdb**	człowiek, mysz, danio pręgowany	ekspresja miRNA i ich mRNA docelowych	bazy danych: Ensembl, miRBase, microCosm, HUGE, KEGG, GO	21177657
	23	**miRNeye**	mysz	ekspresja miRNA w oku myszy	Eksperyment: hybrydyzacja RNA *in situ* z wykorzystaniem modyfikowanych nukleotydów LNA	21171988
	24	**dbDEMC**	Człowiek	ekspresja miRNA w tkankach nowotworowych	literatura	21143814
	25	**miReg**	Człowiek	regulacja ekspresji genów miRNA	literatura	20693604
	26	**PuTmiR**	Człowiek	regulacja ekspresji genów miRNA przez czynniki transkrypcyjne	bazy danych: miRBase, UCSC	20398296
	27	**S-MED**	Człowiek	ekspresja miRNA w sarkomie	eksperymenty z wykorzystaniem systemu BeadArrays	20212452
	28	**PhenomiR**	Człowiek	ekspresja miRNA w chorobach i różnych procesach biologicznych	literatura, bazy danych: OMIM Morbid Map, Gene Ontology, BRENDA Tissue Ontology	20089154
	29	**miRGen**	zwierzęta (11)	regulacja ekspresji miRNA, polimorfizm, mRNA docelowe	literatura, program mathTM tool (szukanie TFBS), bazy danych: mammalian miRNA expression atlas, UCSC, dbSNP	19850714
	30	**TransmiR**	Zwierzęta	regulacja ekspresji miRNA przez czynniki transkrypcyjne	literatura, baza danych UCbase & miRfunc	19786497
	31	**miR2Disease**	Człowiek	ekspresja miRNA w chorobach	literatura, baza danych TarBase	18927107
	32	**GenomeTraFaC**	człowiek, mysz	regulacja ekspresji miRNA przez czynniki transkrypcyjne	bazy danych: Homologene, NCBI, MGI, miRBase	17178752
	33	**miSolRNA**	pomidor, rzodkiewnik pospolity	ekspresja miRNA oraz ich funkcje w szlakach metabolicznych	literatura	21059227
	34	**MirZ (dawniej: mammalian miRNA expression atlas)**	człowiek, mysz, szczur	ekspresja miRNA	eksperymenty sekwencjonowania w technologii NGS	17604727
	35	**mirEX**	rzodkiewnik pospolity	ekspresja miRNA	eksperymenty real-time PCR	22013167
	36	**mimiRNA**	Człowiek	ekspresja miRNA	literatura, programy: TargetScan, RNA22, PicTar, algorytm ExParser, bazy danych: Hypertext cell line database, mammalian miRNA expression atlas, GEO	19933167
	37	**mirConnX**	człowiek, mysz	regulacja ekspresji miRNA	bazy danych: TarBase, miRBase, DBTSS, UCSC, The Eukaryotic Promoter Database, programy: CoreBoost_HM, PITA, miRANDA, TargetScan, RNAhybrid, Pictar	21558324
Polimorfizm	38	**miRvar**	Człowiek	polimorfizm i jego funkcjonalne konsekwencje	literatura, bazy danych: SNPdb, UCSC Genome Browser, miRBase, programy: PHDcleav, RISCbinder	21618345
	39	**Patrocles**	zwierzęta (7)	polimorfizm miRNA i mRNA docelowych	literatura, bazy danych: miRBase, Ensembl, program RNAfold	19906729
	40	**PolymiRTS**	człowiek, mysz	polimorfizm w mRNA docelowych	bazy danych: dbSNP, miRBase	17099235
	41	**dPORE-miRNA**	Człowiek	polimorfizm i regulacja ekspresji miRNA	bazy danych: UCSC, PhenomiR, Tarbase, KEGG, program BIOBASE MATCH	21326606
	42	**dbSMR**	Człowiek	polimorfizm miRNA	bazy danych: miRBase, Ensembl, programy: miRanda, RNAHybrid, TargetScan	19371411

Funkcje	43	**UCbase & miRfunc**	człowiek, mysz, szczur	funkcje miRNA; konserwacja sekwencji miRNA	bazy danych: miRBase, UCSC, NCBI	18945703
	44	**miRNApath**	człowiek, mysz, szczur, kura	udział miRNA w ścieżkach metabolicznych	bazy danych: miRBase, miRGen, miRGen, KEGG	18058708
	45	**miRò**	Człowiek	powiązania miRNA-fenotyp	bazy danych: miRBase, mammalian miRNA expression atlas, miRecords, NCBI, GO, Genetic Association Database, programy: TargetScan, PicTar, miRanda	20157481
	46	**miREnvironment**	zwierzęta, rośliny (17)	powiązania miRNA - fenotyp	literatura	21984757
	47	**miTALOS**	człowiek, mysz	udział miRNA w szlakach sygnalizacyjnych	programy: TargetScan, TargetScan, PicTar, Pita, RNA22, bazy danych: KEGG, NCBI	21441347
Inne	48	**IntmiR**	człowiek, mysz	intronowe miRNA, ich mRNA docelowe i deregulacja w chorobach	brak danych	21423893
	49	**CoGemiR**	zwierzęta (36)	genomika i konserwacja sekwencji miRNA	bazy danych: miRBase, Ensembl, SymAtlas, CoGemiR, program miRNAminer	18837977
	50	**AntagomirBase**	Człowiek	antagomiry (cząsteczki służące do wyciszania ekspresji genów miRNA)	programy: Sfold, mfold	21904438
	51	**HNOCDB**	Człowiek	miRNA powiązane z nowotworami głowy i szyi oraz nowotworem szczęki	literatura	22024348

Bazy danych podzielono na pięć kategorii, w zależności od charakteru przechowywanych w nich danych. Dodatkowo wyszczególniono kategorię *Inne* dla baz IntmiR, CoGemiR, AntagomirBase i HNOCDB ze względu na unikalny charakter danych. *PMID – PubMed ID, identyfikator publikacji w serwisie PubMed.

Map Viewer [43] i Ensembl [44]), które pozwalają śledzić otoczenie genowe pre-miRNA, choć prawie zawsze informacja na temat budowy genu miRNA nie jest dostępna. W bazie danych RFAM [45], która gromadzi dopasowania wielu sekwencji różnych klas RNA, znajdują się zwierzęce, roślinne i wirusowe sekwencje pre-miRNA podzielone na 452 rodziny na podstawie podobieństwa sekwencji. deepBase [46] jest kolekcją małych regulatorowych RNA i gromadzi sekwencje miRNA należące do 7 gatunków. W bazie ASRP (ang. *Arabidopsis thaliana Small RNA Project*) można znaleźć krótkie sekwencje RNA z eksperymentów NGS zmapowane do pre-miRNA u *Arabidopsis thaliana* [30], zaś CSRDB (ang. *Cereal Small RNA Database*) kolekcjonuje małe niekodujące RNA, również z eksperymentów NGS, ale zidentyfikowane u ryżu i kukurydzy [47]. Dodatkowo, informacji o miRNA można szukać w bazach ENCODE [48], RNAdb [49] i ncRNAdb [50], kolekcjonujących niekodujące RNA oraz ncRNAimprint [51], bazie zawierającej RNA będące przedmiotem imprintingu.

ZAUTOMATYZOWANE PRZESZUKIWANIE I POBIERANIE DANYCH

Głównym problemem podczas korzystania z wielu baz danych jako źródła informacji o miRNA, jest brak jednolitego interfejsu wyszukiwania i pobierania potrzebnych informacji. Poszczególne serwery bazodanowe przechowują dane w charakterystyczny dla siebie sposób, co prowadzi do dużego zróżnicowania formatów plików i danych. Odpowiedzią na taki stan rzeczy jest miRMaid [52]. Jest to system ułatwiający wyszukiwanie i ściąganie potrzebnych informacji z różnych serwerów bazodanowych, zaprojektowany do współpracy z bazą miRBase, ale w przyszłości planowane jest rozszerzenie jego funkcjonalności na inne bazy danych miRNA. miRMaid pozwala na dostęp do danych poprzez interfejs oparty na języku Ruby oraz poprzez sieć WWW, korzystając z interfejsu REST (ang. *Representational State Transfer*). Po zainstalowaniu na serwerze, miRMaid może automatycznie pobierać dane z obecnej wersji bazy miRBase, a następnie tworzyć lokalna bazę danych, na komputerze użytkownika.

PODSUMOWANIE

Istnieje kilka czynników decydujących o użyteczności bazy danych dla społeczności naukowej. Są to przede wszystkim: jakość danych, ich ilość, oryginalność, jak również jakość interfejsu. Jakość danych mocno zależy od metody, która posłużyła do ich otrzymania. Jednakże dużą niedogodnością jest to, że nierzadko brak wymiernej, liczbowej informacji na temat jakości danych, jak np. wartości prawdopodobieństwa czy P-value, a jeśli jest, to w jednostkach, które nie pozwalają na porównania z podobnymi bazami danych. Poza tym, należy się liczyć z faktem, że bazy danych posiadają pewną ilość przykładów fałszywie pozytywnych oraz innego rodzaju błędów, zwłaszcza jeśli nie są utrzymywane ręcznie i nie są aktualizowane. Jeśli chodzi o rozmiar bazy danych, to istnieje obecnie tendencja do tworzenia dość dużych i wszechstronnych baz danych, jako że te o wąskiej tematyce, skoncentrowane na przykład na jednym gatunku i jednej tkance, jak np. miRNeye [53], są skierowane jedynie do wąskiego grona specjalistów, przez co ich użyteczność jest mocno ograniczona. W przypadku udostępniania przez bazę danych/serwis danych z zewnętrznych źródeł danych, powinny być one jasno wskazane. Ostatnim kryterium mówiącym o użyteczności bazy danych jest jakość interfejsu. Zdarza się, że baza gromadzi niezwykle ciekawe, oryginalne dane, jednakże posiada nie-

intuicyjny interfejs graficzny lub pojawiają się liczne błędy ze strony serwera czy przeglądarki internetowej. Stworzenie bazy danych, która posiadałaby wysokiej jakości, oryginalne dane dostępne poprzez prosty w obsłudze i nowoczesny interfejs graficzny jest trudnym zadaniem, z którym niektórzy twórcy baz danych miRNA sobie nie poradzili.

W chwili obecnej istnieje 51 baz danych miRNA i coraz szybciej powstają nowe. Są to głównie bazy danych sekwencji miRNA, ich mRNA docelowych, funkcji oraz poziomu ekspresji. Choć ciągle istnieją kierunki badań nad miRNA, które nie doczekały się bazy danych, np. budowa genów miRNA, ważna z punktu widzenia badań nad regulacją ekspresji miRNA oraz ich ewolucją, to liczba baz danych oraz ich niekonsekwentna struktura i niepełna dokumentacja sprawiają, że poruszanie się w tej materii wiąże się z coraz większą trudnością. Prawdopodobnym kierunkiem, w jakim może podążać tworzenie nowych baz danych są zintegrowane systemy, kolekcjonujące dane dostępne dotychczas w różnych bazach danych i udostępniające je w zestandaryzowanym formacie poprzez jednolity interfejs graficzny.

PIŚMIENNICTWO

1. Leung AK, Sharp PA (2010) MicroRNA functions in stress responses. Mol Cell 40: 205-215
2. O'Neill LA, Sheedy FJ, McCoy CE (2011) MicroRNAs: the fine-tuners of Toll-like receptor signaling. Nat Rev Immunol 11: 163-175
3. Kedde M, Agami R (2008) Interplay between microRNAs and RNA--binding proteins determines developmental processes. Cell Cycle 7: 899-903
4. Schotte D, Pieters R, Den Boer ML (2012) MicroRNAs in acute leukemia: from biological players to clinical contributors. Leukemia 26: 1-12
5. Brabletz S, Bajdak K, Meidhof S, Burk U, Niedermann G, Firat E, Wellner U, Dimmler A, Faller G, Schubert J, Brabletz T (2011) The ZEB1/miR-200 feedback loop controls Notch signaling in cancer cells. EMBO J 30: 770-782
6. Yao J, Hennessey T, Flynt A, Lai E, Beal MF, Lin MT (2010) MicroRNA-related cofilin abnormality in Alzheimer's disease. PLoS One 5: e15546
7. Filip A (2007) MikroRNA: nowe mechanizmy regulacji ekspresji genów. Postepy Biochem 53: 413-419
8. Koronacki J, Cwik J (2008) Statystyczne systemy uczące się, Exit, Warszawa
9. Higgs PG, Attwood TK (2008) Bioinformatyka i ewolucja molekularna, Wydawnictwo Naukowe PWN, Warszawa
10. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 26: 407-415
11. Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez JM, Aransay AM (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 37: W68-W76
12. Szcześniak MW, Deorowicz S, Gapski J, Kaczyński Ł, Makałowska I (2012) miRNEST database: an integrative approach in microRNA search and annotation. Nucleic Acids Res 40: D198-D204
13. Mhuantong W, Wichadakul D (2009) MicroPC (microPC): A comprehensive resource for predicting and comparing plant microRNAs. BMC Genomics 10: 366
14. Chi X, Yang Q, Chen X, Wang J, Pan L, Chen M, Yang Z, He Y, Liang X, Yu S Identification and Characterization of microRNAs from Peanut (Arachis hypogaea L.) by High-Throughput Sequencing. PLoS One 6: e27530
15. Chen J, Lozach J, Garcia EW, Barnes B, Luo S, Mikoulitch I, Zhou L, Schroth G, Fan JB (2008) Highly sensitive and specific microRNA expression profiling using BeadArray technology. Nucleic Acids Res 36: e87
16. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res 37: W273-W276
17. Krüger J, Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res 34: W451-454
18. Rusinov V, Baev V, Minkov IN, Tabler M (2005) MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. Nucleic Acids Res 33: W696-700
19. Chen K, Rajewsky N (2006) Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet 38: 1452-1456
20. Kuhn DE, Martin MM, Feldman DS, Terry AV Jr, Nuovo GJ, Elton TS (2008) Experimental validation of miRNA targets. Methods 44: 47-54
21. Griffiths-Jones S (2004) The microRNA Registry. Nucleic Acids Res 32: D109-D111
22. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R (2005) NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res 33: D562-566
23. Zhang Z, Yu J, Li D, Zhang Z, Liu F, Zhou X, Wang T, Ling Y, Su Z (2009) PMRD: plant microRNA database. Nucleic Acids Res 38: D806-813
24. Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ (2010) PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. Genome Biol 11: R6
25. Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, Yao L, Zhang Y, Miao R, Cao Y, Zhao Y, Zhong Y, Zhao H (2010) dbDEMC: a database of differentially expressed miRNAs in human cancers. BMC Genomics 11: S5
26. Hiard S, Charlier C, Coppieters W, Georges M, Baurain D (2010) Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. Nucleic Acids Res 38: D640-651
27. Wang X (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. RNA 14: 1012-1017
28. 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 (2010) miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res 39: D163-169
29. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T (2008) miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 37: D105-D110
30. Gustafson AM, Allen E, Givan S, Smith D, Carrington JC, Kasschau KD (2005) ASRP: the Arabidopsis Small RNA Project Database. Nucleic Acids Res 33: D637-640
31. Schmeier S, Schaefer U, MacPherson CR, Bajic VB (2011) dPORE-miRNA: polymorphic regulation of microRNA genes. PLoS One 6: e16657
32. Maselli V, Di Bernardo D, Banfi S (2008) CoGemiR: a comparative genomics microRNA database. BMC Genomics 9: 457
33. Zhang Y, Guan DG, Yang JH, Shao P, Zhou H, Qu LH (2010) ncRNAimprint: a comprehensive database of mammalian imprinted noncoding RNAs. RNA 16: 1889-1901
34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-402
35. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948
36. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40: D130-D135
37. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 37: D98-D104

