Sari Agriculture Sciences and Natural Resources University

Research on Animal Production, Vol.14, No. 41, Autumn 2023 p: 33-44

"Research Paper"

Classification of microRNA precursors using reduced features of dinucleotide repeats in cattle (*Bos Taurus*)

Atefeh Seyeddokht[1], Javad Rahmaninia[2] and Hasan Karami[3]

1- Assistant professor Animal Science Research Department, Khorasan Razavi Agricultural and Natural Resources Research and Education Center, AREEO, Mashhad, Iran, (Corresponding author: atefeh.seyeddokht@gmail.com)
2- Assistant professor Animal Science Research Institute of Iran, Agricultural Research, Education and Extension Organization (AREEO), Karaj, Iran
3- MSc Animal Science Research Institute of Iran, Agricultural Research, Education and Extension Organization (AREEO), Karaj, Iran

Received: 31 May, 2023 Accepted: 19 September, 2023

Abstract

Introduction and Objective: The latest major advances in transcriptomics technologies, especially next-generation sequencing technologies and advanced bioinformatics tools, allows deeper exploration of messenger RNAs (mRNAs) and non-coding RNAs (ncRNAs), including miRNAs. These technologies have offered important chance for a deeper study of miRNA association in farm animal diseases, as well as livestock productivity and welfare. Since the discovery of lin-4 and let-7, many microRNAs have been identified in farm animal species and deposited in miRNA databases. miRNA can be used as biomarkers in the context of farm animal disease diagnostics, prediction, and therapeutic purposes, for the management of livestock diseases. By the sequencing of *Bos Taurus* (cattle) genome, we have an opportunity to discover novel miRNAs in this species. However, the experimental determination of miRNA sequence and structure is both expensive and time-consuming, therefore, computational and machine learning-based approaches have been adopted to predict novel microRNAs in the *Bos Taurus* (cattle) genome.

Material and Methods: Finding an accurate method for Identification of miRNA molecules can help for understanding of regulatory processes. Currently, computational methods based on learning algorithms have been extensively applied for miRNA prediction.

Inspired by the work of predecessors, we proposed an improved computational model based on random forest (RF) for identifying real miRNA precursor sequences (pre-miRNAs). First, the occurrence frequencies of the dinucleotide of pre-miRNAs genes, and the percentage of G+C content were calculated. The observed dinucleotide composition was calculated as the structural features of the sequence composition for each miRNA gene. A total of cattle (*Bos Taurus*) dinucleotide compositions with their genomic G+C contents for 1064 genes encoding miRNA and non-miRNA sequences were calculated. In the next step two classification models based on machine learning approach were trained to identify real and pseudo bovine pre-miRNAs. One set of 17 optimized features related to sequence structures were used to train the models. These models were trained and validated with 10-fold cross validation method.

Results: Our goal was to investigate the predictive performance of RNA features in distinguishing pre-miRNAs from pseudo hairpins. Our model achieved 99% precision, and 97.9% MCC using *Bos Taurus* datasets.

Conclusion: Computational methods of Artificial intelligence can detect novel potential miRNAs in the bovine genome, some of which to have previously undetected in this genome. As a result, it seems necessary to use computational methods to identify these regulatory RNAs in livestock for breeding purposes. Our discoveries support that dinucleotide features will be beneficial to achieve the highest accuracies for miRNA sequences prediction.

Keywords: Bioinformatic, Cattle (*Bos Taurus*), Computational identification, Machine learning, miRNA.

[DOR: 20.1001.1.22518622.1402.14.41.4.9]

[DOI: 10.61186/rap.14.41.33]

دانشگاه علوم کشاورزی و منابع طبیعی ساری

"مقاله پژوهشی"

طبقه‌بندی پیش‌سازهای microRNA در گاو (*Bos Taurus*) با استفاده از ویژگی‌های کاهش یافته تکرارهای دو نوکلئوتیدی

عاطفه سیددخت[۱]، جواد رحمانی‌نیا[۲] و حسن کرمی[۳]

۱- استادیار بخش تحقیقات علوم دامی، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان خراسان رضوی، سازمان تحقیقات، آموزش و ترویج کشاورزی، مشهد، ایران، (نویسنده مسوول: atefeh.seyeddokht@gmail.com)
۲- استادیار مؤسسه تحقیقات علوم دامی کشور، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران
۳- مربی مؤسسه تحقیقات علوم دامی کشور، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران
تاریخ دریافت: ۱۴۰۲/۳/۱۰ تاریخ پذیرش: ۱۴۰۲/۶/۲۸
صفحه: ۳۳ تا ۴۴

چکیده مبسوط

مقدمه و هدف: توسعه مداوم فناوری‌های مولکولی مورد استفاده برای تجزیه و تحلیل رونوشت‌ها، به‌ویژه فناوری‌های توالی‌یابی نسل بعدی و ابزارهای پیشرفته بیوانفورماتیک، امکان کاوش عمیق‌تر RNA های پیام‌رسان (mRNAs) و RNAهای غیرکد کننده (ncRNA) از جمله miRNAها را فراهم می‌کند. این فناوری‌ها فرصت‌های بزرگی را برای اکتشاف عمیق‌تر دخالت miRNAها در بیماری‌های حیوانات مزرعه و همچنین بهره‌وری و رفاه دام ارائه نمودند. از زمان کشف lin-4 و let-7، هزاران miRNA در گونه‌های حیوانات مزرعه شناسایی و در پایگاه‌های داده miRNA ثبت شده‌اند. miRNAها را می‌توان به‌عنوان نشانگرهای زیستی، اهداف تشخیصی، پیش‌آگهی و درمانی برای مدیریت بیماری‌های دام استفاده کرد. با تعیین توالی ژنوم گاو (*Bos Taurus*)، فرصتی برای کشف miRNAهای جدید در این گونه فراهم خواهد شد. از آنجایی‌که تعیین توالی و ساختار miRNA ها به‌صورت آزمایشگاهی هزینه‌بر و زمان‌بر است، بنابراین این پژوهش با هدف استفاده از روش‌های محاسباتی مبتنی بر یادگیری ماشین به‌منظور پیش‌بینی microRNAها در ژنوم گاو انجام شد.

مواد و روش‌ها: یافتن روشی دقیق برای شناسایی مولکول‌های miRNA می‌تواند به درک فرآیندهای تنظیمی کمک کند. در حال حاضر روش‌های محاسباتی مبتنی بر الگوریتم‌های یادگیری به‌طور گسترده برای پیش‌بینی miRNAها استفاده می‌شوند. با الهام از سایر تحقیقات انجام شده در زمینه شناسایی miRNAها، یک مدل محاسباتی بهبود یافته یادگیری ماشین برای شناسایی توالی‌های پیش‌ساز miRNAهای واقعی (pre-miRNA) پیشنهاد شد. در مرحله اول فراوانی توالی‌های دو نوکلئوتیدی ژن‌های pre-miRNA و محتوای بازهای سیتوزین و گوانین (G+C) در توالی‌ها در نظر گرفته شد. ترکیب دی‌نوکلئوتیدی مشاهده شده به‌عنوان ویژگی‌های ساختاری ترکیب توالی برای هر ژن miRNA محاسبه شد. مجموع ترکیبات دو نوکلئوتیدی در گونه گاو (*Bos Taurus*) با محتویات ژنومی G+C برای ۱۰۶۴ توالی miRNA و توالی‌های غیر miRNA محاسبه شد. در مرحله بعد دو مدل طبقه‌بندی مبتنی بر رویکرد یادگیری ماشین برای شناسایی pre-miRNA های واقعی و شبه واقعی آموزش داده شدند. مجموعه‌ای از ۱۷ ویژگی بهینه شده مربوط به ساختارهای توالی برای آموزش مدل‌ها استفاده شد. این مدل‌ها با روش اعتبارسنجی متقاطع ۱۰ تایی آموزش یافتند و اعتبارسنجی شدند.