38. Lindberg C (1990) The Unified Medical Language System (UMLS) of the National Library of Medicine. J Am Med Rec Assoc 61: 40-42
39. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37: D793-D796
40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29
41. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res 37: D588-D592
42. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, James Kent W (2012) The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res 40: D918-923
43. Wolfsberg TG (2007) Using the NCBI Map Viewer to browse genomic sequence data. Curr Protoc Bioinformatics 1: 1.5
44. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M (2002) The Ensembl genome database project. Nucleic Acids Res 30: 38-41
45. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. Nucleic Acids Res 31: 439-441
46. Yang JH, Shao P, Zhou H, Chen YQ, Qu LH (2010) deepBase: a database for deeply annotating and mining deep sequencing data. Nucleic Acids Res 38: D123-D130
47. Johnson C, Bowman L, Adai AT, Vance V, Sundaresan V (2007) CSRDB: a small RNA integrated database and browser resource for cereals. Nucleic Acids Res 35: D829-D833
48. Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, Karolchik D, Barber GP, Clawson H, Diekhans M, Fujita PA, Goldman M, Gravell RC, Harte RA, Hinrichs AS, Kirkup VM, Kuhn RM, Learned K, Maddren M, Meyer LR, Pohl A, Rhead B, Wong MC, Zweig AS, Haussler D, Kent WJ (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res 40: D912-D917
49. Pang KC, Stephen S, Engström PG, Tajul-Arifin K, Chen W, Wahlestedt C, Lenhard B, Hayashizaki Y, Mattick JS (2005) RNAdb - a comprehensive mammalian noncoding RNA database. Nucleic Acids Res 33: D125-D130
50. Szymanski M, Erdmann VA, Barciszewski J (2007) Noncoding RNAs database (ncRNAdb). Nucleic Acids Res 35: D162-D164
51. Zhang Y, Guan DG, Yang JH, Shao P, Zhou H, Qu LH (2010) ncRNAimprint: a comprehensive database of mammalian imprinted noncoding RNAs. RNA 16: 1889-1901
52. Jacobsen A, Krogh A, Kauppinen S, Lindow M (2010) miRMaid: a unified programming interface for microRNA data resources. BMC Bioinformatics 11: 29
53. Karali M, Peluso I, Gennarino VA, Bilio M, Verde R, Lago G, Dollé P, Banfi S (2010) miRNeye: a microRNA expression atlas of the mouse eye. BMC Genomics 11: 715

microRNA databases

Michał Szcześniak, Elżbieta Owczarkowska, Jakub Gapski, Izabela Makałowska✉

Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznan, Poznan

Key words: database, microRNA, miRBase
✉**e-mail**: izabel@amu.edu.pl

SUMMARY

microRNAs (miRNAs) are small RNAs that play key roles in regulation of cellular processes and therefore could largely contribute to solving many problems in medicine, biotechnology, and other biological sciences. As a result, the numbers of research projects and publications on miRNAs are constantly growing, which is accompanied by mounting up amounts of new data and databases need to be created for data storage. There are 51 dedicated miRNA databases at the moment, which makes it quite difficult for the users to find relevant data. Moreover, such problems as insufficient documentation, low quality of data or flaws in the graphical interface make the things even worse. However, there are positive signs, including standardization of database interfaces, a tendency to create integrated systems that collect data from a number of databases and present it in a uniform format, and emergence of systems for automated data search and download.

VI. ZAŁĄCZNIKI

Primate and Rodent Specific Intron Gains and the Origin of Retrogenes with Splice Variants

Michał W Szcześniak,†[,1] Joanna Ciomborowska,†[,1] Witold Nowak,[2] Igor B Rogozin,[3] and Izabela Makałowska*[,1]

[1]Laboratory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

[2]Laboratory of Molecular Techniques, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

[3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

†These authors contributed equally

***Corresponding author**: E-mail: izabel@amu.edu.pl.

Associate editor: Helen Piontkivska

Abstract

Retroposition, a leading mechanism for gene duplication, is an important process shaping the evolution of genomes. Retrogenes are also involved in the gene structure evolution as a major player in the process of intron deletion. Here, we demonstrate the role of retrogenes in intron gain in mammals. We identified one case of "intronization," the transformation of exonic sequences into an intron, in the primate specific retrogene *RNF113B* and two independent "intronization" events in the retrogene *DCAF12L2*, one in the common ancestor of primates and rodents and another one in the rodent lineage. Intron gain resulted from the origin of new splice variants, and both genes have two transcript forms, one with retained intron and one with the intron spliced out. Evolution of these genes, especially *RNF113B*, has been very dynamic and has been accompanied by several additional events including parental gene loss, secondary retroposition, and exaptation of transposable elements.

Key words: intron gain, gene structure evolution, splice variant, RNF113, DCAF12.

The majority of protein-coding genes in eukaryotes are interrupted by introns that are removed from the pre-mRNA by a RNA–protein complex called the spliceosome (Cavalier-Smith 1985; Crick 1979). Introns and the splicing machinery have been found in all eukaryotic species with fully sequenced genomes (Chow et al. 1977; Roy and Gilbert 2006). Comparative genomic studies have revealed striking conservation of intron positions in distant eukaryotes such as animals and plants (Fedorov et al. 2002; Rogozin et al. 2003; Carmel et al. 2007). On the other hand, many genome-wide comparisons of eukaryotic species demonstrated multiple intron losses and intron gains (Roy et al. 2003; Cho et al. 2004; Qiu et al. 2004; Coulombe-Huntington and Majewski 2007b; Li et al. 2009). However, it was found that intron gain is a very rare event in vertebrate evolution (Loh et al. 2007) and no intron gains into intact conserved coding regions of mammalian genes are known (Roy et al. 2003; Coulombe-Huntington and Majewski 2007a).

Comparative gene structure studies have not revealed any intron gain into existing exons in mammals. The only reported new introns were acquired, by and large, by either a fusion of retrogene with host genes or de novo from the genomic environment as a result of new exon capture (O'Neill et al. 1998; Vinckenbosch et al. 2006; Sela et al. 2007; Baertsch et al. 2008; Fablet et al. 2009). Here, we report two retrogenes, *RNF113B* and *DCAF12*, where the exon sequence was split by creation of a new intron as the result of mutations and emergence of new splice sites. The introns discovered by us represent cases of intron creation via recruitment of exonic sequence (intronization) proposed by Irimia et al. (2008) and Lahn and Page (1999).

Evolution of Introns

RNF113A is a retrogene encoding a ring finger protein of unknown function and is present in the genomes of all vertebrates. Interestingly, in mammalian genomes, only intronless copy exist, whereas in all other vertebrates, a ten-exon parental gene is present and no retrogenes were detected. Genomic sequence analysis showed that there are two copies of *RNF113* in primates, rodents, carnivores, and even-toed ungulates and only one in the genomes of the other mammals we studied. The first copy of *RNF113* was retroposed into the intronic region of *NDUFA1* gene in the genome of the mammalian ancestor. Following the retroposition, the parental gene was lost. This likely took place before the divergence of Prototheria (Monotremes) and Theria (Marsupials and Placentals) because in the genomes of all species representing these lineages, the multiexon form of *RNF113* is absent. After the mammalian radiation the *RNF113A* retrogene was duplicated, by retropositions or segmental duplications, in several lineages. Analysis of genomic locations of these copies suggests that the duplication events were independent in each lineage. For example, in rodents, the *RNF113* copy (*RNF113A2*) was inserted into an intron of

Letter

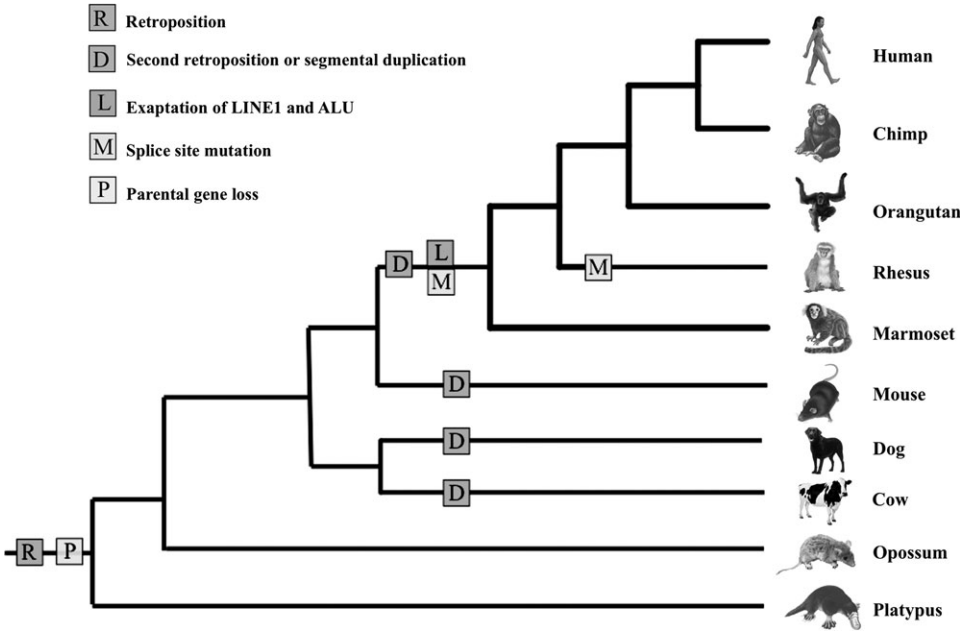


Fig. 1. Schematic tree representing major events during evolution of *RNF113* gene in mammalian lineage. Color version of the figure can be found in supplementary Data (Supplementary Material online).

the *2900006K08Rik* gene, whereas the primate specific gene, *RNF113B*, was copied into an intron of the *FARP1* gene. The primate specific duplication happened before Old World Monkeys and New World Monkeys diverged (fig. 1).

After the retroposition/duplication, the primate specific *RNF113B* gene underwent rapid evolution including intron gain. The presence of the intron is surprising, however, it is supported by several GenBank mRNA sequences (accession numbers: AF539427, BC025388, and BC017585). To confirm the existence of the intron and learn about its origin, we compared *RNF113B* sequences from available primate genomes (human, marmoset, macaque, orangutan, and chimpanzee) with sequences of other mammalian *RNF113A* genes. Sequence alignment revealed that the intron of *RNF113B* is not a de novo insertion but rather originated from the exonic sequence (fig. 2*a*). A double point mutation, AG → GT, generated the donor site (fig. 2*a*). The origin of acceptor site is not so clear. One possibility is that a point mutation, GG → AG, created acceptor site. Another option is that the acceptor site was brought during the exonization of L1 element, merged at the 3′ end of *RNF113B* (fig. 2*b*). The newly generated splice sites together with the branch site and the polypyrimidine tract likely enabled recognition of the new intron by the U2 spliceosome (fig. 2*a*). The 105 bp intron contains 59 nucleotides of previously coding sequence and 46 nucleotides from the 3′ UTR.

Generation of splice sites most probably occurred in the primate specific *RNF113B* copy since neither human *RNF113A* gene, which gave a rise to primate *RNF113B*, nor *RNF113A* genes from other mammals have AG or GT at the donor and acceptor positions. Splicing signals were formed before the Old World Monkeys and New Monkeys split. Interestingly, loss of the splicing boundaries subsequently converted the intron into a "retained intron" in some primates. In rhesus, for example, acceptor was lost due to a point mutation (AG → AA change) (fig. 1*b*).

The creation of splicing signals was accompanied not only by exaptation of an L1 element but also by exonization of an Alu element. The L1 element inserted within the 3′ end of the gene could have contributed the acceptor site and provided a new polyA signal used for the new splice variant (fig. 2*a*). The complete AluSx element transposed upstream the gene was exapted at the 5′ end and most probably delivered some regulatory elements.

Sequencing of the human *RNF113B* cDNA using primers flanking the intronic sequence revealed that *RNF113B* produces two variant transcripts. One variant has two exons, as described above, and the other one is a single exon transcript similar to *RNF113A*. Consequently, most primates have three transcripts of *RNF113*: one from the *RNF113A* retrogene and two from the *RNF113B* (fig. 2*b*). Rodents, cow, and dog have two transcripts, each coming from different copy of *RNF113*, and all other mammals have only one *RNF113* transcript. The presence of the splice variants in the retrogene is very surprising and has only been reported once before (Lahn and Page 1999).

A second case involves *DCAF12* (DDB1 and CUL4 associated factor 12), which encodes a WD repeat-containing protein that interacts with the COP9 signalosome (Jin et al. 2006). Although the gene is present in vertebrate and insect genomes, only placental mammals have retrocopies of this gene. One copy, *DCAF12L2*, has the same location in all placental mammals and therefore most likely was retroposed in the placental mammals ancestor. Another copy, *DCAF12L1*, is present only in Euarchontoglires (a clade which includes rodents and primates). It likely emerged as a result of tandem duplication of *DCAF12L2* as it is located next to the *DCAG12L1* gene. There were two events that changed the

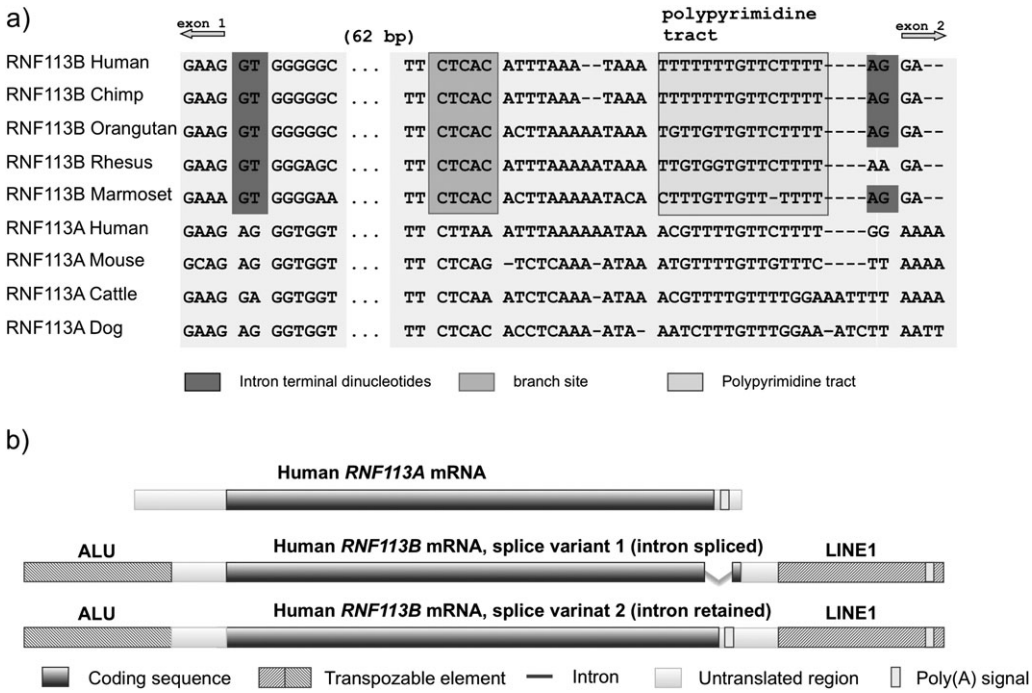


Fig. 2. (*a*) Alignment of mammalian *RNF113A* and primate *RNF113B* genomic sequences at the acceptor and donor sites. (*b*) Structure of human *RNF113A* mRNA and two splice variants of *RNF113B*. Color version of the figure can be found in supplementary Data (Supplementary Material online).