یافته‌ها: هدف بررسی عملکرد پیش‌بینی طبقه‌بندی کننده‌ها براساس ویژگی‌های RNA در تشخیص pre-miRNA ها از سایر توالی‌ها بود. مدل آنالیز شده در این پژوهش با استفاده از مجموعه داده‌های گاو (*Bos Taurus*) به‌دقت ۹۹ درصد و ضریب همبستگی متیو ۹۷/۹ درصد دست یافت.

نتیجه‌گیری: روش‌های محاسباتی هوش مصنوعی می‌توانند miRNAهای بالقوه جدیدی را در ژنوم گاو شناسایی کنند که برخی از آنها قبلاً در این ژنوم شناسایی نشده بودند. در نتیجه لزوم استفاده از روش‌های محاسباتی جهت شناسایی این RNAهای تنظیمی در دام‌ها جهت اهداف اصلاحی ضروری به‌نظر می‌رسد. نتایج این پروژه نشان داد که تنها با استفاده از ویژگی‌های ساختاری دو نوکلئوتیدی می‌توان در پیش‌بینی توالی‌های miRNA به‌دقت بالایی دست یافت.

واژه‌های کلیدی: بیوانفورماتیک، شناسایی محاسباتی، گاو (*Bos Taurus*)، یادگیری ماشینی، miRNA

مقدمه

اهلی شدن گاو (*Bos Taurus* و *Bos indicus*) حدود ۸۰۰۰ تا ۱۰۰۰۰ سال پیش (Barazandeh et al. 2019; Moradian *et al*. 2019) در دو رویداد جداگانه انجام شد: یکی در منطقه هلال بارور (Fertile Crescent) (منشاء گاو تورین) و دیگری در دره سند (منشاء گاو ایندیسین یا زبو) (Barazandeh et al. 2016; Moradian et al. 2019). از آن زمان، طیف گسترده‌ای از رویدادهای انتخاب طبیعی و مصنوعی، ویژگی‌های مهم گاو مانند سازگاری با محیط‌های مختلف، تولیدمثل، فرم بدن، رفتار، مقاومت در برابر بیماری‌ها و انگل‌ها و ویژگی‌های اقتصادی مطلوب را به‌شدت تغییر داده است (Barazandeh et al. 2016; Bordbar et al., 2022).

microRNAها[۱] گروه بزرگی از RNAهای کوتاه ۱۹ تا ۲۵ نوکلئوتیدی هستند که به پروتئین ترجمه نمی‌شوند و تنظیم بیان ژن‌ها را بر عهده دارند و از اجزای کلیدی تنظیم بیان ژن در گیاهان و جانوران محسوب می‌شوند. این مولکول‌ها در طی تکامل حفظ شده‌اند و تنظیم بیان ژن را در مرحله پس از رونویسی[۲] انجام می‌دهند. تنظیم و کاهش بیان ژن توسط miRNAها می‌تواند یا از طریق القای برش[۳] mRNAهای هدف باشد یا از طریق سرکوب ترجمه[۴] آن‌ها انجام گیرد. مسیر عملکردی mRNAها متعلق به شبکه گسترده تنظیم بیان ژن است که تحت عنوان RNAi شناخته می‌شود. در چند سال گذشته پیشرفت‌های بسیاری در روش‌های خاموش‌سازی ژن براساس RNAi ایجاد شده است که محققان از آن‌ها جهت خاموش‌سازی ژن‌ها بهره می‌برند. امروزه مشخص شده است

1-miRNAs
2- post transcriptional
3- cleavage
4- transcriptional repression

که می‌توان از RNAi در زمینه‌های مختلفی از قبیل شناسایی مسیرهای پیام‌رسانی[1]، شناسایی عملکرد ژن‌ها و درمان بیماری‌ها استفاده کرد (Castel and Martienssen, 2013; De Souza et al., 2009; Do et al., 2021; Tzelos et al., 2022).

تعداد بسیاری از miRNAها در حیوانات معرفی شده‌اند. با وجود اینکه miRNAها نقش مهمی در بررسی عملکردهای زیستی حیوانات دارند و منابع ژنومی خوبی برای درک مکانیسم‌های بیان ژن در دام‌ها هستند، اما در مورد miRNAها در نشخوارکنندگان اطلاعات کمی وجود دارد. در این میان روش‌های بیوانفورماتیک و همچنین مقایسه ژنومی به‌علت ماهیت حفاظتی، روش و منبع مفیدی جهت کشف و شناسایی miRNAها هستند. در بین نشخوارکنندگان گاو یک حیوان مفید به‌لحاظ تولیدات بوده و همچنین یک حیوان مدل ایده‌آل برای مطالعات بیولوژی مرتبط با انسان است. لذا تحقیق و مطالعه در مورد آن امری ضروری و سودمند است. miRNAها عناصر بسیار مهم ژنتیکی هستند که قادر به تأثیر بر بیان ژن‌های کلیدی یا ژن‌های تنظیم‌کننده بیان در مسیرهای مؤثر بر صفات اقتصادی دام‌ها می‌باشند. تعداد miRNAهای شناخته شده اختصاصی در گونه‌های دامی رو به افزایش است و تحقیقات نشان داده‌اند که این توالی‌ها به‌شدت محافظت شده هستند. روش‌های مختلفی برای شناسایی miRNAها در گونه‌های مختلف مورد استفاده قرار می‌گیرد، از جمله استفاده از داده‌های NGS و ریزآرایه‌ها، تجزیه و تحلیل EST و یاGSS (Akhtar et al., 2016; Huang et al., 2015). مهمترین ویژگی‌هایی که برای شناسایی بیوانفورماتیکی miRNA مورد استفاده قرار می‌گیرد، طول و حافظت شدگی توالی‌ها در بین گونه‌های مختلف و ویژگی‌های ساختاری نظیر ساختار سنجاق سر و حداقل انرژی تاخوردگی است (de Sousa et al., 2021; Fan et al., 2021). تجزیه و تحلیل محاسباتی داده‌های miRNA می‌تواند در پیش‌بینی miRNAهای محافظت شده در گونه‌های مختلف نشخوارکنندگانی مانند گاو مورد استفاده قرار گیرد (de Sousa et al., 2021). در این مطالعه از یک روش محاسباتی براساس ویژگی‌های دی‌نوکلئوتیدی، برای شناسایی miRNAهای گاو استفاده شد.

از آنجایی‌که در آینده انجام برنامه‌های اصلاح نژادی در جهت افزایش بهره‌وری از این پستاندار به‌سمت تحقیقات مولکولی متمایل خواهد شد، در نتیجه با استفاده از روش‌های جدید شناسایی miRNAها می‌توان به پیشرفت سریع‌تری در فرایندهای اصلاحی دست یافت و همچنین مسیر شناسایی فرایندهای تنظیم فعالیت‌های سلولی را نیز بهبود بخشید. تعداد miRNAهای شناسایی شده پیش‌ساز و بالغ به‌ترتیب ۱۰۶۴ و ۱۰۲۵ و مطالعات مرتبط با miRNA در گونه گاو به‌عنوان یک حیوان مزرعه ۸۷۰ مطالعه تا سال ۲۰۲۱ گزارش شده است (Do et al., 2021). در مقایسه با رویکردهای رگرسیون، روش‌های مختلفی با عنوان سیستم‌های فازی عصبی و شبکه‌های عصبی مصنوعی (ANN) برای حل مشکلات ناشی از روش‌های رگرسیون سنتی وجود دارد (Ghotbaldini et al., 2017). شبکه‌های عصبی مصنوعی الگوریتم‌های یادگیری و مدل‌های ریاضی هستند که توانایی پردازش اطلاعات مغز انسان را تقلید می‌کنند و می‌توانند برای داده‌های غیر خطی و پیچیده استفاده شوند، حتی اگر داده‌ها نادقیق باشند (Pour Hamidi et al., 2017). این شبکه‌ها شامل مجموعه‌ای از اجزای پردازشی هستند که به‌عنوان نورون‌ها یا گره‌ها نیز شناخته می‌شوند که عملکرد آن‌ها بر اساس نورون‌های بیولوژیکی است (Pour Hamidi et al., 2017). این واحدها در لایه‌هایی تشکیل می‌شوند که اطلاعات ورودی را پردازش کرده و به لایه زیر منتقل می‌کنند (Pour Hamidi et al., 2017). توانایی شبکه در پردازش در نقاط قوت (یا وزن‌های) اتصال بین واحدی که از طریق فرآیند انطباق با مجموعه‌ای از الگوهای آموزشی به‌دست می‌آیند، انباشته می‌شود (Ghotbaldini et al., 2019).

مواد و روش‌ها

miRBase

با توجه به گسترش روزافزون تعداد miRNAهای کشف شده و کاربرد روش‌های بیوانفورماتیک برای پیش‌بینی miRNAها لزوم ایجاد پایگاه‌هایی که اطلاعات مربوط به آنها را جمع‌آوری و سازماندهی نمایند، بسیار ضروری می‌باشد. در این راستا تعدادی پایگاه داده که تمامی اطلاعات علمی مربوط به miRNAها را جمع‌آوری و ارائه می‌کنند، پدید آمده است. یکی از مهمترین این پایگاه‌ها miRBase (http://microrna.sanger.ac.uk) است. در پایگاه miRBase، تا نوامبر ۲۰۲۲ تعداد miRNAهای کشف شده برای گونه انسان ۱۹۱۷ و برای گونه گاو ۱۰۶۴عدد گزارش شده است. در این پژوهش از توالی‌های miRNA پیش‌ساز گاو (*Bos Taurus*) نسخه miRBase ۲۲ استفاده شد.

از آنجایی‌که توزیع نامتعادل می‌تواند بر عملکرد طبقه‌بندی کننده تأثیر بگذارد، تعادل بین تعداد نمونه‌ها در مجموعه‌های آموزشی مثبت و منفی حفظ شدند (Lertampaiporn et al., 2014).

راه‌اندازی آزمایش

برای مجموعه داده‌های آموزشی گاو یک اعتبارسنجی متقاطع ده‌تایی[2] انجام شد. به‌طور تصادفی ۸۰ درصد از داده‌ها برای تشکیل مجموعه آموزشی انتخاب گردید. ۲۰ درصد باقی‌مانده به‌منظور ارزیابی اعتباریابی الگوریتم‌ها به‌عنوان مجموعه تست استفاده شدند (Wang et al., 2019). به‌منظور ارزیابی مقایسه مدل‌های بیزی و جنگل تصادفی، حساسیت[3] (SE یاTP rate)، امتیاز F و ضریب همبستگی متیو[4] جهت ارزیابی عملکرد مدل‌های پیش‌بینی استفاده شدند. ویژگی‌هایی شامل یک ویژگی درصد بازهای گوانین و سیتوزین (%G+C) و ۱۶ ویژگی دی‌نوکلئوتیدی شامل فراوانی توالی آدنوزین–آدنوزین (%AA)، فراوانی توالی آدنوزین–سیتوزین (%AC)، فراوانی توالی آدنوزین–گوانین (%AG)، فراوانی توالی

1- signaling
2-10-Fold Cross Validation
3-Sensitivity (true positive rate)
4- Matthews Correlation Coefficient (MCC)

آدنوزین–اوراسیل (%AU)، فراوانی توالی سیتوزین–آدنوزین (%CA)، فراوانی توالی سیتوزین–سیتوزین (%CC)، فراوانی توالی سیتوزین–گوانین (%CG)، فراوانی توالی سیتوزین–اوراسیل (%CU)، فراوانی توالی گوانین–آدنوزین (%GA)، فراوانی توالی گوانین–سیتوزین (%GC)، فراوانی توالی گوانین–گوانین (%GG)، فراوانی توالی گوانین–اوراسیل (%GU)، فراوانی توالی اوراسیل–آدنوزین (%UA)، فراوانی توالی اوراسیل–سیتوزین (%UC) و فراوانی توالی اوراسیل–گوانین (%UG) در مجموعه داده‌های آموزشی و تست آنالیز شدند.

محاسبه توالی‌های دی‌نوکلئوتیدی

نحوه محاسبه درصد بازهای گوانین و سیتوزین (%G+C) و ۱۶ ویژگی دی‌نوکلئوتیدی به‌ترتیب در زیر آورده شده است:

برای ویژگی درصد بازهای گوانین و سیتوزین از فرمول محاسباتی زیر استفاده شد (L= طول رشته):

$$\%(C+G)=(|C|+|G|)/L*100 \quad (۱)$$

برای ۱۶ ویژگی ساختاری دی‌نوکلئوتیدی روش زیر به‌کار گرفته شد (X و Y نوع نوکلئوتید می‌باشد):

$$\%XY=|XY|/(L)*100 \quad (۲)$$

در شکل زیر به‌عنوان مثال نمونه‌ای از سه توالی miRNA پیش‌ساز، که بر متابولیسم چربی و افزایش طول اسید چرب شیر گاو مؤثر هستند، در نماد نقطه براکت نشان داده شده است (شکل ۱) (Chen et al., 2019).

>bta-mir-155 MI0009752
UGUUAAUGCUAAUCGUGAUAGGGGUUUUUACCUCGGACUGACUCCUACAUGUUAGCAUUAACA
((((((((((.((((.(((((((.....)))))))))))))))))))))
>bta-mir-103-1 MI0004736
CUGCCCUCGGCUUCUUUACAGUGCUGCCUUGUUGCAUAUGGAUCAAGCAGCAUUGUACAGGGCU
AUGAAGGC
..(((.(((((((.((.((((((((((.(((.....)))))))))))))).)))))))..)).)))
>bta-mir-103-2 MI0005456
UUGUGCUUUCAGCUUCUUUACAGUGCUGCCUUGUAGCAUUCAGGUCAAGCAGCAUUGUACAGGG
CUAUGAAAGAAC
.....(((((((((.((.((((((((((((.....)))).....)))))))))).)))))..)))))...

شکل ۱- سه توالی pre-miRNA در گاو (*Bos Taurus*)
Figure 1. Three pre-miRNA sequences in cattle (*Bos Taurus*)

بررسی دسته‌بندی کننده‌ها

تا به امروزه روش‌های دسته‌بندی زیادی به‌وجود آمده‌اند و هرکدام از آن‌ها دارای کاربردهای زیاد و گوناگونی در علوم زیستی می‌باشند. مسئله پیش‌بینی ژن‌های miRNA را می‌توان یک مسئله دسته‌بندی دودویی[1] درنظر گرفت (Chen et al., 2016). برای حل مسئله دسته‌بندی مدل‌های هوشمند مختلفی ارائه شده است. دو روشی که در دسته‌بندی ژن‌های miRNA گاو در این پروژه مورد استفاده قرار گرفتند، به‌طور خلاصه بررسی می‌شوند.