splicing pattern in *DCAF12L2*. First, an intronization event occurred in the common ancestor of primates and rodents. Second, an alternative donor site emerged in rodents only (fig. 3). The limited available data and sequence divergence make any conclusions in regard to the exact pattern of splice site evolution infeasible. However, there is convincing experimental evidence confirming both splicing events (fig. 3): splicing at the shared rodent–primate intron, boundaries are confirmed by two expressed sequence tags (ESTs) (AK034343 and AK047360), and usage of the rodent

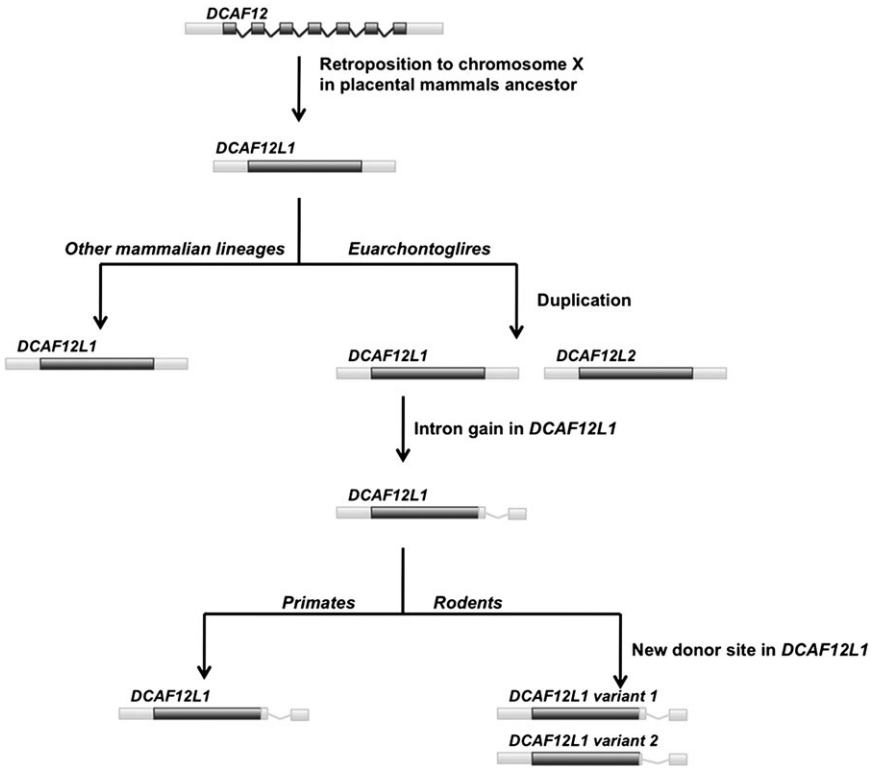


Fig. 3. Pattern of *DCAF12* duplication and "intronization" events in mammalian genomes. Color version of the figure can be found in supplementary Data (Supplementary Material online).

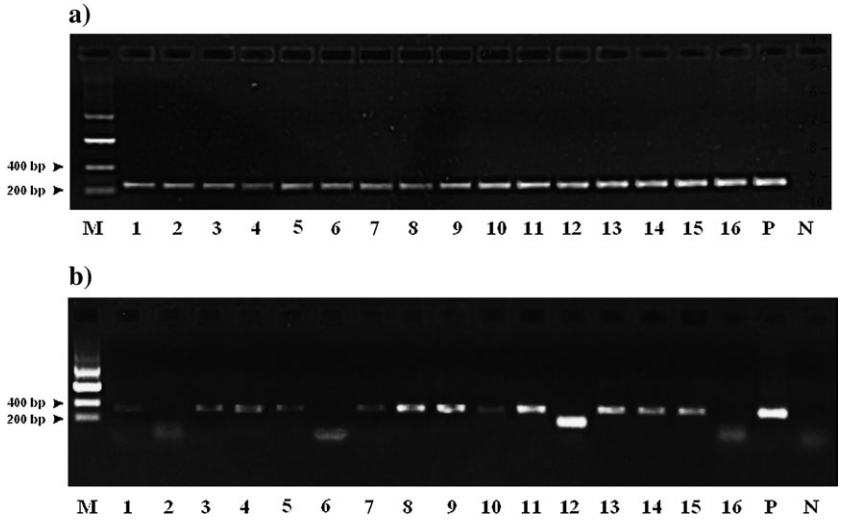


Fig. 4. Expression pattern of *RNF113A* and two forms of *RNF113B* (195 bp product with intron spliced; 295 bp product-form with intron retained) in 16 human tissues: 1: heart, 2: brain, 3: placenta, 4: lung, 5: liver, 6: skeletal muscle, 7: kidney, 8: pancreas, 9: spleen, 10: thymus, 11: prostate, 12: testis, 13: ovary, 14: small intestine w/o mucosal lining, 15: colon, 16: peripheral leukocytes, P: genomic DNA, and N: water.

alternative donor site is confirmed by four ESTs (AK038557, BC068319, AK034472, and AK039767).

Retrogene Expression

Numerous studies revealed a tendency of retrogenes to be expressed exclusively in testis. It was suggested that the hypertranscription present in the meiotic and postmeiotic spermatogenic cells makes possible transcription of DNA that is usually not transcribed. This may facilitate transcription of retrocopies in the testis during their early evolution (reviewed in (Kaessmann et al. 2009). Another hypothesis explains the high expression of retrogenes in testis by the fact that these are, in most cases, retrocopies of spermatogenesis-related genes located on the X chromosome. Because the X chromosome is inactivated during meiosis, retroposition to autosomes enables escape from inactivation and expression during spermatogenesis (Turner 2007).

The retroposition of both genes studied here, *RNF113* and *DCAF12*, was in the opposite direction, from autosomes to chromosome X. In the case of *RNF113*, the parental gene is detectable by sequence similarity as an apparent pseudogene on chromosome 9. The parental multiexon *DCAF12* gene is coincidentally also located on chromosome 9. *RNF113A* and both *DCAF12* retrogenes are on chromosome X. We surveyed the expression pattern of all human *RNF113* transcripts (one from *RNF113A* and two from *RNF113B*) in 16 human tissues (fig. 4) (for methods, see Supplementary Material online). *RNF113A* was expressed in all studied tissues, including testes. Interestingly, *RNF113B* exhibited tissue-specific splicing; while the unspliced form of *RNF113B* was expressed in all tissues but testis, the spliced variant was expressed in testis, prostate, thymus, and lung. Both *RNF113B* splice variants were present in thymus, prostate, and lung, but in all of these tissues, the form with the intron spliced out had much lower expression level than the single exon primary form. Relatively high expression of the new form of *RNF113B*, form with the intron spliced out, was observed only in testis.

According to the EST data, the human *DCAF12* gene is widely expressed. EST sequences present in the dbEST database represent almost 40 libraries and show the highest expression in testis and trachea. The retrogene *DCAF12L1* is expressed only in kidney and testis and a second human retrogene, *DCAF12L2*, is expressed in eye and testis. Therefore, both retrogenes show very different expression patterns than their parental genes, with very limited and low expression level and notable expression in testis.

Conclusions

Retroposition, a major mechanism for gene duplication, is an important process shaping the evolution of genomes (Brosius 1991; Marques et al. 2005). Our study confirms the unusual role of retrogenes in shaping the genomes and underscores the importance of mobile elements in evolution. It also reveals that retrogenes may be responsible for a wealth of species-specific features including species-specfic introns and splice variants.

Previous analyses of introns in the vertebrate genomes did not uncover any intron gain in mammals (Roy et al. 2003). Our study clearly shows that creation of introns has occurred during mammalian evolution. The failure of previous studies to find intron gains can be explained by the fact that they were focused on different intron gain mechanisms and did not consider exon intronization. In addition, they looked at conserved among studied species genes, while we focused on young and in many cases lineage-specific retrogenes.

Interestingly, the retrogenes studied here exhibit testis-specific expression typically associated with genes escaping from the X chromosome despite their opposite history (retroposition from autosome to X). This biased expression pattern may not be exclusively related to meiotic genes, sex chromosome inactivation, and dosage compensation

(Marques et al. 2005; Vinckenbosch et al. 2006; Potrzebowski et al. 2008). The same pattern of high expression level in testis is observed in young, primate-specific splice variant of retrogene *RNF113B* as well as in both retroposed copies of *DCAF12* retroposed on the human X chromosome. The older, unspliced variant of *RNF113B*, as well as an earlier retrocopy *RNF113A*, displays more diverse expression patterns. Therefore, testis-specific expression could be a common feature of all newly evolved transcripts regardless of their chromosomal localization and may reflect a transcriptional noise due to "hypertranscription" in testis, facilitating the activation of new transcripts (Kleene et al. 1998).

The small number of observed intron gain in retrogenes may reflect that this is a rare event. Alternatively, the low number of observations could reflect the difficulties in identification of such events. One major complication lies in annotation problems and the common expectation that retrogenes do not have introns. Genome-wide comparative studies currently underway have already showed that intron gain in retrogenes could be more frequent than we expected but that annotations remain a major obstacle in uncovering this phenomenon.

Supplementary Material

Supplementary Data are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Jurgen Brosius and two reviewers for their comments and insightful suggestions. I.B.R. was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/U.S. Department of Health and Human Services.

References

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. 2008. Retrocopy contributions to the evolution of the human genome. *BMC Genomics*. 9:466.

Brosius J. 1991. Retroposons–seeds of evolution. *Science* 251:753.

Carmel L, Wolf YI, Rogozin IB, Koonin EV. 2007. Three distinct models of intron dynamics in the evolution of eukaryotes. *Genome Res*. 17:1034–1044.

Cavalier-Smith T. 1985. Selfish DNA and the origin of introns. *Nature* 315:283–284.

Cho S, Jin SW, Cohen A, Ellis RE. 2004. A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. *Genome Res*. 14:1207–1220.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 12:1–8.

Coulombe-Huntington J, Majewski J. 2007b. Characterization of intron loss events in mammals. *Genome Res*. 17:23–32.

Coulombe-Huntington J, Majewski J. 2007a. Intron loss and gain in *Drosophila*. *Mol Biol Evol*. 24:2842–2850.

Crick F. 1979. Split genes and RNA splicing. *Science* 204:264–271.

Fablet M, Bueno M, Potrzebowski L, Kaessmann H. 2009. Evolutionary origin and functions of retrogene introns. *Mol Biol Evol*. 26:2147–2156.

Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc Natl Acad Sci U S A*. 99:16128–16133.

Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW. 2008. Origin of introns by 'intronization' of exonic sequences. *Trends Genet*. 24:378–381.

Jin J, Arias EE, Chen J, Harper JW, Walter JC. 2006. A family of diverse Cul4-Ddb1-interacting proteins includes Cdt2, which is required for S phase destruction of the replication factor Cdt1. *Mol Cell*. 23:709–721.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet*. 10:19–31.

Kleene KC, Mulligan E, Steiger D, Donohue K, Mastrangelo MA. 1998. The mouse gene encoding the testis-specific isoform of Poly(A) binding protein (Pabp2) is an expressed retroposon: intimations that gene expression in spermatogenic cells facilitates the creation of new genes. *J Mol Evol*. 47:275–281.

Lahn BT, Page DC. 1999. Retroposition of autosomal mRNA yielded testis specific gene family on human Y chromosome. *Nat Genet*. 21:429–433.

Li W, Tucker AE, Sung W, Thomas WK, Lynch M. 2009. Extensive, recent intron gains in Daphnia populations. *Science* 326:1260–1262.

Loh Y-H, Brenner S, Venkatesh B. 2007. Investigation of loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon). *Mol Biol Evol*. 25:526–535.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol*. 3:e357.

O'Neill RJ, Brennan FE, Delbridge ML, Crozier RH, Graves JA. 1998. De novo insertion of an intron into the mammalian sex determining gene, SRY. *Proc Natl Acad Sci U S A*. 95: 1653–1657.

Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jegou B, Kaessmann H. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol*. 6:e80.

Qiu WG, Schisler N, Stoltzfus A. 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol*. 21:1252–1263.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*. 13:1512–1517.

Roy SW, Fedorov A, Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci U S A*. 100:7158–7162.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet*. 7:211–221.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. 2007. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol*. 8:R127.

Turner JM. 2007. Meiotic sex chromosome inactivation. *Development* 134:1823–1831.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A*. 103:3220–3225.

Published online 19 October 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D191–D197***
doi:10.1093/nar/gkr878

mirEX: a platform for comparative exploration of plant pri-miRNA expression data

Dawid Bielewicz[1], Jakub Dolata[1], Andrzej Zielezinski[2], Sylwia Alaba[2], Bogna Szarzynska[1], Michal W. Szczesniak[2], Artur Jarmolowski[1], Zofia Szweykowska-Kulinska[1,2,*] and Wojciech M. Karlowski[2,*]

[1]Department of Gene Expression and [2]Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Umultowska 89, 61-614 Poznan, Poland

Received August 1, 2011; Revised and Accepted September 29, 2011

ABSTRACT

mirEX is a comprehensive platform for comparative analysis of primary microRNA expression data. RT–qPCR-based gene expression profiles are stored in a universal and expandable database scheme and wrapped by an intuitive user-friendly interface. A new way of accessing gene expression data in mirEX includes a simple mouse operated querying system and dynamic graphs for data mining analyses. In contrast to other publicly available databases, the mirEX interface allows a simultaneous comparison of expression levels between various microRNA genes in diverse organs and developmental stages. Currently, mirEX integrates information about the expression profile of 190 *Arabidopsis thaliana* pri-miRNAs in seven different developmental stages: seeds, seedlings and various organs of mature plants. Additionally, by providing RNA structural models, publicly available deep sequencing results, experimental procedure details and careful selection of auxiliary data in the form of web links, mirEX can function as a one-stop solution for *Arabidopsis* microRNA information. A web-based mirEX interface can be accessed at http://bioinfo.amu.edu.pl/mirex.

INTRODUCTION

MicroRNAs are the key post-transcriptional regulators of gene expression in Eukaryota. They control gene expression by targeting the cleavage of cognate mRNAs or by inhibiting their translation (1–3). Therefore, when studying the biology of any organism, it is of utmost importance to quantify precisely the expression level of each particular microRNA gene during organ development. Northern hybridization, microarray analysis, deep sequencing approaches and real-time quantitative PCR (RT–qPCR) are standard techniques used to accomplish this task as described previously (4–7). The RT–qPCR is considered a gold standard method in precise quantification of gene transcript levels (8).

MicroRNA genes that encode transcripts which are processed to the same or similar mature microRNA species are grouped in families. In *Arabidopsis thaliana*, the number of such family members varies between 1 (e.g. miR163, miR173) and 14 (miR169) [miRBase release 17 (9)]. MicroRNA genes from the same family, although represented by identical or almost identical mature microRNAs, differ considerably in gene organization and sequence. In many cases it is only possible to observe the expression of all family members as a group, rather than the individual members due to the large sequence conservation when using northern hybridization or deep sequencing approaches. However, individual members of a given microRNA family may be expressed in different developmental stages or in response to various biotic/abiotic stimuli (10–12).