الگوریتم جنگل تصادفی

جنگل تصادفی[2] یک الگوریتم یادگیری ماشینی با قابلیت استفاده آسان است که اغلب اوقات نتایج بسیار خوبی را حتی بدون تنظیم پارامترهای آن فراهم می‌کند. این الگوریتم به‌دلیل سادگی در قابلیت استفاده، هم برای دسته‌بندی (Classification) و هم رگرسیون (Regression)، یکی از پرکاربردترین الگوریتم‌های یادگیری ماشین محسوب می‌شود. جنگل تصادفی به‌عنوان یک مدل یادگیری ماشین نظارت شده، یاد می‌گیرد که در فاز آموزشی (training) یا برازش مدل، داده‌ها را (توالی‌های آموزشی) به خروجی‌ها (احتمال پیش‌بینی توالی) نگاشت کند. در طول آموزش، داده‌های آموزشی به مدل داده می‌شوند که مرتبط با دامنه مسأله هستند و مقدار صحیحی (که در این مسئله miRNA بودن است) که مدل باید بیاموزد تا بتواند پیش‌بینی کند. مدل، روابط میان داده‌ها را (که با عنوان ویژگی‌ها در یادگیری ماشین شناخته شده‌اند) و مقادیری که می‌خواهیم آن‌ها را پیش‌بینی کند (به آن‌ها هدف گفته می‌شود) می‌آموزد. این الگوریتم، به‌جای جست‌وجو به‌دنبال مهم‌ترین ویژگی‌ها هنگام تقسیم کردن یک گره (Node)، به‌دنبال بهترین ویژگی‌ها در میان مجموعه تصادفی از ویژگی‌ها می‌گردد. این امر منجر به تنوع زیاد و در نهایت مدل بهتر می‌شود. بنابراین در جنگل تصادفی، تنها یک زیر مجموعه از ویژگی‌ها توسط الگوریتم برای تقسیم یک گره در نظر گرفته می‌شود (Yao et al., 2019).

شبکه‌های بیزی

شبکه‌های بیزی[3] از تعدادی معیار استفاده می‌کند که توزیع‌های متمایزی را بین miRNAهای شناخته شده و RNAهای کوچک دیگر نشان می‌دهد، تا یک طبقه‌بندی‌کننده بیزی را آموزش دهد و احتمال این را که توالی خوانده شده یک miRNA است، تخمین بزند. ویژگی‌هایی که در این پژوهش در نظر گرفتیم، شامل فراوانی‌های دی نوکلئوتیدی است. هیچ یک از این معیارها به‌تنهایی برای تعیین قطعی اینکه آیا یک توالی miRNA است یا خیر کافی نیست. طبقه‌بندی‌کننده بیزی فرض می‌کند که این معیارها مستقل هستند و بر اساس حاصل ضرب نسبت‌های شانس آن‌ها، احتمال اینکه یک توالی شناسایی شده یک miRNA باشد را محاسبه می‌کند. توانایی طبقه‌بندی‌کننده بیزی حاصل هم به‌صورت محاسباتی و هم آزمایشی تعیین می‌شود. سپس با استفاده از معیارهای حساسیت و دقت نشان می‌دهیم که طبقه‌بندی‌کننده با مقادیر بسیار بالای این معیارها عملکرد مناسبی دارد. همچنین نشان می‌دهیم که

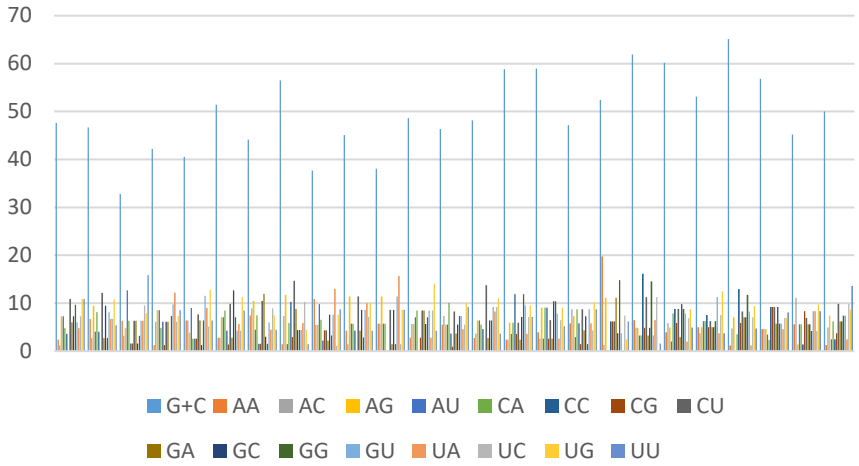
1-binary classification problem
2-Random Forest
3-Bayesian Networks

miRNAهای شناسایی شده با امتیاز بالا نسبت به نمونه‌های با امتیاز پایین از نظر نرخ مثبت درست، بیشتر آزمایش‌های ارزیابی را پشت سر می‌گذارند. به‌طور خاص پیش‌بینی می‌شود که بسیاری از کاندیدها دارای امتیاز بالای نرخ مثبت درست، miRNAهایی شناخته شده با صحت بالا هستند (Douglass et al., 2016; Magyar, 2018).

نتایج و بحث

نتایج آنالیز ویژگی‌های فراوانی جفت نوکلئوتیدی و درصد G+C در شناسایی miRNA

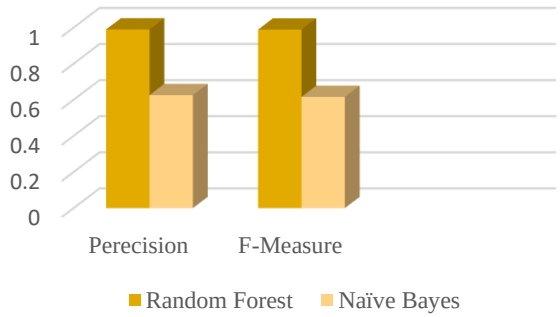
ویژگی‌های استخراج شده برای ۱۰۶۴ توالی miRNA گاو و همچنین توالی‌های غیرکدکننده دیگر توسط دو الگوریتم بیزی و جنگل تصادفی آموزش داده شدند. فراوانی این دی‌نوکلئوتیدها برای ۲۵ توالی miRNA گاو در شکل ۲ نمایش داده‌شده است.



شکل ۲- نمونه‌ای از فراوانی توالی‌های دو نوکلوتیدی و G+C (درصد) برای ۲۵ توالی پیش‌ساز miRNA گاو

Figure 2. Example of abundance of dinucleotide and G+C sequences (%) for 25 bovine miRNA precursor sequences.

توالی‌های miRNA برای ویژگی‌های جفت نوکلئوتیدی و درصد G+C به دو گروه تقسیم بندی شدند (۸۰ درصد آموزش و ۲۰ درصد توالی‌های تست) و در هر گروه مدل‌های بهینه‌سازی شده اجرا شدند و نتایج خروجی‌های مدل‌های بیزی و جنگل تصادفی مورد مقایسه قرار گرفتند.



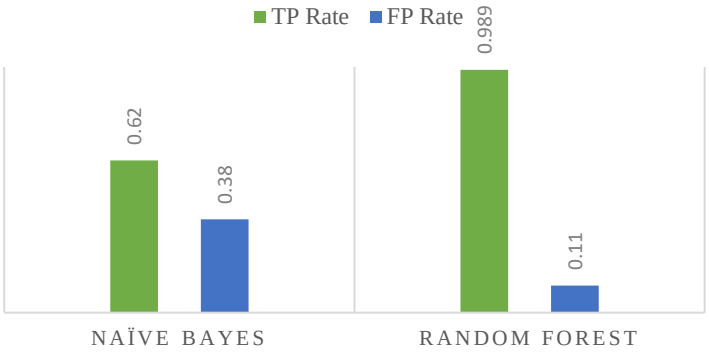
شکل ۳- آماره F و دقت با استفاده از دو مدل مختلف

Figure 3. Comparative results with F-measure, and Precision between different models on training datasets

نتایج به‌دست آمده در شکل ۳ نشان می‌دهد که در توالی‌های miRNA ژنومی برای گونه گاو، الگوریتم جنگل تصادفی می‌تواند با دقت[1] بالاتری (۰/۹۹) توالی‌های miRNA پیش‌ساز را نسبت به مدل بیزی (۰/۶۲۶) به‌درستی پیش‌بینی کند. از نظر آماره F مدل جنگل تصادفی ۹۸/۹ و مدل بیزی ۶۱/۶ درصد داده‌ها را به‌درستی پیش‌بینی کرد. در طول آنالیزها از داده‌های اعتبارسنجی برای ارزیابی میزان بهینه بودن ساختار یک مدل نسبت به مدل دیگر استفاده شد. مقادیر اولیه پارامترهای مدل نقش مهمی در حرکت مدل به‌سمت جواب بهینه دارند. با توجه به اینکه داده‌های آموزشی ممکن است نتیجه مدل را اریب کنند، از روش اعتبارسنجی متقاطع ۱۰ تایی استفاده شد.

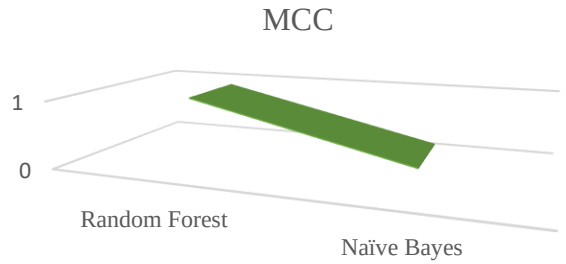
1-Perecision

با توجه به مقادیر حساسیت (نرخ مثبت درست)[1] و مثبت غلط[2] می‌توان بیان نمود که الگوریتم جنگل تصادفی با استفاده از ویژگی‌های دی‌نوکلئوتیدی با مقدار نرخ مثبت غلط ۰/۱۱، میزان خطای پیش‌بینی کمتری نسبت به مدل بیزی دارد. علاوه بر نرخ مثبت درست و کاذب که برای تخمین عملکرد پیش‌بینی طبقه‌بندی کننده استفاده می‌شود، ضریب همبستگی متیو[3] (MCC) نیز برای ارزیابی بیشتر عملکرد مدل‌های یادگیری ماشینی مورد استفاده مقایسه شدند (شکل ۴).



شکل ۴- مقایسه میزان حساسیت (نرخ مثبت درست) و مثبت کاذب توسط دو مدل مختلف
Figure 4. Comparison of false positive and true positive rates for two distinct methods based on an independent training set

عملکرد دو الگوریتم مختلف با استفاده از شاخص ضریب همبستگی متیو نیز مقایسه شد (شکل ۵). ضریب همبستگی متیو، یک روش آماری قابل اعتمادتر است که تنها در صورتی امتیاز بالایی ایجاد می‌کند که پیش‌بینی نتایج خوبی در هر چهار دسته ماتریس خطا[4] (مثبت درست، منفی غلط، منفی درست و مثبت غلط) متناسب با اندازه عناصر مثبت و اندازه عناصر منفی در مجموعه داده به‌دست آورد. براساس داده‌های ورودی به دو مدل، بیشترین و کمترین مقدار ضریب همبستگی متیو ۰/۹۷۹ در الگوریتم جنگل تصادفی و ۰/۲۴۶ در الگوریتم بیزی بدست آمد. در نتیجه رویه جنگل تصادفی عملکرد بهتری از نظر ضریب همبستگی متیو نسبت به روش بیزی داشت. پن و همکاران (۲۰۱۹) ضریب همبستگی متیو را با استفاده از ماشین بردار پشتیبان ۰/۸۸ گزارش کردند (Pan et al., 2019).



شکل ۵- مقایسه ضریب همبستگی متیو برآورد شده در دو مدل مختلف
Figure 5. Matthew's correlation coefficient for cattle miRNAs using random forest and baysian classifier employing only dinucleotid-based features.

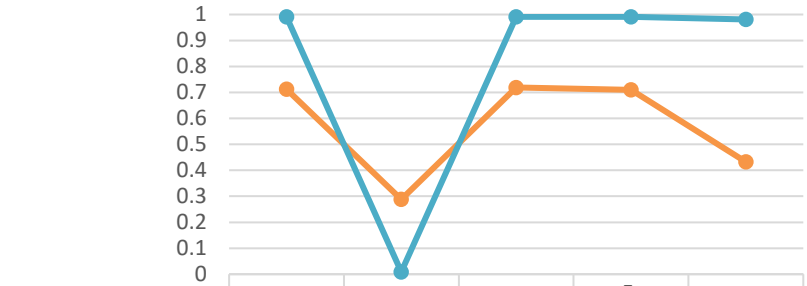
عملکرد الگوریتم جنگل تصادفی در پیش‌بینی

جنگل تصادفی یک الگوریتم یادگیری نظارت شده محسوب می‌شود. همانطور که از نام آن مشهود است، این الگوریتم جنگلی را به‌طور تصادفی می‌سازد. جنگل ساخته شده، در واقع گروهی از درخت‌های تصمیم (Decision Trees) است. کار ساخت جنگل با استفاده از درخت‌ها اغلب اوقات به‌روش کیسه‌گذاری (Bagging) انجام می‌شود. ایده اصلی روش کیسه‌گذاری آن است که ترکیبی از مدل‌های یادگیری، نتایج کلی مدل را بهبود می‌دهد. به بیان ساده جنگل تصادفی چندین درخت تصمیم ساخته و آن‌ها را با یکدیگر ادغام می‌کند تا پیش‌بینی‌های صحیح‌تر و پایدارتری حاصل شوند (Zhang et al., 2016). در این پروژه پارامترهای مدل جنگل تصادفی با حداکثر قعر ۰، تعداد درختان ۱۰ و سید ۱ و با اعتبارسنجی متقابل ده‌تایی تنظیم شدند.

1-Sensitivity (true positive rate)
2- False positive (FP) rate
3- Matthews's correlation coefficient (MCC)
4- Confusion matrix