Our resource contains information regarding the expression of 190 *A. thaliana* microRNA genes at the level of primary microRNA (pri-miRNA) in different developmental stages obtained using a RT–qPCR high-throughput platform. Many databases that store and allow access to microRNA gene expression profiles from variety of organisms already exist [e.g. (13–18)], however,

*To whom correspondence should be addressed. Tel: +48 61 829 5841; Fax: +48 61 829 5949; Email: wmk@amu.edu.pl
Correspondence may also be addressed to Zofia Szweykowska-Kulinska. Tel: +48 61 8295766; Fax: +48 61 829 5949; Email: zofszwey@amu.edu.pl
Present address:
Bogna Szarzynska, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, 46 rue d'Ulm, 75230 ParisCedex 05, France.

The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

they have various limitations and restrictions. One of the limitations includes the lack of tools and/or data for comparisons between developmental stages and various genes at the same time. With this in mind, we developed the mirEX platform as a comprehensive starting point for a comparative investigation of microRNA genes expression. Our database offers to the scientific community easily accessible data and is of interest to researchers working on the specific microRNA function, the expression profile of entire microRNA family members during a particular organ/developmental stage, or on microRNA biogenesis. Additionally, by creating the mirEX interface, we hope to propose a new database interface standard for comparative microRNA gene expression data mining.

WEB TOOLS AND DATA

Expression datasets

In the current release, the mirEX database includes expression data of microRNA genes from *A. thaliana*. To collect the data, we prepared a real-time PCR platform for 190 primary microRNA sequences. For each microRNA, a pair of specific primers amplifying a single product was designed; 10 microRNAs primer sequences were taken from Pant *et al.* (19). The qPCR reactions for one biological replicate included 12 different controls and standards, were carried out in parallel as described previously (20) using RNA template isolated from six developmental stages of *A. thaliana* Columbia-0 wild-type plants. RNA from seeds and siliques was isolated according to Ref. (21). The detailed procedure of running the platform can be found in the mirEX database documentation.

The plant material used in expression profiling experiments includes: seeds, 10-day-old seedlings, 14-day-old seedlings, 25-day-old plants, 35-day-old plants, 42-day-old plants (rosette leaves and stems) and 53-day-old plants (rosette leaves, stems, inflorescence and siliques). For better visualization of the individual developmental stages, we included pictures of actual plants used in our experiments. Our high-throughput platform contains an original and novel set of data on microRNA gene expression in dormant seeds. The PP2A (At1g13320) and actin (At3g18780) cDNAs were used as an expression reference (22). The mean measurements of three replicas were used to calculate the fold-change value and presented in a form of $\log_{10}$. In a case when the value of correlation coefficient of the three replicas was <0.995, such data was labeled as 'low quality' and is not shown by default on the graphs, but indicated in gray in the data tables. Since the level of expression of most of the primary microRNAs is lower than the reference genes, the data shown on the graphs in mirEX is rescaled in order to avoid presenting it as a negative value. Rescaling shifts the zero value of the graph's y-axis to the basal expression level of the whole experiment. This allows showing the expression profile in a positive data range, but does not change the actual values.

Web interface

The mirEX database interface is designed to be used by a bench scientist on an everyday basis. Following a simplicity rule, the interface of mirEX has been built on only two types of result windows and a simple two-step querying system.

There are two ways to access data in mirEX: by searching for a particular microRNA or microRNA family, or by browsing the database content. A single input window allows searching for an individual or a family of microRNA by providing the numerical part of its identification (ID). When entered ID corresponds to a record of a single microRNA, the resulting page shows all of the details for this particular sequence (Figure 1). In the case of whole family ID, the search results are presented in the form of a line graph, showing expression data for all stages available in the database (Figure 2). The table located below the graph contains numerical values for data presented in the graph, including low quality measurements. The IDs shown in the table allow quick access to detailed information about particular microRNA.

Browsing the data in mirEX starts at the two-step selection process. The first step includes selection of developmental stages, which is followed by the selection of microRNAs. In mirEX, there are three ways to select microRNAs: (i) by typing their IDs in the input window, (ii) by uploading a file with a coma-separated list of microRNA IDs (that can be created in the first step) and (iii) by selecting individual or groups of sequences from the tree-like expandable menu. During the process of entering information using a keyboard, the mirEX interface will provide the list of available microRNAs. It is possible to mix all of the input methods—the resulting page will refer to all entered or selected IDs.

Presentation of search/browse results may differ depending on the number of selected developmental stages or microRNA genes (see 'Data mining' section). By default, the graph of expression profiles for various stages and microRNAs shows only high quality data. However, it is a user-defined option in mirEX whether the low quality measurements are presented on the graph. The expression values can be displayed using two reference genes: actin and PP2A.

The results for a single microRNA are presented in a record window in a form that is divided into distinct sections (Figure 1). The upper part of the record is dedicated to expression information presented in the form of a graph and two tables. When applicable, the graph contains information about the expression of other members of the microRNA family, that can be dynamically turned on and off. Additionally, specific shortcuts allow quick access to the records of any member of such family. The two tables, containing either reference-calculated or raw data, can be hidden to reduce the amount of information presented at any given moment on the single record window. Each graph or table in the mirEX database can be saved for later use by selecting an appropriate button.

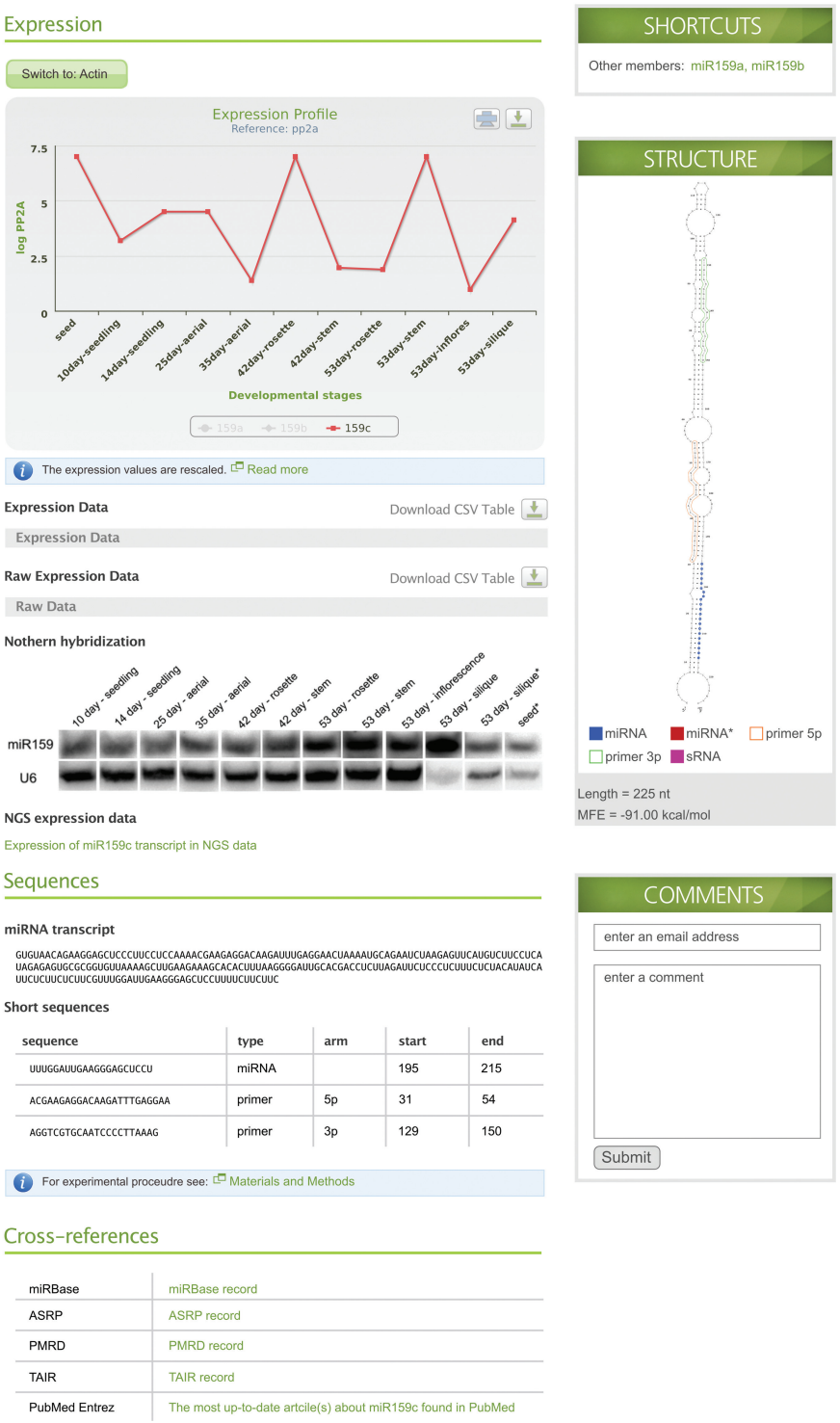
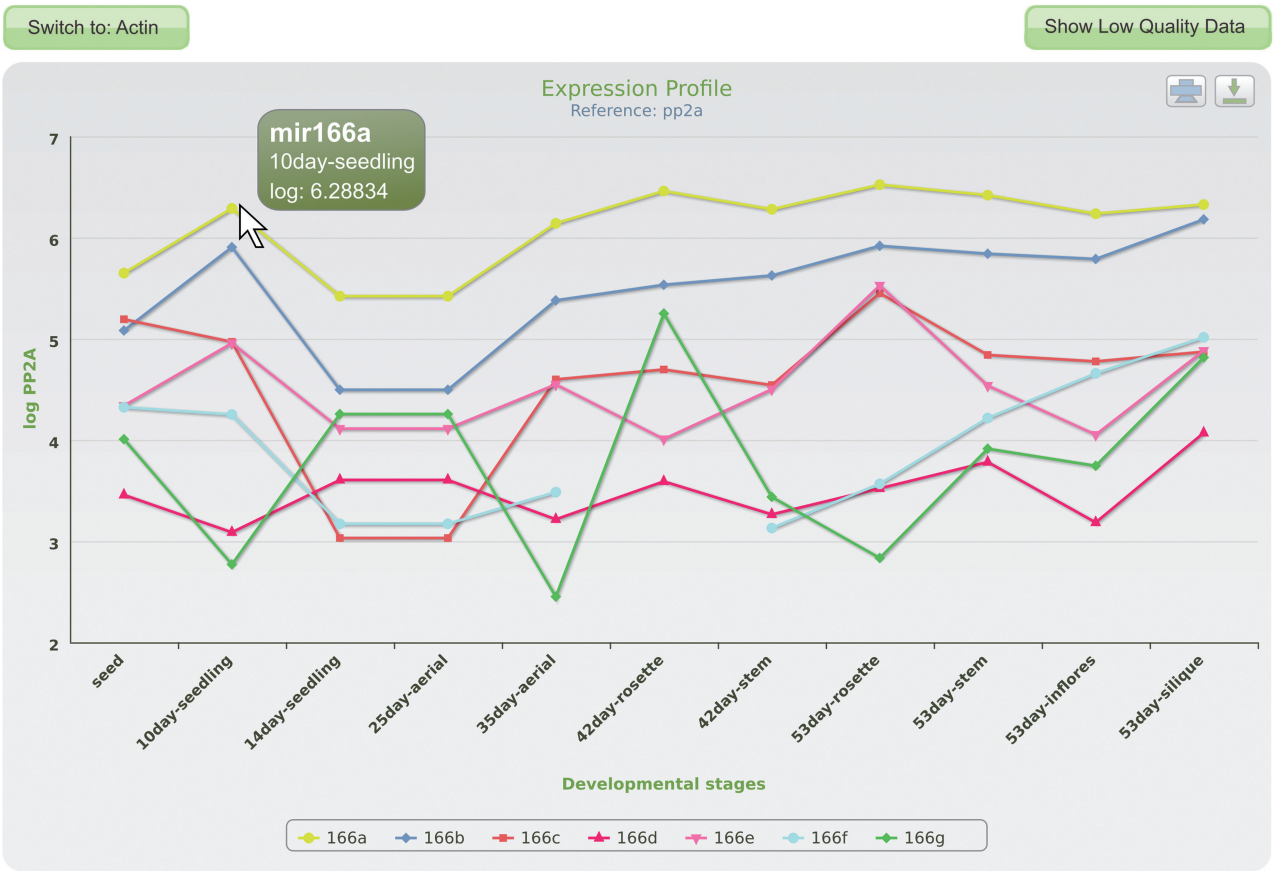


Figure 1. Example of mirEX record window for ath-miR159c. This window is divided into distinct sections representing: expression data, sequences and external references. The right panel contains shortcuts to microRNA family members' records, structure of the RNA transcript molecule with labeled mature sequences and RT–qPCR primers and webform for user comments.

Additionally, below the quantitative expression data, each record contains a web link to a text-based map of publicly available high-throughput Next-Generation Sequencing (NGS) short reads aligned to primary microRNA transcript. The information included here contains sequence, location, length and number of the mapping reads. On a separate window, it is also possible to access the original record of the used NGS data.

The record window contains a graphical presentation of the structure of microRNA transcript with designations of

Expression



The expression values are rescaled. Read more

Table report

miRNA	seed	10day seedling	14day seedling	25day aerial	35day aerial	42day rosette	42day stem	53day rosette	53day stem	53day inflores	53day silique
166a	5.64921	6.28834	5.42083	5.42083	6.14107	6.45961	6.27966	6.52257	6.41905	6.23591	6.32645
166b	5.08157	5.90528	4.49621	4.49621	5.37973	5.53315	5.62505	5.91811	5.84033	5.78834	6.17987
166c	5.19281	4.96945	3.03096	3.03096	4.59848	4.69646	4.54260	5.45045	4.83994	4.77668	4.87147
166d	3.45695	3.08791	3.60479	3.60479	3.21583	3.59028	3.26447	3.52176	3.78184	3.18464	4.06962
166e	4.33894	4.95676	4.11181	4.11181	4.54865	4.00851	4.50100	5.52666	4.53720	4.05214	4.88188
166f	4.32301	4.25127	3.17132	3.17132	3.48388	4.57686	3.12957	3.56610	4.21639	4.65809	5.01434
166g	4.00704	2.77047	4.25605	4.25605	2.45153	5.24896	3.43708	2.83375	3.91363	3.74440	4.81495

Low quality data are indicated in gray. Read more

Figure 2. Example of mirEX report window for ath-miR166 gene family. The report window contains graphical presentation of the expression levels, as well as actual numerical values presented in tabular format. Holding a mouse pointer over any data point allows access to details of expression measurements. The datasets for a particular microRNA can be dynamically turned on and off by clicking ID in the legend of the graph.

mature molecule(s) and primers used to assess its expression level by RT–qPCR. When available, the record data will be modified to include results from northern hybridization with mature microRNAs. However, it has to be noted that in most cases, such hybridization represents the level of the whole microRNA gene family expression, and will be not specific for an individual microRNA gene. Moreover, the record window contains a table with sequences relevant to particular microRNA and a link to experimental procedures.

The last part of the record window includes references to external databases. Our aim was to avoid replication of the data available in other databases. For information such as gene structure or available papers, we direct the user to specialized databases, for example TAIR (23) or PubMed (24). The number of external references will grow in time, when new resources for microRNA biology are available, and in response to user requests.

To maintain a close relation with users of our database, we provide in the record window a simple tool for entering comments. The comment may be of any nature: concerning the biological aspect of the presented data or any general issue.

Although the interface is very intuitive and simple, at the main page we provide a video tutorial on all of the capabilities and various ways available to explore the mirEX tools and data. Moreover, every window in the mirEX interface contains context-specific help guides.

A complete 'road map' of the mirEX interface is included as Supplementary Figure S1.