اعتبارسنجی و ارزیابی عملکرد داده‌های تست

دو مدل جنگل تصادفی و بیزی در مجموعه داده‌های تست شامل ۲۱۲ توالی برای شناسایی miRNA های پیش‌ساز گاو اجرا شد. نتایج آنالیز داده‌های تست برای ویژگی‌های جفت نوکلئوتیدی در شکل ۶ نشان داده شده‌اند. با توجه به شکل ۶ می‌توان نتیجه گرفت که الگوریتم جنگل تصادفی از دقت بالایی در پیش‌بینی داده‌های تست برخوردار است. یکی از روش‌های شناسایی miRNAها (Tran et al., 2015) miRBoost است. این روش از مجموعه‌ای شامل ۱۸۷ ویژگی، ویژگی‌های مناسب را استخراج می‌کند و پس از آموزش داده‌ها با استفاده از تکنیک‌های بهبود مؤلفه SVM، طبقه‌بندی را انجام می‌دهد. روش دیگر (Batuwita and Palade, 2009) microPred متمایزترین ویژگی‌ها را انتخاب می‌کند، برای آموزش از طبقه‌بندی‌کننده SVM با استفاده از روش فیلترینگ عمل می‌کند، مشکل عدم تعادل کلاس در مجموعه داده را مدیریت می‌کند و از اعتبارسنجی متقاطع به‌منظور ارزیابی کارایی طبقه‌بندی استفاده می‌نماید. SVM مورد استفاده در این دو روش مدل طبقه‌بندی دودویی است که می‌تواند به‌عنوان طبقه‌بندی‌کننده خطی با بیشترین فاصله در فضای ویژگی تعریف شود. استراتژی یادگیری در این روش، فاصله را به حداکثر رسانده و در نهایت به راه‌حلی برای یک مسئله برنامه‌نویسی درجه دوم محدب تبدیل می‌کند. پارک و همکاران (۲۰۱۷) ساختار ثانویه را با توالی pre-miRNA ترکیب کردند، تا یک ماتریس ۱۶ بعدی را تشکیل دهند، سپس نتایج را برای بهبود مدل‌سازی وابستگی دوربرد[1] که برای آنالیز داده‌های مکانی یا سری زمانی ایجاد می‌شود، به RNN ارسال کردند. بزرگترین مزیت روش آنها این بود که به ویژگی‌های دست‌ساز نیاز نداشت (Park et al., 2017). دو و همکاران (Do et al, 2018) با استفاده از ساختار ثانویه کدگذاری شده توسط قالب ماتریس جفت شده به‌عنوان ورودی به شبکه دو بعدی کانولوشن برای دستیابی به استخراج خودکار ویژگی، یک روش چند کاناله دو بعدی مشترک جدید را برای شناسایی pre-miRNAها پیشنهاد کردند. این ویژگی‌ها توسط روش یادگیری عمیق به لایه‌های کاملاً متصل برای طبقه‌بندی وارد شدند (Do et al., 2018).



	TP Rate	FP Rate	Perecision	F-Measure	MCC
Naïve Bayes	0.712	0.288	0.719	0.71	0.432
Random Forest	0.991	0.009	0.991	0.991	0.981

شکل ۶- خروجی آنالیز داده‌های تست در miRNAهای گاو
Figure 6. Prediction results for test dataset in catlle miRNAs.

تشخیص بیوانفورماتیکی miRNAها به استفاده از نوع مدل‌های پیش‌بینی و ویژگی‌های مختص miRNA های پیش‌ساز متکی است. با این‌حال طول کوتاه ژن‌های miRNA و نبود ویژگی‌های مشخص برای این توالی‌ها این کار را پیچیده می‌کند (Lopes et al., 2016). علاوه‌بر این miRNAها در بسیاری از فرآیندهای زیستی مهم از جمله رشد، انتقال سیگنال و تخریب پروتئین نقش دارند (Pritchard et al., 2012). اخیراً معرفی مدل‌های پیش‌بینی miRNAهای پیش‌ساز به‌دلیل ارتباط دقیق فرایندهای زیستی با بیوژنز miRNA و طرح RNAهای مداخله‌گر کوتاه، به یک موضوع داغ در تحقیقات miRNA تبدیل شده است. روش‌های آزمایشگاهی سنتی تشخیصی برای شناسایی miRNAها مانند توالی‌یابی تراشه گران و وقت‌گیر هستند (Bentwich, 2005; Liao et al., 2014; Peng et al., 2018). در دوران پس از توالی‌یابی ژنوم، تعداد زیادی از توالی‌های ژنومی در دسترس قرار گرفته‌اند و فرصتی را برای شناسایی pre-miRNAها در مقیاس بزرگ توسط تکنیک‌های محاسباتی فراهم می‌کنند (Li et al., 2010).

در سال‌های اخیر روش‌های محاسباتی زیادی برای شناسایی pre-miRNAها پیشنهاد شده‌اند که بیشتر آنها بر اساس الگوریتم‌های یادگیری ماشین یا مدل‌های آماری هستند. روش‌های مبتنی بر یادگیری ماشینی معمولاً شناسایی pre-miRNA را به‌عنوان یک مشکل طبقه‌بندی دودویی مدل‌سازی می‌کنند، تا miRNAهای واقعی و توالی‌های مشابه pre-miRNA را تشخیص دهند. الگوریتم‌های مبتنی بر یادگیری ماشینی که به‌طور گسترده مورد استفاده قرار می‌گیرند، شامل ماشین‌های بردار پشتیبان (SVM) و نقشه خود سازمان‌دهی (SOM) شبکه‌های عصبی پس انتشار[2]،

1- long-term dependency modeling

2- Self-organizing map (SOM) neural networks

برنامه‌ریزی ژنتیکی خطی، مدل مارکوف پنهان، جنگل تصادفی، تمایز کوواریانس، بیز ساده (Lopes et al., 2014) و یادگیری عمیق هستند. یوسف و همکاران (Yousef et al., 2006) و پنگ و همکاران (Peng et al., 2018) از یک طبقه‌بندی کننده بیزی برای شناسایی pre-miRNA استفاده کردند که در شناسایی pre-miRNAها در ژنوم گونه‌های مختلف اثر بخش بود (Peng et al., 2018; Yousef et al., 2006). زو و همکاران (Xue et al, 2005) یک پیش‌بینی‌کننده سه‌گانه SVM را برای شناسایی ویژگی‌های ساختاری سنجاق سر miRNA پیش‌ساز پیشنهاد کردند که عملکرد پیش‌بینی آن در مقایسه با روشی که با استفاده از طبقه‌بندی‌کننده MiPred مبتنی بر جنگل تصادفی بود، تا ۱۰ درصد بهبود یافت (Xue et al., 2005). علاوه‌بر این استگمایر و همکاران (Stegmayer (۲۰۱۶) یک پیش‌بینی عمیق SOM را برای حل مشکل عدم تعادل نمونه‌های pre-miRNA مثبت و منفی پیشنهاد کردند (Stegmayer et al., 2016).

عملکرد روش‌های مبتنی بر یادگیری ماشینی به‌میزان زیادی به نوع ویژگی‌های استخراج شده مرتبط است (Ren et al., 2018; W. Zhang & Wang, 2017). ساختار ثانویه و روش‌های مبتنی بر اطلاعات توالی، ازجمله روش‌های معرفی ویژگی‌های متعارف هستند (Demirci and Allmer, 2017; Wei et al., 2017; Yousef et al., 2017). به‌عنوان مثال زو و همکاران (۲۰۰۵) یک ویژگی ۳۲ بعدی از توالی‌های سه‌تایی حاوی اطلاعات ساختار ثانویه را برای بیان بهتر توالی‌های pre-miRNA پیشنهاد کردند (Xue et al., 2005). جیانگ و همکاران (۲۰۰۷) برای توالی‌های تصادفی بازآرایی انجام دادند که در به‌دست آوردن مقادیر انرژی توالی‌های pre-miRNA مفید بود. با این حال روش آنها بسیار کند بود (Jiang et al., 2007). علاوه بر این وی و همکاران (۲۰۱۴) و چن و همکاران (۲۰۱۶) ویژگی‌های پیشنهاد شده توسط زو و همکاران (۲۰۰۵) را به ویژگی‌های ۹۸ بعدی pre-miRNA گسترش دادند که منجر به دقت پیش‌بینی بیشتری برای شناسایی pre-miRNA ها شد (Chen et al., 2016; Wei et al., 2014).