Data mining

The usefulness of the mirEX database lies in the ability to compare expression between different developmental stages and various microRNA genes. By using the 'browse' button, the user can, in two simple steps, select available developmental stages and a single or sets of microRNAs. Depending on the number of selected stages, the expression results are presented in the form of a bar graph (for one-stage analysis) or a line-based graph (for many stages). The bar graph presentation allows analysis of all pri-miRNAs at once with a zoom-in on an individual gene.

The line graph displays by default, data for only 24 microRNAs selected from the list. However, this restriction can be turned off. The line graph offers an additional option to dynamically remove and/or add any of the selected microRNA genes.

Another, quick way to access comparative expression profiles are shortcuts. In general, the shortcuts represent a new and easy way to explore data stored in the mirEX database. They are located in the same area on every page, and their content follows user selections. The use of shortcuts allows quick access to specific sets of data, e.g. the most expressed genes, or records representing microRNAs belonging to the same family. Periodically, we will modify the shortcuts in response to comments and readapt them according to the most frequently issued queries.

The strategy to use primary microRNA expression data allows fine exploration of gene activities of members of microRNA families. In some families, the expression pattern of individual genes in all developmental stages and organs is similar. However, the expression level of the individual pri-miRNAs may differ considerably: the fold change in the primary transcript levels within one microRNA gene family may range between 10 and 10^5 (e.g. miR160, miR162, miR164, miR165, miR167, miR168, miR394, miR396, miR398, miR447 families, Figure 3A). Moreover, a few individual pri-miRNAs show profound differences in their expression depending on the developmental stage and/or organ studied (e.g. miR156, miR157, miR158, miR159, miR166, miR169, miR397, miR399, miR404, Figure 3B). These differences may reflect either the existence of promoter regulatory elements that are responsive to developmental stimuli, or the various rate of pri-miRNA maturation during plant growth and organ formation. MicroRNA families containing a single representative also may exhibit dramatic changes in their expression levels (reaching even 10^6-fold change) during plant growth and organ development (e.g. miR173, miR774, miR776, miR778, miR780, miR783, Figure 3C). Conversely, there are also pri-miRNAs that show relatively small fluctuations of expression levels during plant growth and organ formation. In conclusion, our data shows that each microRNA gene has its own characteristic expression profile reflecting its spatial and temporal regulation that can be followed and comparatively analyzed using tools implemented in the mirEX platform.

CONCLUSIONS

By creating the mirEX platform, we provide the scientific community with a high quality pri-miRNA expression data in seven developmental stages represented by 11 distinct organs of *A. thaliana*. This is a new and user-friendly platform designed to explore expression data in various developmental stages for a large number of related genes. The querying system has been limited to two simple steps, which allow access to any type of data stored in mirEX. Additionally, the mirEX interface does not require the use of a keyboard—the database user interface is fully mouse operated. This makes the process of selection and comparing data very easy and effective. Moreover, graphs presenting expression data are designed to accommodate user selection dynamically, which makes exploration of the mirEX content even more efficient.

The data currently incorporated in mirEX represents the starting point in our database development. The modular character of the database design makes it possible for further mirEX expansion to incorporate new species and datasets. Following, we plan to include *Arabidopsis* microRNA genes discovered in the future, various mutants and microRNA expression profiles from other plant species. The work on barley microRNA expression profiling is already underway. Incorporation of data from other species will broaden the available tool set allowing comparative analyses within and between species.

By designing the mirEX interface, we would like to propose a new trend in biological databases for simplicity and user-friendliness. Additionally, we put special attention to the web browser compatibility issue of the interface by testing all of the most popular tools. Careful selection of informatics techniques resulted in the platform that can be accessed even via iOS on mobile devices without loosing any of the functionality and interface features.

Carefully selected links to external databases, prevent the user interface from overloading with data, yet creates

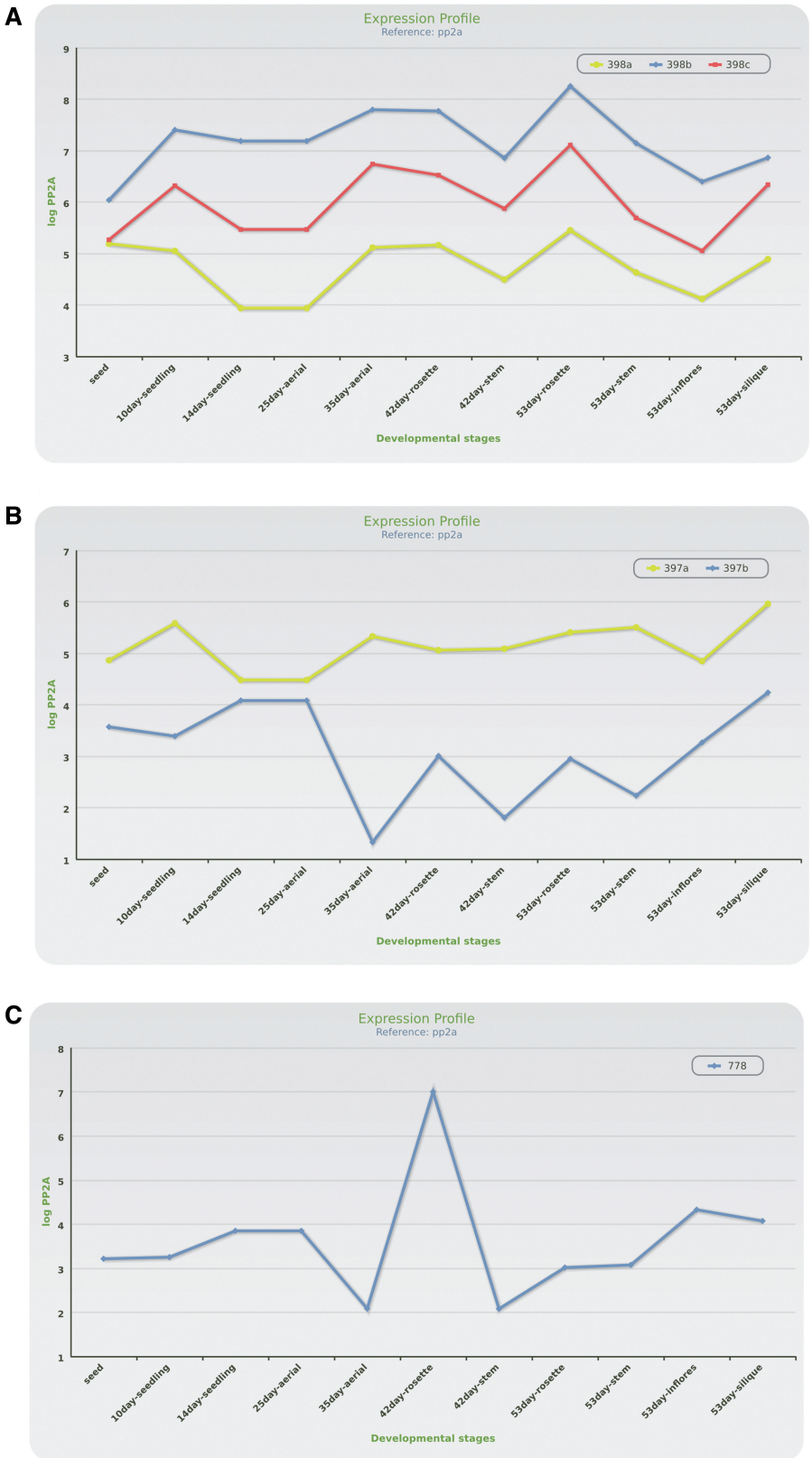


Figure 3. Examples of various microRNA gene expression profiles. The presented graphs were created with options available in mirEX interface and represent (**A**) miR398 and (**B**) miR397 gene family (see text for details). (**C**) Example of miR778 showing dramatic changes in expression level during growth and organ development.

the opportunity to easy access-related information from the most significant databases in the field. In every day laboratory work, this approach proved to be very efficient, and made the mirEX platform a one-stop information center for *Arabidopsis* microRNA data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figure S1.

ACKNOWLEDGEMENTS

We thank Professor Eva Czarnecka for the helpful suggestions in writing this manuscript and Lance Verner for proofreading it.

FUNDING

The Polish Ministry of Science and Higher Education (grant 3011/B/P01/2009/37); the Faculty of Biology Adam Mickiewicz University in Poznan, Poland; the Foundation for Polish Science (FNP) within the International PhD Program co-financed from European Union Regional Development Fund (MPD 2010/3 to D.B. and J.D.). Funding for open access charge: The Faculty of Biology Adam Mickiewicz University.

Conflict of interest statement. None declared.

REFERENCES

1. Reinhart,B., Weinstein,E., Rhoades,M., Bartel,B. and Bartel,D. (2002) MicroRNAs in plants. *Genes Dev.*, **16**, 1616–1626.
2. Llave,C., Xie,Z., Kasschau,K.D. and Carrington,J.C. (2002) Cleavage of scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science*, **297**, 2053–2056.
3. Brodersen,P., Sakvarelidze-Achard,L., Bruun-Rasmussen,M., Dunoyer,P., Yamamoto,Y.Y., Sieburth,L. and Voinnet,O. (2008) Widespread translational inhibition by plant miRNAs and siRNAs. *Science*, **320**, 1185–1190.
4. Chambers,C. and Shuai,B. (2009) Profiling microRNA expression in Arabidopsis pollen using microRNA array and real-time PCR. *BMC Plant Biol.*, **9**, 87.
5. Lee,H., Yoo,S.J., Lee,J.H., Kim,W., Yoo,S.K., Fitzgerald,H., Carrington,J.C. and Ahn,J.H. (2010) Genetic framework for flowering-time regulation by ambient temperature-responsive miRNAs in Arabidopsis. *Nucleic Acids Res.*, **38**, 3081–3093.
6. Laubinger,S., Sachsenberg,T., Zeller,G., Busch,W., Lohmann,J.U., Rätsch,G. and Weigel,D. (2008) Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana. *Proc. Natl Acad. Sci. USA*, **105**, 8795–8800.
7. Meyers,B.C., Tej,S.S., Vu,T.H., Haudenschild,C.D., Agrawal,V., Edberg,S.B., Ghazal,H. and Decola,S. (2004) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Res.*, **14**, 1641–1653.
8. Graeber,K., Linkies,A., Wood,A.T.A. and Leubner-Metzger,G. (2011) A guideline to family-wide comparative state-of-the-art quantitative RT-PCR analysis exemplified with a Brassicaceae cross-species seed germination case study. *Plant Cell*, **23**, 2045–2063.
9. Kozomara,A. and Griffiths-Jones,S. (2010) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
10. Zhao,B., Liang,R., Ge,L., Li,W., Xiao,H., Lin,H., Ruan,K. and Jin,Y. (2007) Identification of drought-induced microRNAs in rice. *Biochem. Biophys. Res. Commun.*, **354**, 585–590.
11. Moldovan,D., Spriggs,A., Yang,J., Pogson,B.J., Dennis,E.S. and Wilson,I.W. (2010) Hypoxia-responsive microRNAs and trans-acting small interfering RNAs in Arabidopsis. *J. Exp. Bot.*, **61**, 165–177.
12. Zhao,M., Ding,H., Zhu,J.-K., Zhang,F. and Li,W.-X. (2011) Involvement of miR169 in the nitrogen-starvation responses in Arabidopsis. *New Phytol.*, **190**, 906–915.
13. Backman,T.W.H., Sullivan,C.M., Cumbie,J.S., Miller,Z.A., Chapman,E.J., Fahlgren,N., Givan,S.A., Carrington,J.C. and Kasschau,K.D. (2008) Update of ASRP: the Arabidopsis small RNA project database. *Nucleic Acids Res.*, **36**, D982–D985.
14. Meng,Y., Gou,L., Chen,D., Mao,C., Jin,Y., Wu,P. and Chen,M. (2010) PmiRKB: a plant microRNA knowledge base. *Nucleic Acids Res.*, **39**, D181–D187.
15. Betel,D., Wilson,M., Gabow,A., Marks,D.S. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–D153.
16. Cho,S., Jun,Y., Lee,S., Choi,H.-S., Jung,S., Jang,Y., Park,C., Kim,S., Lee,S. and Kim,W. (2010) miRGator v2.0: an integrated system for functional investigation of microRNAs. *Nucleic Acids Res.*, **39**, D158–D162.
17. Kaya,K.D., Karakülah,G., Yakicier,C.M., Acar,A.C. and Konu,O. (2011) mESAdb: microRNA expression and sequence analysis database. *Nucleic Acids Res.*, **39**, D170–D180.
18. Ritchie,W., Flamant,S. and Rasko,J.E.J. (2010) mimiRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets. *Bioinformatics*, **26**, 223–227.
19. Pant,B., Musialak-Lange,M., Nuc,P., May,P., Walther,D. and Scheible,W. (2009) Identification of nutrient-responsive Arabidopsis and rapeseed microRNAs by comprehensive real-time PCR profiling and small RNA sequencing. *Plant Physiol.*, **150**, 1541–1555.
20. Szarzynska,B., Sobkowiak,L., Pant,B., Balazadeh,S., Scheible,W., Mueller-Roeber,B., Jarmolowski,A. and Szweykowska-Kulinska,Z. (2009) Gene structures and processing of Arabidopsis thaliana HYL1-dependent pri-miRNAs. *Nucleic Acids Res.*, **37**, 3083–3093.
21. Oñate-Sánchez,L. and Vicente-Carbajosa,J. (2008) DNA-free RNA isolation protocols for Arabidopsis thaliana, including seeds and siliques. *BMC Res. Notes*, **1**, 93.
22. Czechowski,T., Stitt,M., Altmann,T., Udvardi,M.K. and Scheible,W.-R. (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol.*, **139**, 5–17.
23. Lamesch,P., Dreher,K., Swarbreck,D., Sasidharan,R., Reiser,L. and Huala,E. (2010) Using the Arabidopsis information resource (TAIR) to find information about Arabidopsis genes. *Curr. Protoc. Bioinformatics*, Chapter 1, Unit1.11.
24. Lu,Z. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, **2011**, doi:10.1093/database/baq036.