از آنجایی‌که اکثر pre-miRNA ها ساختار سنجاق‌سر ساقه–حلقه ویژه‌ای دارند (Xue et al., 2005)، بنابراین ساختار ثانویه یک ویژگی مهم مورد استفاده در روش‌های محاسباتی محسوب می‌شود. اخیراً لیو و همکاران چندین روش را برای پیش‌بینی pre-miRNA ها بر اساس ساختار ثانویه، به نام‌های iMiRNA-PseDPC، iMcRNA-PseSSC، miRNA-dis و deKmer پیشنهاد کرده‌اند (Fu et al., 2019). برخی از محققان (Khan et al., 2017; Yousef et al., 2017) به‌منظور بهبود دقت پیش‌بینی pre-miRNA ها ابعاد ویژگی‌ها را با ترکیب چند منبع ویژگی افزایش دادند. با افزایش ابعاد ماتریس ویژگی، اطلاعات اضافی و نویزهای قابل‌توجهی نیز گنجانده شده است که ممکن است دقت پیش‌بینی را کاهش داده و الگوریتم را کند سازد. بنابراین معمولاً لازم است برای حذف ویژگی‌های غیرمرتبط یا اضافی، انتخاب ویژگی انجام شود. یک روش انتخاب ویژگی مناسب می‌تواند به‌طور مؤثر زمان اجرای آموزش مدل را کاهش دهد و عملکرد پیش‌بینی را بهبود بخشد (Wang et al., 2011). چندین ابزار بیوانفورماتیک به‌منظور تسهیل اغلب فرآیندهای محاسباتی برای تولید اطلاعات ویژگی‌های توالی‌های عددی توسعه داده شده است (Liu et al., 2015).

توسعه الگوریتمی که بتواند ویژگی مؤثر برای توالی‌های pre-miRNA را معرفی می‌کند، کاری چالش‌برانگیز است. روش‌های موجود مشکلات متعددی دارند و ممکن است برای تمایز بین pre-miRNA و توالی‌های غیر از pre-miRNA به‌اندازه کافی آموزنده نباشد. روش‌های استخراج ویژگی اغلب معمولاً فقط فراوانی یا اطلاعات توالی بازهای pre-miRNA را در نظر می‌گیرند که رویکرد پروژه حاضر در شناسایی محاسباتی pre-miRNAها بود. لازم به ذکر است که روش‌های استخراج ویژگی مبتنی بر ساختارهای ثانویه معمولاً فقط ویژگی‌های کلی را در نظر می‌گیرند و روش‌های ترکیب اطلاعات ویژگی چند منبعی و ادغام الگوریتم‌های انتخاب ویژگی برای کاهش ابعاد ماتریس ویژگی‌ها از نظر زمان محاسباتی ناکارآمد هستند (Khan et al., 2017; Yousef et al., 2017).

در سال ۲۰۰۵، چنگهای و همکاران روشی را برای طبقه‌بندی mRNA های واقعی و mRNA های کاذب توسط SVM با استفاده از ویژگی‌های ساختار توالی محلی تعریف کردند. آنها به دقت ۹۰ درصد در داده‌های انسانی دست یافتند. به طور مشابه رحمان و همکاران (۲۰۱۲) یک طبقه‌بند چندلایه شبکه عصبی مصنوعی را با آموزش ۱۷ پارامتر برای پیش‌بینی pre-miRNA واقعی از miRNAهای کاذب پیشنهاد کردند و به طور میانگین حساسیت ۹۷/۴۰ درصد و اختصاصیت ۹۵/۸۵ درصد را به دست آوردند. این روش همچنین با چهار روش طبقه‌بندی پیشرفته دیگر به‌نام‌های MiPred، miRabela، microPred و Triplet-SVM مقایسه شد. مشابه با این تحقیقات، طبقه‌بندی‌کننده جدیدی توسط یو و همکاران (۲۰۱۵) برای پیش‌بینی miRNAهای تنظیمی توسعه یافت. در این مطالعه، آنها نشان دادند که روش‌های پیشرفته برای تعیین pre-miRNAها کافی هستند. با این‌حال محققان سیستمی را برای بهبود دقت کارایی شناسایی pre-miRNA ایجاد کردند که می‌تواند ویژگی‌های ساختار ثانویه–حلقه و چند ساقه را با استفاده از شبکه‌های عصبی به‌کار برد. مجموعه داده واقعی pre-miRNA برای ساخت موفقیت‌آمیز این طبقه‌بندی کننده جدید برای مدیریت مشکلات عدم تعادل کلاس استفاده شد. روش اعتبارسنجی متقابل پنج‌تایی نیز برای ارزیابی عملکرد طبقه‌بندی‌کننده پیشنهاد شده استفاده شد. زو و همکاران (۲۰۰۵) از یک رویکرد ترکیبی برای پیش‌بینی ژن‌های RNA غیرکدکننده کوتاه استفاده کردند. در مدل رحمان و همکاران (۲۰۱۲) یک رویکرد یادگیری ماشینی شبکه عصبی مصنوعی نظارت شده برای پیش‌بینی miRNAهای جدید که به‌عنوان pre-miRNAها شناخته می‌شود، با استفاده از مجموعه‌ای از نواحی توالی‌های کدکننده انسانی (CDS) توسعه یافت. در این تحقیق نتایج به‌دست‌آمده آنها با ۹۹/۹ درصد صحت (ACC)، ۹۹/۸ درصد حساسیت (SN) و ۱۰۰

درصد اختصاصیت (SP) پیش‌بینی قابل اعتمادتری ارائه کرد (Abbas et al., 2016).

روش ما با چهار روش یادگیری ماشینی دیگر که در این بخش بحث شده است (جدول ۱)، مقایسه شد. نمونه‌های مثبت این پروژه از پایگاه miRbase ویرایش ۲۲ به‌دست آمد و نمونه‌های منفی از NCBI (http://www.ncbi.nlm.nih. gov)، RNAcentral (https://rnacentral.org) و snoRNA-LBME-db (http://www-snorna.biotoul.fr/) گرفته شد. نمونه‌های منفی عمدتاً از مناطق بیرونی ژن‌های کدکننده پروتئین‌ها و RNA های غیرکدکننده‌ای که miRNA نبودند، مانند tRNA، siRNA، snRNA و snoRNA تشکیل شده‌اند. به‌منظور بهبود کیفیت داده‌ها و جلوگیری از بیش برازش، عناصر با حاشیه‌نویسی اشتباه و توالی‌های تکراری حذف شدند. برای مجموعه داده پروژه حاضر، ۲۰ درصد داده‌ها به‌عنوان مجموعه داده تست و ۸۰ درصد باقی‌مانده را برای اجرای اعتبارسنجی متقاطع ده‌تایی برای آموزش و انتخاب مدل استفاده شد. با استفاده از دو مدل آموزش داده شده با مجموعه داده‌های گاوی و ۱۷ ویژگی دو نوکلئوتیدی پیش‌بینی انجام شد.

قدرت پیش‌بینی مدل جدید پیشنهادی

عملکرد الگوریتم‌های یادگیری ماشینی معمولاً به وظیفه‌ای که برای اجرای آن به‌کار برده می‌شوند، بستگی دارد. الگوریتم‌های مختلف می‌توانند از ویژگی‌ها و روابط مختلف در یک مجموعه داده معین بهره ببرند. در این پروژه دو طبقه‌بندی‌کننده بر اساس ۱۷ ویژگی دو نوکلئوتیدی ساخته شد و عملکرد آنها با استفاده از اعتبارسنجی متقاطع ۱۰ تایی ارزیابی شد. روش جنگل تصادفی از تکنیک بیزی برای پیش‌بینی miRNAها بهتر عمل کرد. عملکرد بهتر جنگل تصادفی را می‌توان به‌دلیل قدرت کیسه‌بندی تکنیک بوت استرپ[1] و فرآیند انتخاب ویژگی تصادفی به‌منظور ساخت مجموعه مدل درخت تصمیم‌گیری دانست. داده‌های این پروژه شامل توالی‌های miRNAهای گاو بود و الگوریتم جنگل تصادفی نسبت به روش بیزی مناسب‌تر بود، زیرا ناهمگونی زیرخانواده‌های RNAهای غیر کدکننده توسط مجموعه‌ای از درختان شناسایی شد. به‌منظور شناسایی miRNAها استفاده از ویژگی‌های مختلف می‌تواند مفید باشند. بنابراین با استفاده از انواع مختلف داده‌ها می‌توان حساسیت روش‌های محاسباتی را برای شناسایی miRNA بهبود بخشید (Lu et al., 2011).