POSTĘPY BIOLOGII KOMÓRKI TOM 36 2009 NR 1 (23–35)

REGULACJA ALTERNATYWNEGO SPLICINGU

REGULATION OF ALTERNATIVE SPLICING

Michał SZCZEŚNIAK, Zofia SZWEYKOWSKA-KULIŃSKA

Zakład Ekspresji Genów, Instytut Biologii Molekularnej i Biotechnologii,
Wydział Biologii, Uniwersytet im. A. Mickiewicza w Poznaniu

Streszczenie: Alternatywny splicing to proces, w którym z jednego pre-mRNA powstaje więcej niż jedna izoforma mRNA. Rodzaj powstającej izoformy mRNA jest wynikiem działania dość skomplikowanych mechanizmów regulacyjnych, które pozwalają na uzyskanie na przykład tkankowo-specyficznego wzoru splicingu bądź zmian komórkowego profilu splicingu na różnych etapach rozwoju. Jest to możliwe dzięki współdziałaniu szeregu elementów, spośród których należy wymienić: sekwencję nukleotydową oraz strukturę drugorzędową pre-mRNA, czynniki splicingowe oraz dodatkowe czynniki białkowe i niebiałkowe. Oprócz podstawowych sygnałów splicingowych (miejsca splicingowe, miejsce rozgałęzienia, trakt polipirymidynowy), w pre-mRNA ulegającym alternatywnemu splicingowi znajdują się elementy regulatorowe. Należą tutaj intronowe sekwencje wzmacniające (ISE) i wyciszające (ISS) oraz egzonowe sekwencje wzmacniające (ESE) i wyciszające (ESS). Ich funkcja w regulacji alternatywnego splicingu najogólniej polega na wiązaniu odpowiednich czynników działających w trans, które z kolei wpływają na wybór miejsca splicingowego przez spliceosom. Struktura drugorzędowa pre-mRNA wpływa na alternatywny splicing, decydując o dostępności cząsteczki dla czynników splicingowych. Po pierwsze, utworzenie struktury drugorzędowej wiąże się z powstaniem dwuniciowych fragmentów RNA, które są rozpoznawane przez niektóre czynniki działające w trans. Poza tym struktura drugorzędowa zmienia wzajemne ułożenie elementów cis w przestrzeni, co stwarza dodatkowe możliwości regulatorowe. Podstawową rolę w regulacji alternatywnego splicingu spełniają jednak czynniki działające w trans. Należą do nich między innymi białka SR i hnRNP. Białka SR przyłączają się przeważnie do sekwencji wzmacniających, pełniąc rolę aktywatorów splicingu. Stopień fosforylacji tych białek zmienia się wraz z przebiegiem splicingu i jest przedmiotem złożonej regulacji. Białka hnRNP przyłączają się do sekwencji wyciszających i są inhibitorami splicingu. Ich najlepiej scharakteryzowanym przedstawicielem jest białko PTB, które – jako inhibitor splicingu – uczestniczy przede wszystkim w splicingu tkankowo-specyficznym. W rzeczywistości białka aktywatorowe i inhibitorowe splicingu działają jednocześnie, a ostatecznie o wzorze splicingu decyduje rodzaj związanych białek, ich zdolność do oddziaływania z innymi białkami (w tym ze składnikami spliceosomu), jak również czynniki pozasplicingowe. Transkrypcja jest procesem dość silnie sprzężonym z alternatywnym splicingiem. Główną rolę odgrywa tutaj polimeraza RNA II, a zwłaszcza jej domena C-końcowa: CTD. Domena ta jest odpowiedzialna m.in. za rozmieszczenie czynników splicingowych i transkrypcyjnych w jądrze. Poznano szereg czynników białkowych sprzęgających splicing i transkrypcję poprzez regulowanie stopnia ufosforylowania domeny CTD. Funkcje polimerazy RNA II w splicingu zależą ponadto od rozpoznanego przez nią promotora transkrybowanego genu. Promotor może bowiem decydować o zdolności białek SR do wiązania domeny CTD lub o

procesywności polimerazy RNA II (procesywność polimerazy niekiedy wpływa na przebieg splicingu). Wzajemna regulacja splicingu i transkrypcji jest możliwa dzięki czasowemu i przestrzennemu sprzężeniu obu procesów. Splicing zostaje zahamowany podczas mitozy; efekt ten jest w głównej mierze skutkiem zmian stopnia ufosforylowania niektórych czynników splicingowych. Na przykład białko SRp38 ulega defosforylacji wraz z początkiem mitozy i w tej postaci zakłóca funkcje białek SR na wczesnym etapie splicingu. Fosforylacja tego czynnika już po zakończeniu mitozy, przywraca komórce zdolność do przeprowadzania splicingu. Splicing jest regulowany przez jeszcze inne procesy w komórce. Na przykład niektóre czynniki splicingowe wykorzystują proces degradacji mRNA niosących przedwczesny kodon stop do regulacji poziomu swojej ekspresji, jak ma to miejsce w przypadku białka PTB. Na komórkę działają różnego rodzaju bodźce zewnątrzkomórkowe, jak na przykład czynniki wzrostu, hormony i czynniki wywołujące depolaryzację błony komórkowej. Jedną z odpowiedzi komórki na te bodźce jest zmiana profilu splicingu alternatywnego. Angażowane są tutaj szlaki przekazywania sygnałów, które zmieniają stopień ufosforylowania odpowiednich czynników działających w trans (głównie białek SR). Splicing można regulować poprzez wprowadzanie do komórki pewnych związków chemicznych. Należą tutaj niskocząsteczkowe inhibitory splicingu – związki te hamują splicing na różnych jego etapach i pewne nadzieje wiąże się z wykorzystaniem ich w leczeniu schorzeń wywołanych nieprawidłowym splicingiem.

Słowa kluczowe: alternatywny splicing, białka SR, hnRNP, transkrypcja.

Summary: Alternative splicing is a process in which more than one isoform of mRNA can be produced from one pre-mRNA. Quite complex regulatory mechanisms decide which isoform is produced in a given case. These mechanisms lead e.g. to a tissue-specific or stadium-dependent splicing, which is possible due to cooperation of regulatory factors. These factors can be considered (though somewhat artificially) at several levels: pre-mRNA sequence, secondary structure of pre-mRNA, trans-acting factors and additional protein and non-protein factors. The pre-mRNA molecule apart from such splicing signals as splicing sites, branch site or polypyrimidine tract, contains cis-acting regulatory elements. These include Intron Splicing Enhancers (ISE), Intron Splicing Silencers (ISS), Exon Splicing Enhancers (ESE) and Exon Splicing Silencers (ESS). In the splicing regulation they are bound by the trans-acting factors, which in turn affect the splice site recognition by the spliceosom. The role of the secondary stucture of pre-mRNA in the regulation of alternative splicing is that it modulates the accessibility of cis-acting elements for trans-acting factors. Firstly, some RNA-binding factors recognize only double-stranded fragments of RNA and changes of the secondary structure modulate their function. Secondly, the secondary structure changes the localization of cis-acting elements in space which is another possibility for splicing regulation. However, trans-acting factors play a central role in alternative splicing. They include among others SR and hnRNP proteins. SR proteins usually bind to enhancers, being splicing activators. Phosphorylation state of SR proteins changes throughout the splicing process and is subject to a complex modulation. hnRNP proteins bind to silencers, playing a role of splicing inhibitors. Their best known member is probably PTB (Polypyrimidine Tract Binding protein), a splicing inhibitor that is engaged mainly in tissue-specific splicing. In fact splicing activators and inhibitors act together in splicing regulation and the final effect depends on the amounts of these factors, their ability to interact with other proteins (including the components of spliceosom) and even proteins engaged in other cellular processes. Transcription is quite tightly coupled with alternative splicing. RNA Pol II plays a central role there, especially its C-Terminal Domain (CTD). CTD is responsible for the nuclear localization of splicing and transcription factors. Multiple factors coupling splicing and transcription are known – they usually phosphorylate or dephosphorylate the CTD domain. The function of RNA Pol II in splicing depends also on the promoter it recognizes. The promoter may decide about the ability of SR proteins to bind to CTD or about the processivity of polymerase (that in some cases affects splicing as well). Such a co-regulation of splicing and transcription is possible due to spatial and temporal coupling of the processes. Splicing is inhibited during mitosis; this effect is achieved mainly through changes in phosphorylation state of some splicing factors. For example SRp38 protein is dephosphorylated at the beginning of mitosis and in this form (as dSRp38) it can affect proper function of SR proteins at an early stage of splicing. Phosphorylation of dSRp38 after mitosis makes a cell able to conduct splicing again. Splicing is modulated by even more cellular processes. For example Nonsense-Mediated mRNA Decay (NMD) is a way of degradation of alternative splicing products that contain Premature Termination Codon (PTC).

Some splicing factors use NMD to regulate the level of its own expression, e.g. PTB. The cell is affected by extracellular stimuli, such as growth factors, hormones and factors leading to the depolarisation of a cell membrane. The modulation of alternative splicing is one of the ways a cell can use to respond to these factors. Signal transduction pathways are engaged in this process, changing the phosphorylation state of trans-acting factors (mainly SR proteins). Splicing can also be regulated artificially through the introduction of chemical compounds to the cell. These factors include low molecular weight splicing inhibitors that affect splicing at diferent stages of the process. Such inhibitors seem to be promising in treatment of diseases caused by abnormal splicing.

Key words: alternative splicing, SR proteins, hnRNP, transcription.

Wykaz skrótów: **ASF/SF2** (*Alternative Splicing Factor/Splicing Factor 2*) – czynnik splicingowy; **Clk** (*CDC-like kinase*) – kinaza Clk; **DCL1** (*Dicer-Like 1)* – rybonukleaza DCL 1; **DSCAM** (*Down Syndrome Cell Adhesion Molecule)* – receptor DSCAM; **FGF** (*Fibroblast Growth Factor*) – czynnik wzrostu fibroblastów; **hnRNP** *(heterogeneous nuclear Ribonucleoprotein)* – heterogeniczna jądrowa rybonukleoproteina; **HYL1** (*HYponastic Leaves 1)* – białko HYL 1; **snRNP** (*small nuclear Ribonucleoprotein*) – mała jądrowa rybonukleoproteina.

WSTĘP

Splicing pre-mRNA, jest jednym z etapów ekspresji genów. Zachodzi w jądrze i występuje prawopodobnie u wszystkich Eukaryota. Polega na wycinaniu odpowiednich fragmentów pre-mRNA i łączeniu pozostałych elementów z utworzeniem cząsteczki mRNA. W przypadku tzw. *konstytutywnego splicingu*, elementami wycinanymi są zawsze introny, zaś mRNA jest składany z egzonów. Podstawowe informacje na temat budowania kompleksu splicingowego i przebiegu reakcji splicingowych można znaleźć w podręcznikach biochemii [30].

Alternatywny splicing, w odróżnieniu od konstytutywnego, może generować różne izoformy mRNA z tego samego pre-mRNA. Wyróżnia się następujące rodzaje alternatywnego splicingu:

- zachowanie intronu lub wycięcie egzonu;
- alternatywne miejsce splicingowe 5' lub 3';
- alternatywny egzon pierwszy (alternatywne miejsce inicjacji transkrypcji);
- alternatywny egzon terminalny (alternatywne miejsce poliadenylacji).

Ocenia się, że pre-mRNA 20–30% genów u roślin, a ponad 60% u człowieka, podlega alternatywnemu splicingowi [11, 26, 29, 32]. Proces ten odgrywa ważną rolę w regulacji ekspresji genów: wpływa na jakościowy i ilościowy profil białek. Jednocześnie sam alternatywny splicing jest przedmiotem wieloczynnikowej regulacji, która może zachodzić na różnych etapach procesu. Zasadniczo splicing jest regulowany przez czynniki działające w trans, które wiążą się z substratowym pre-mRNA i zmieniają zdolność spliceosomu do rozpoznawania odpowiednich miejsc cięcia.

Funkcje cząsteczki pre-mRNA i czynników splicingowych w regulacji alternatywnego splicingu są modyfikowane przez dodatkowe czynniki sprzęgające splicing z innymi procesami komórkowymi i reakcją na działanie bodźców zewnątrzkomórkowych.

STRUKTURA pre-mRNA A ALTERNATYWNY SPLICING

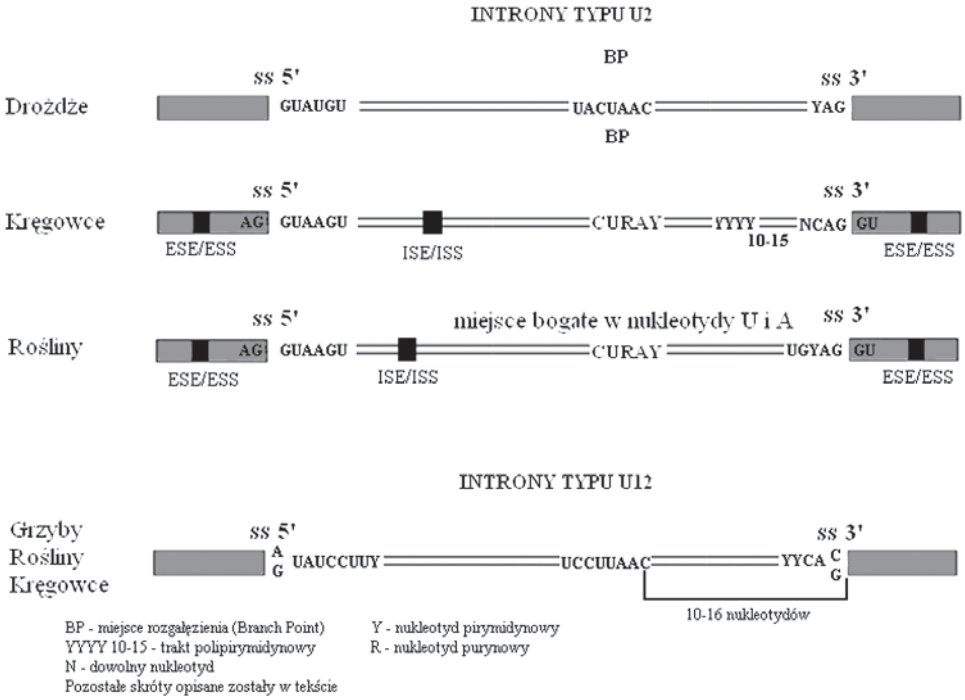
Cząsteczka pre-mRNA zawiera szereg motywów sekwencyjnych i strukturalnych, które są nieodzowne w przebiegu splicingu tak konstytutywnego, jak i alternatywnego. Motywy te są zlokalizowane w intronach i w egzonach. Należą tutaj przede wszystkim miejsca splicingowe po stronie 5' intronu (5' ss) oraz po jego stronie 3' (3' ss), miejsce rozgałęzienia czy też trakt polipirymidynowy. Istotą alternatywnego splicingu jest wykorzystywanie konkurujących miejsc splicingowych. W wyborze tych miejsc uczestniczą obecne w intronach i egzonach czynniki działające w cis, określane mianem SRE (ang. *Splicing Regulatory Elements*). Modulują one przebieg splicingu, oddziałując z odpowiednimi czynnikami działającymi w trans (m.in. białka SR i hnRNP). Do elementów SRE zalicza się:

- ISE (ang. *Intron Splicing Enhancer*) i ESE (ang. *Exon Splicing Enhancer*), stanowiące odpowiednio sekwencje wzmacniające obecne w intronach (ISE) i egzonach (ESE);
- intronowy element ISS (ang. *Intron Splicing Silencer*) oraz egzonowy element ESS (ang. *Exon Splicing Silencer*), pełniące funkcje sekwencji wyciszających (ryc. 1) [1, 29, 32, 35].

W regulacji alternatywnego splicingu sekwencje wzmacniające oddziałują głównie z białkami SR, zaś sekwencje wyciszające z białkami hnRNP. Sytuacja może nie być tak jednoznaczna, gdyż na przykład niektóre białka SR wiążą się do sekwencji wyciszających [32]. Oprócz tego udowodniono, że na różnych etapach splicingu powinowactwo białek do elementów SRE może ulegać zmianie (wskutek np. wzajemnego oddziaływania czynników splicingowych działających w trans) [32]. Funkcje elementów SRE w regulacji alternatywnego splicingu zależą w różnym stopniu również od innych procesów komórkowych oraz od reakcji na działanie bodźców zewnątrzkomórkowych.

W przypadku alternatywnego splicingu polegającego na wycinaniu egzonów lub zatrzymywaniu intronów w transkrypcie, kluczową rolę spełnia definiowanie fragmentów sekwencji pre-mRNA jako introny lub egzony. U ssaków, gdzie egzony są z reguły znacznie krótsze niż introny, przeważa definiowanie tych pierwszych i obserwuje się częstsze przypadki usuwania z transkryptu egzonu niż zachowywania intronu. U większości pozostałych Metazoa, w tym u roślin, przeważa natomiast definiowanie intronów, a alternatywny splicing polega częściej na zatrzymywaniu intronu w mRNA [31, 32].

Cząsteczka pre-mRNA moduluje przebieg alternatywnego splicingu w jeszcze jeden sposób – poprzez zmiany konformacyjne [14, 24, 26]. Istnieje kilka mechanizmów wyjaśniających to zjawisko. Zgodnie z pierwszym mechanizmem, struktura drugorzędowa pre-mRNA wpływa na dostępność miejsc splicingowych oraz elementów działających w cis dla czynników splicingowych i dla samego spliceosomu. Wykazano wpływ struktury drugorzędowej pre-mRNA na wiązanie czynników regulujących alternatywny splicing (B52, SRp55, NOVA-1) [4]. Niektóre sekwencje wyciszające tworzą strukturę drugorzędową pre-mRNA, która utrudnia rozpoznanie



RYCINA 1. Elementy działające w cis obecne w pre-mRNA i istotne z punktu widzenia alternatywnego splicingu [1]. W splicingu intronów mogą uczestniczyć dwa różne spliceosomy: spliceosom typu U2 i typu U12; introny typu U12 stanowią mniej niż 0,4% wszystkich intronów i są słabiej scharakteryzowane niż introny typu U2 [13] (wg [1], zmodyfikowane)
FIGURE 1. Cis-elements present in pre-mRNA, crucial in alternative splicing regulation [1]. Two types of spliceosomes can participate in intron splicing - type U2 and U12. The latter is much less prevalent and participates in less than 0.4% of all splicing events [13] (according to [1], changed)

sąsiednich sekwencji wzmacniających przez białka SR [29]. Inny mechanizm zakłada natomiast, że poprzez przyjęcie przez pre-mRNA odpowiedniej struktury drugorzędowej, elementy działające w cis, położone stosunkowo daleko na nici pre-mRNA mogą zostać przybliżone do kompleksu splicingowego, co umożliwia im pełnienie odpowiednich funkcji w regulacji splicingu [21].

CZYNNIKI DZIAŁAJĄCE W TRANS W REGULACJI ALTERNATYWNEGO SPLICINGU

Podstawowymi czynnikami działającymi w trans w regulacji alternatywnego splicingu są białka SR oraz białka hnRNP. U podstaw ich funkcji leży oddziaływanie z elementami działającymi w cis. Białka SR uczestniczą zarówno w konstytutywnym, jak i alternatywnym splicingu. U *A. thaliana* zidentyfikowano 19 białek SR, prawie dwukrotnie więcej niż u człowieka [26, 31, 36]. Oddziałując z elementami ISE i ESE

pełnią one funkcje aktywatorów splicingu. Uczestniczą przy tym na różnych etapach formowania spliceosomu, jak również w samych reakcjach splicingowych. Białka SR promują powstanie tzw. kompleksu E (ang. *Early complex*), zawierającego U1 snRNP oraz białko wiążące się z traktem polipirymidynowym – U2AF (ang. *U2 snRNP Auxiliary Factor*). Poprzez stymulację przyłączania U1 snRNP, białka te wpływają na wybór miejsca splicingowego 5'. Natomiast promując wiązanie U2AF, uczestniczą w wyborze 3' ss [15, 21, 28]. Dodatkowo białka SR przyłączone do egzonowych sekwencji wzmacniających wpływają na interakcję U2 snRNP z miejscem rozgałęzienia [25]. Białka SR uczestniczą także w przyłączaniu U4/U6/U5 snRNP do pre-mRNA, czemu towarzyszy przejście pre-spliceosomu w spliceosom (w tych oddziaływaniach, jak i wielu innych z udziałem białek SR, pośredniczy obecna w tych białkach domena RS bogata w powtórzenia serynowo-argininowe) [15]. Decydujące znaczenie dla funkcji białek SR ma ich stopień ufosforylowania. Fosforylacja tych białek zachodzi głównie w obrębie reszt seryny, w domenie RS; zmienia ona charakter oddziaływań typu białko-białko i białko-RNA, jak również rozmieszczenie białek SR w komórce, czego konsekwencją jest odpowiednia zmiana przebiegu alternatywnego splicingu [29]. Kolejne zmiany stanu ufosforylowania poszczególnych białek SR warunkują przejście z jednego etapu splicingu do drugiego. O ile ufosforylowane białka SR zapewniają prawidłowy charakter oddziaływań między białkami na wczesnym etapie splicingu, to gdy funkcjonalny spliceosom jest już uformowany, zachodzi defosforylacja większości białek SR [10].

Białka hnRNP są czynnikami działającymi w trans, związanymi z sekwencjami wyciszającymi. Stanowią dość dużą grupę regulatorów splicingu i charakteryzują się zróżnicowanym sposobem działania. Często obserwuje się oddziaływanie białek hnRNP (zwykle o charakterze antagonistycznym) z innymi czynnikami splicingowymi, takimi jak: SC35 i ASF/SF2 u człowieka [21].

Najlepiej poznanym przedstawicielem białek hnRNP jest białko PTB (ang. *Polypyrimidine Tract Binding protein*), zwane również hnRNP I. PTB jest represorem splicingu. Przyłącza się do fragmentów pre-mRNA bogatych w reszty C i U (np. UUCU i UCUCU). Lokalizacja tych bogatych w zasady pirymidynowe elementów w pre-mRNA ma wpływ na przebieg splicingu. Przykładowo funkcja PTB jako represora splicingu jest słaba, jeśli miejsce wiązania dla PTB znajduje się przy końcu 3' alternatywnie wycinanego egzonu [2]. Skuteczne hamowanie splicingu przez PTB wymaga czasami obecności drugiego miejsca wiązania, na przykład w obrębie regulowanego egzonu. To drugie miejsce, nawet jeśli PTB nie ma do niego wysokiego powinowactwa, może zainicjować powstanie kompleksu wielu cząsteczek PTB, który skuteczniej hamuje splicing. Z drugiej strony, dla egzonów podlegających mało wydajnemu splicingowi, represja splicingu za pośrednictwem PTB może mieć miejsce już w obecności pojedynczego miejsca wiążącego to białko [2].

U ssaków znanych jest wiele egzonów, których obecność w dojrzałym mRNA jest kontrolowana działaniem PTB, m.in. w przypadku pre-mRNA aktyniny, tropomiozyny, troponiny, c-src, receptorów FGF1 i 2 oraz IgM [2]. W przypadku pre-mRNA c-src regulacji przez PTB ulega egzon N1. W komórkach innych niż nerwowe, wycięciu egzonu N1 z transkryptu zapobiega białko PTB. W komórkach

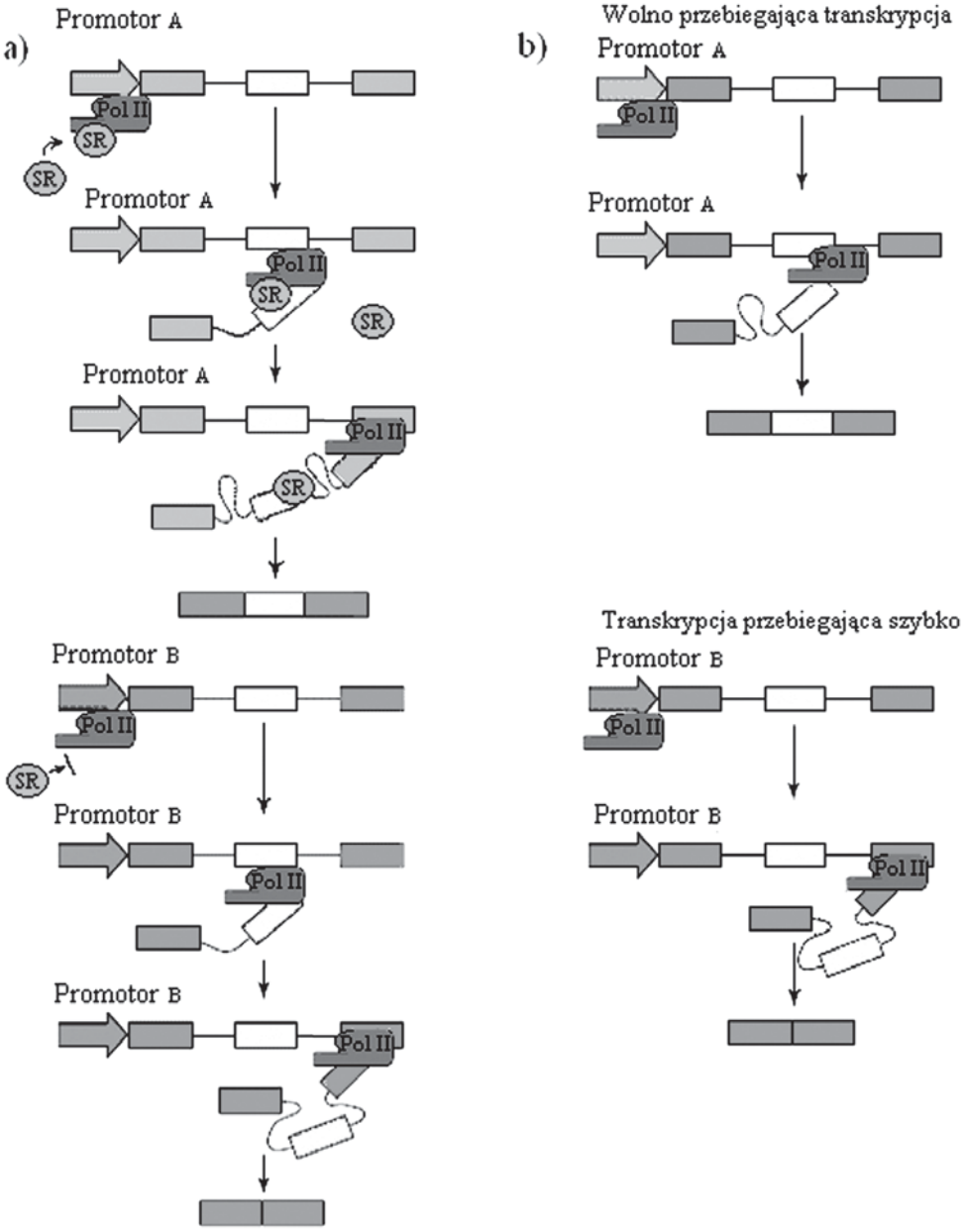
nerwowych natomiast, gdzie egzon N1 jest tracony, zlokalizowano białko nPTB (neuronowe PTB) o działaniu przeciwnym do PTB [2, 34].

Oprócz białek SR i hnRNP zidentyfikowano szereg dodatkowych czynników działających w trans i uczestniczących w regulacji alternatywnego splicingu, np. białka z rodziny CELF (ang. *CUG-BP and ETR-3 Like Factors*) u zwierząt. Białka CELF, zwane również białkami Brunopodobnymi (ang. *Bruno-like proteins*) mogą pełnić funkcję represorów bądź aktywatorów splicingu – podobna prawidłowość dotyczy, w mniejszym lub większym stopniu, także innych czynników działających w trans. U podstaw antagonistycznej funkcji czynników splicingowych stoi konkurowanie o miejsce wiązania w obrębie pre-mRNA. Przykładem jest interakcja białek ASF/SF2 i hnRNP A1. ASF/SF2 zwiększa powinowactwo U1 snRNP do 5' ss. Pod nieobecność hnRNP A1, U1 snRNP przyłącza się do obu alternatywnych 5' ss intronu. W tej sytuacji splicing zachodzi w tym 5' ss, które generuje krótszy intron. Obecność hnRNP A1 przeważnie przeszkadza w wiązaniu U1 snRNP i wówczas dochodzi do wyboru drugiego miejsca splicingowego 5' [37].

Prawidłowy przebieg alternatywnego splicingu wymaga precyzyjnej współpracy wielu czynników splicingowych; współpraca ta jest zaburzona w przypadku niektórych chorób genetycznych człowieka. Przykładem jest dystrofia miotoniczna typu 1 (DM1). U chorych na DM1 obserwuje się mutację w genie białka DMPK (ang. *Dystrophia Myotonica-Protein Kinase*), polegającą na pojawieniu się wielokrotnie powtórzonego elementu CTG w 3' UTR tego genu [20,27]. Transkrypty zmutowanego genu kumulują się w jądrze, zamiast być transportowane do cytoplazmy. Wówczas czynniki splicingowe, które wykazują powinowactwo do powtórzeń CUG (np. białka z rodziny MBNL), przyłączają się do zmutowanego RNA zamiast do swojego substratowego pre-mRNA. Skutkuje to zmianą we wzorze splicingu kilkudziesięciu do kilkuset pre-mRNA i produkcją białek, które normalnie nie występują w danym typie tkanki [27]. W konsekwencji pojawiają się u chorego takie objawy, jak: zanik mięśni, powstawanie katarakty czy oporność na insulinę. W powyższym przykładzie zwraca uwagę fakt, że białkiem, którego dysfunkcja tak znacząco zmienia komórkowy profil splicingu jest kinaza białkowa (DMPK), a nie czynnik splicingowy.

MODULOWANIE PRZEBIEGU ALTERNATYWNEGO SPLICINGU PRZEZ PROCESY WEWNĄTRZKOMÓRKOWE

Podstawowe znaczenie w przebiegu alternatywnego splicingu ma sekwencja nukleotydowa substratowego pre-mRNA (obecność miejsc wiązania dla czynników splicingowych czy struktura drugorzędowa) oraz funkcje czynników splicingowych działających w trans. Nie są to jednak jedyni gracze w regulacji splicingu – coraz więcej dowodów świadczy o tym, że alternatywny splicing jest silnie sprzężony z różnymi procesami komórkowymi, takimi jak: transkrypcja, mitoza i degradacja mRNA, niosącymi przedwczesny kodon stop.



RYCINA 2. Dwa modele regulowania alternatywnego splicingu przez polimerazę RNA II. Pierwszy (a) zakłada, że wybór jednego z alternatywnych promotorów decyduje o przyłączaniu białek SR. W drugim (b) wybór jednego z alternatywnych promotorów decyduje o szybkości i procesywności polimerazy RNA II (wg [14], zmodyfikowane)

FIGURE 2. Two models for regulation of alternative splicing by RNA Polymerase II. In the first one (a), alternative promoters have different effect on ability of SR proteins to bind to pre-mRNA molecule. According to the second model (b), chosen promoter decides about processivity of RNA Pol II (according to [14], changed)

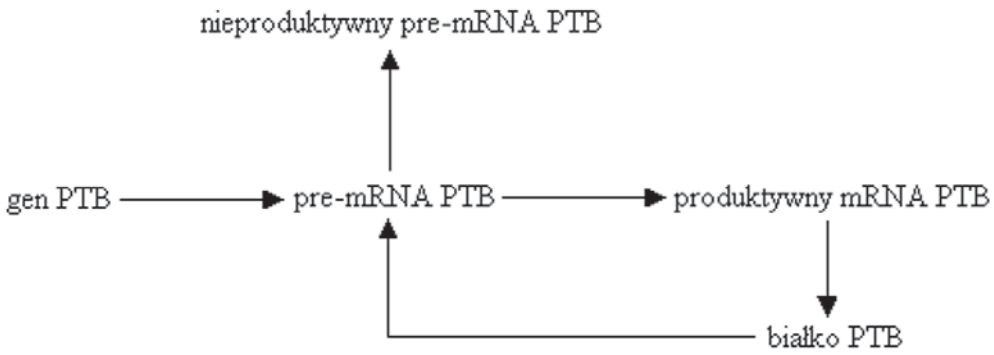
Główną rolę w sprzężeniu alternatywnego splicingu i transkrypcji pełni stopień ufosforylowania domeny C-końcowej – CTD (ang. *C-Terminal Domain*) polimerazy RNA II. W sprzężeniu tym uczestniczy szereg białek, z kinazami na czele (np. kinazy KIN28, CTK1, BUR1 i SRB10 u *S. cerevisiae*) [33]. Domena CTD ma wpływ przede wszystkim na rozmieszczenie czynników splicingowych i transkrypcyjnych w jądrze. Inicjacja transkrypcji z udziałem polimerazy RNA II powoduje nagromadzenie czynników splicingowych (m.in. białek SR) w miejscach, w których zachodzi transkrypcja. Podobnego przemieszczenia czynników splicingowych nie obserwuje się po usunięciu bądź skróceniu domeny CTD, ma wówczas miejsce znaczne ograniczenie wydajności splicingu. Poza tym, przeciwciała skierowane przeciwko Pol RNA II bądź samej domenie CTD powodują koimmunoprecypitację białek SR, hamując splicing [5, 11].

Funkcje polimerazy RNA II w regulacji alternatywnego splicingu należy rozpatrywać również w kontekście alternatywnych promotorów transkrypcji. Znane są tu przynajmniej dwa modele. W pierwszym modelu przyjmuje się, że wybór danego promotora decyduje o zdolności białek SR do wiązania się z domeną CTD polimerazy RNA II (ryc. 2a). Drugi model zakłada zaś zróżnicowaną procesywność polimerazy: jeśli wybrany promotor powoduje małą procesywność enzymu, to częstszym efektem może być usuwanie alternatywnego intronu, z jednoczesnym zachowaniem alternatywnego egzonu. W przeciwnym wypadku wydajne miejsce 3' ss intronu położonego poniżej skutecznie konkuruje z mało wydajnym miejscem 3' ss intronu leżącego powyżej, co skutkuje usunięciem alternatywnego egzonu (ryc. 2b) [14]. Oba te modele nie wykluczają się wzajemnie. Możliwe że w niektórych przypadkach uwzględnienie ich obu okaże się kluczowe w zrozumieniu mechanizmu regulacji alternatywnego splicingu.

Splicing podlega również regulacji związanej z przebiegiem cyklu komórkowego. Należy tu przede wszystkim mitotyczna inhibicja splicingu, z którą u człowieka wiąże się defosforylacja białka SRp38 – czynnika splicingowego z rodziny białek SR [3]. SRp38 jest nietypowym przedstawicielem rodziny białek SR – w stanie nieufosforylowanym jest inhibitorem splicingu. Na wczesnym etapie mitozy SRp38 ulega defosforylacji (powstaje dSRp38). dSRp38 wchodzi w słabe oddziaływania z innymi białkami SR, zakłócając ich funkcje na wczesnym etapie splicingu. Wskutek tego nie powstaje kompleks prespliceosomowy i dochodzi do blokady splicingu. Dopiero zadziałanie odpowiedniej kinazy po zakończeniu mitozy, fosforylującej dSRp38, może przywrócić komórce zdolność do przeprowadzania reakcji splicingowych [3].

Przypuszcza się, że niektóre z pozostałych białek SR, których stan ufosforylowania zmienia się podczas mitozy, mogą pełnić podobną funkcję. Co więcej, SRp38 nie jest białkiem powszechnie występującym u organizmów, co sugeruje istnienie innych czynników hamujących splicing podczas mitozy.

Splicing jest powiązany z procesem degradacji mRNA niosących przedwczesny kodon stop – NMD (ang. *Nonsense-Mediated Decay*) [6, 7, 17]. W procesie tym degradacji ulegają nieprawidłowe izoformy mRNA. Niektóre czynniki splicingowe wykorzystują NMD do regulowania poziomu swojej ekspresji np. białko PTB (ryc.



RYCINA 3. Autoregulacja alternatywnego splicingu na przykładzie białka PTB
FIGURE 3. PTB protein – an example of splicing autoregulation

3) [25]. Wykazano również związek między (alternatywnym) splicingiem a kompleksem białek wiążących się z *Cap* – CBC (ang. *Cap-Binding Complex*), który stanowi rusztowanie, ułatwiające formowanie spliceosomu. CBC uczestniczy na przykład w rozpoznawaniu miejsc 5' ss przez U1 snRNP podczas formowania spliceosomowego kompleksu E [12]. Ponadto u *A. thaliana* dostrzeżono wpływ białka SE (SERRATE) na splicing – czynnika uczestniczącego ponadto w obróbce pri-miRNA, obok DCL1 i HYL1 [12]. Powyższe wybrane przykłady obrazują jak bardzo przebieg splicingu zależy od innych procesów – w ten sposób komórka sprawniej dostosowuje wzór splicingu do bieżących potrzeb.

CZYNNIKI POZAKOMÓRKOWE WPŁYWAJĄCE NA PRZEBIEG ALTERNATYWNEGO SPLICINGU

Alternatywny splicing może być modulowany przez różnorodne czynniki zewnętrzne, takie jak: hormony, czynniki wzrostu, odpowiedź immunologiczna, czynniki wywołujące depolaryzację błony komórkowej czy stres komórkowy. Zachodzi to poprzez zmiany w syntezie i degradacji białek regulatorowych splicingu, ich komórkowej lokalizacji, jak również zdolności do tworzenia funkcjonalnych kompleksów [11, 23].

Jednym z najlepiej poznanych przykładów jest wpływ insuliny na alternatywny splicing pre-mRNA kinazy białkowej C – PKC (ang. *Protein Kinase C*) u ssaków. Białka PKCβI i PKCβII są produktami jednego genu i różnią się 50–52 C-końcowymi aminokwasami. Aktywacja receptora insuliny przez insulinę indukuje u szczura fosforylację białka SRp40. Białko to przyłącza się wówczas do pre-mRNA kinazy PKC poniżej regulowanego egzonu i egzon ten zostaje wycięty z transkryptu: powstaje PKCβI [9]. W przeciwnym wypadku, gdy nie dochodzi do aktywacji receptora insuliny, ma miejsce zatrzymanie egzonu i powstaje mRNA kodujący

C-końcowy fragment PKCβII (który decyduje o lokalizacji komórkowej i specyficzności substratowej białka) [22].

Udaje się również zmieniać wzór alternatywnego splicingu, wprowadzając do komórki niskocząsteczkowe inhibitory kinaz białkowych i innych regulatorów splicingu. Przykładowo, oparty na benzotiazolu związek o numerze katalogowym TG003 jest inhibitorem kinazy Clk1, przez co hamuje zależny od Clk1 splicing w komórkach ssaków. Poza tym związek ten zatrzymuje fosforylację białek SR. Stwierdzono, że TG003 obniża procesy fosforylacji zachodzące z udziałem kinaz Clk u żab z rodzaju *Xenopus*. Stwarza to szansę na ewentualne wykorzystanie tego związku w terapii niektórych schorzeń człowieka, u których podstaw stoi nieprawidłowy splicing z udziałem kinaz Clk [9].

Dla kinaz SRPK (ang. *SR Protein Kinase*) zidentyfikowano natomiast inhibitor o nazwie SRPIN340, zaś NB-506 (pochodna indolokarbazolu) jest inhibitorem topoizomerazy DNA I i hamuje fosforylację czynnika splicingowego ASF/SF2. Traktowanie komórek mysiej linii komórkowej p388 związkiem NB-506 zmienia wzór splicingu pre-mRNA wielu białek [9]. Co więcej możliwa jest regulacja splicingu z użyciem syntetycznych oligonukleotydów RNA. Takie oligonukleotydy przyłączają się do miejsc zawierających miejsca splicingowe na pre-mRNA i specyficznie hamują alternatywny splicing. Być może podobny mechanizm występuje w naturze – taka obserwacja mogłaby pomóc w lepszym zrozumieniu regulacji alternatywnego splicingu, jako że obecna wiedza zdaje się nie wyjaśniać całej złożoności procesu wyboru miejsc splicingowych [8, 16].

PODSUMOWANIE

Splicing jest procesem regulowanym w sposób niezwykle złożony – angażowane są tutaj różnorodne czynniki, powiązane częstokroć siecią wzajemnych oddziaływań. Pozwala to na precyzyjne dostosowanie profilu powstających cząsteczek mRNA do bieżących potrzeb komórki. Wiąże się to z wyborem jednej lub kilku form splicingowych, a bywa, że liczba przewidzianych bioinformatycznie izoform mRNA (uzyskiwanych z jednego pre-mRNA) sięga dziesiątek, a nawet tysięcy. Klasycznym przykładem jest tutaj pre-mRNA białka DSCAM u *D. melanogaster*, gdzie liczba potencjalnych izoform mRNA sięga 38 000 [10, 11]. W przypadku błędów w regulacji alternatywnego splicingu (spowodowanych np. mutacjami punktowymi w elementach SRE) mogą powstawać białka niefunkcjonalne, białka o funkcji zmienionej bądź wręcz przeciwnej. U człowieka nieprawidłowy splicing prowadzi do szeregu chorób, włączając nowotwory [10, 18, 19, 20, 27, 29]. Dobre zrozumienie regulacji alternatywnego splicingu być może pozwoli na szerokie wykorzystanie tego procesu w medycynie i biotechnologii.

LITERATURA

[1] AKERMAN M, MANDEL-GUTFREUND Y. Alternative splicing regulation at tandem 3' splice sites. *Nucl Acids Res* 2006; **34**: 23–31.
[2] AMIR-AHMADY B, BLACK DL. Exon repression by polypyrimidine tract binding protein. *RNA* 2005; **11**(5): 699–716.
[3] BLENCOWE BJ. Splicing Regulation: The Cell Cycle Connection. *Curr Biol* 2003; **13**: 149–151.
[4] BURATTI E, BARALLE FE. Influence of RNA Secondary Structure on the pre-mRNA Splicing Process. *Mol Cell Biol* 2004; **24**: 10505–10514.
[5] CACERES JF, KORNBLIHTT AR. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet* 2002; **18**: 186–193.
[6] CUCCURESE M, RUSSO G. Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucl Acids Res* 2005; **33**(18): 5965–5977.
[7] DZIKIEWICZ A, SZWEYKOWSKA-KULIŃSKA Z. Degradacja mRNA niosących przedwczesny kodon stop (NMD) – na straży jakości mRNA. *Post Biochemii* 2006; **52**(4): 390–398.
[8] GENDRON D, CHABOT B. Modulation of 5' splice site selection using tailed oligonucleotides carrying splicing signals. *BMC Biotechnol* 2006; **6**: 5.
[9] HAGIWARA M. Alternative splicing: A new drug target of the post-genome era. *Biochem Biophys Acta* 2005; **1754**: 324–331.
[10] HÄSLER J, STRUB K. Alu elements as regulators of gene expression. *Nucl Acids Res* 2006; **34**(19): 5491–5497.
[11] KORNBLIHTT AR, NOGUÉS G. Multiple links between transcription and splicing. *RNA* 2004; **10**(10): 1489–1498.
[12] LAUBINGER S, WEIGEL D. Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in *Arabidopsis thaliana. Proc Natl Acad Sci USA* 2008; **105**(25): 8795–8800.
[13] LEWANDOWSKA D, JARMOLOWSKI A. Determinants of plant U12-dependent intron splicing efficiency. *Plant Cell* 2004; **16**: 1340–1352.
[14] LIAN Y, GARNER HR. Evidence for the regulation of alternative splicing via complementary DNA sequence repeats. *Bioinformatics* 2005; **21**(8): 1358–1364.
[15] MAKAROVA O, LÜHRMANN R. The 65 and 110 kDa SR-related proteins of the U4/U6-U5 tri-snRNP are essential for the assembly of mature spliceosomes. *EMBO J* 2001; **20**: 2553–2563.
[16] MATTICK JS. A new paradigm for developmental biology. *J Exp Biol* 2007; **210**: 1526–1547.
[17] METZSTEIN M, KRASNOW M. Functions of the Nonsense-Mediated mRNA Decay Pathway in *Drosophila* Development. *PLoS Genet* 2006; **2**(12): e180.
[18] MÖRÖY T, HEYD F. The impact of alternative splicing *in vivo*: Mouse models show the way. *RNA* 2007; **13**(8): 1155–1171.
[19] ORENGO JP, COOPER TA. Alternative splicing in disease. *Adv Exp Med Biol* 2007; **623**: 212–223.
[20] OSBORNE RJ, THORNTON CA. RNA-dominant diseases. *Hum Mol Genet* 2006; **15**: 162–169.
[21] PARK J, PARISKY K. Identification of alternative splicing regulators by RNA interference in *Drosophila. Proc Natl Acad Sci USA* 2004; **101**(45):15974–15979.
[22] PATEL NA, COOPER D. Insulin regulates alternative splicing of protein kinase C (PKC) ßII through a PI-3 kinase dependent pathway involving the nuclear serine/arginine-rich splicing factor, SRp40, in skeletal muscle cells. *J Biol Chem* 2001; **276**(25): 22648–22654.
[23] PELISCH F, SREBROW A. Cross-talk between Signaling Pathways Regulates Alternative Splicing. *J Biol Chem* 2005; **280**(27): 25461–25469.
[24] POZZOLI U, SIRONI M. Silencers regulate both constitutive and alternative splicing events in mammals. *Cell Mol Life Sci* 2005; **62**: 1579–1604.
[25] ROOKE N, BLACK DL. Roles for SR Proteins and hnRNP A1 in the Regulation of c-src Exon N1. *Moll Cell Biol* 2003; **23**(6): 1874–1884.
[26] SCHINDLER S, REDDY A SN. Alternative splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes. *BCM Genomics* 2008;
[27] SMITH KP, LAWRENCE JB. Defining early steps in mRNA transport: mutant mRNA in myotonic dystrophy type I is blocked at entry into SC-35 domains. *J Cell Biol* 2007; **178**: 951–964.

[28] SMOLIŃSKI D, WRÓBEL B. Organizacja systemu splicingowego w komórkach linii generatywnej. Kosmos. *Problemy nauk biologicznych* 2003; **52**: 481–492.
[29] SREBROW A, KORNBLIHTT AR. The connection between splicing and cancer. *J Cell Sci* 2006; **119**(13): 2635–2641.
[30] TYMOCZKO JL, BERG JM, STRYER L. Biochemia. Warszawa 2005.
[31] WANG B, BRENDEL V. Genomewide comparative analysis of alternative splicing in plants. *Proc Nat Acad Sci USA* 2006; **103**: 7175–7180.
[32] WANG Z, BURGE C. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* 2008; **14**: 802–813.
[33] WILCOX C, HANES SD. Genetic Interactions With C-Terminal Domain (CTD) Kinases and the CTD of RNA Pol II Suggest a Role for ESS1 in Transcription Initiation and Elongation in *Saccharomyces cerevisiae. Genetics* 2004; **167**: 93–105.
[34] WOODLEY L, VALCARCEL J. Regulation of alternative pre-mRNA splicing. *Brief Funct Gen Prot* 2002; **1**: 266–277.
[35] ZAHLER A. Alternative splicing in *C. elegans. WormBook* 2005; **26**: 1–13.
[36] ZHANG Z, KRAINER A. Involvement of SR Proteins in mRNA Surveillance. *Mol Cell* 2004; **16**: 597–607.
[37] www.eurasnet.info/files/presentations/RNA%20Splicing.ppt.

Redaktor prowadzący – Maria Olszewska

Otrzymano: 12.05.2008 r.
Przyjęto: 08.12. 2008 r.
Zakład Ekspresji Genów, Instytut Biologii Molekularnej i Biotechnologii,
Wydział Biologii,Uniwersytet im A. Mickiewicza w Poznaniu,
ul. Umultowska 89, 61-614 Poznań,
e-mail: zofszwey@amu.edu.pl