همانطور که مشاهده می‌شود، تاکنون روش‌های متفاوتی به‌منظور شناسایی ژن‌های miRNA در گونه‌های مختلف ارائه شده است (Das et al., 2018) (جدول ۱). در بین این روش‌ها بیشترین مدلی که برای شناسایی ژن‌های miRNA تاکنون مورد استفاده قرار گرفته است، روش ماشین بردار پشتیبان است. با این‌حال نتایج این پژوهش بر روی miRNAهای شناخته شده ژنوم گاو رویه جنگل تصادفی را به‌دلیل بالاترین مقادیر دقت و صحت نسبت به سایر روش‌ها پیشنهاد می‌دهد. رویه به‌کار گرفته شده در این پژوهش دقت ۹۹ درصد و ضریب همبستگی متیو ۹۸ درصد را برای داده‌های آزمون با استفاده از ویژگی‌های استخراج شده دو نوکلئوتیدی نشان داد.

جدول ۱- مقایسه توانایی پیش‌بینی فن‌آوری‌های هوشمند بیوانفورماتیکی موجود با الگوریتم پیشنهاد شده در پروژه حاضر

Table 1. Comparing the proposed method with other state-of-the-art predictors on an independent dataset

روش طبقه‌بندی Classification methods	حساسیت Sensitivity	اختصاصیت specificity	صحت Acuracy
iMcRNA-PseSSC	0.896	0.832	0.864
miRNA-deKmer	0.732	0.742	0.737
miRNA-dis	0.958	0.804	0.881
iMiRNA-PseDPC	0.954	0.828	0.891
PremipreD	0.890	0.922	0.906

نتیجه‌گیری کلی

برای شناسایی تنوع در خانواده‌های miRNA، ویژگی‌های مبتنی بر توالی به‌طور گسترده برای شناسایی miRNAها استفاده می‌شوند، از جمله ویژگی‌هایی مانند kmer (یک ویژگی مبتنی بر توالی). این ویژگی‌های متداول به‌اندازه کافی مناسب هستند تا بتوانند miRNAها را شناسایی کنند. تاکنون ویژگی‌های ترکیب نوکلئوتیدی (kmer)، با موفقیت برای پیش‌بینی ژن‌های miRNA مورد استفاده قرار گرفتند (به‌عنوان مثال در smyRNA و ncRNAscout). یک سیگنال آماری معنی‌دار بر اساس محتوای GC نیز به‌تنهایی می‌تواند برای تشخیص miRNAها استفاده شود. باید در نظر داشت که ویژگی‌های براساس توالی در گونه‌های مختلف متفاوت هستند و سیگنال‌های مرتبط با آن‌ها عموماً برای شناسایی ژن miRNA در بیشتر ارگانیسم‌ها کافی نیستند و بسته به نوع گونه ممکن است نتایج متفاوتی را شاهد باشیم.

حقوق مؤلف

مقاله مذکور مستخرج از پروژه "آنالیز بیوانفورماتیکی پیش‌سازهای miRNA در گاو (*Bos Taurus*)" با شماره مصوب: ۹۶۰۳۱۲-۹۶۰۰۷-۹۶۰۱-۰۴۵-۱۳۵۷-۱۳-۱۳۴۸ و شماره فروست ۶۳۴۵۶ مؤسسه تحقیقات علوم دامی کشور است.

منابع

Abbas, Q., Raza, S., Biyabani, A., & Jaffar, M. (2016). A review of computational methods for finding non-coding RNA genes. *Genes*, *7*(12), 113. https://doi.org/10.3390/genes7120113

Akhtar, M. M., Micolucci, L., Islam, M. S., Olivieri, F., & Procopio, A. D. (2016). Bioinformatic tools for microRNA Finding Non-Coding RNA Genes. *Nucleic Acids Research*, *44*(1), 24–44.

1-Bagging

https://doi.org/10.1093/nar/gkv1221
Barazandeh, A., Mohammadabadi, M., Ghaderi-Zefrehei, M., Rafeie, F., & Imumorin, I. G. (2019). Whole genome comparative analysis of CpG islands in camelid and other mammalian genomes. *Mammalian Biology*, *98*, 73–79.
Barazandeh, A., Mohammadabadi, M. R., Ghaderi-Zefrehei, M., & Nezamabadipour, H. (2016). Predicting CpG islands and their relationship with genomic feature in cattle by hidden markov model algorithm. *Iranian Journal of Applied Animal Science*, *6*(3), 571–579.
Barazandeh, A., Mohammadabadi, M. R., Ghaderi-Zefrehei, M., & Nezamabadi-Pour, H. (2016). Genome-wide analysis of CpG islands in some livestock genomes and their relationship with genomic features. *Czech Journal of Animal Science*, *61*(11), 487–495.
Batuwita, R., & Palade, V. (2009). microPred: Effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, *25*(8), 989–995. https://doi.org/10.1093/bioinformatics/btp107
Bentwich, I. (2005). Prediction and validation of microRNAs and their targets. *FEBS Letters*, *579*(26), 5904–5910.
Bordbar, F., Mohammadabadi, M., Jensen, J., Xu, L., Li, J., & Zhang, L. (2022). Identification of candidate genes regulating carcass depth and hind leg circumference in simmental beef cattle using Illumina Bovine Beadchip and next-generation sequencing analyses. *Animals*, *12*(9), 1103.
Castel, S. E., & Martienssen, R. a. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature Reviews. Genetics*, *14*(2), 100–112. https://doi.org/10.1038/nrg3355
Chen, D., Du, Y., Chen, H., Fan, Y., Fan, X., Zhu, Z., Wang, J. J., Xiong, C., Zheng, Y., Hou, C., Parveen, A., Mustafa, S. H., Yadav, P., Kumar, A., Wang, H., Ma, Y., Dong, C., Li, C., Wang, J. J., & Ciaudo, C. (2019). Applications of Machine Learning in miRNA Discovery and Target Prediction. *Nucleic Acids Research*, *10*(1), 414656. https://doi.org/10.1186/s12967-019-2009-x
Chen, J., Wang, X., & Liu, B. (2016). iMiRNA-SSF: Improving the Identification of MicroRNA Precursors by Combining Negative Sets with Different Distributions. *Scientific Reports*, *6*(October 2015), 19062. https://doi.org/10.1038/srep19062
Das, S. G., Chakraborty, H. J., & Datta, A. (2018). PremipreD: precursor miRNA prediction by support vector machine approach. *Trends in Bioinformatics*, *11*(1), 17–24.
de ON Lopes, I., Schliep, A., & de LF de Carvalho, A. C. P. (2014). The discriminant power of RNA features for pre-miRNA recognition. *BMC Bioinformatics*, *15*(1), 1–11.
de Sousa, M. A. P., de Athayde, F. R. F., Maldonado, M. B. C., de Lima, A. O., Fortes, M. R. S., & Lopes, F. L. (2021). Single nucleotide polymorphisms affect miRNA target prediction in bovine. *PLoS ONE*, *16*(4 April), 1–17. https://doi.org/10.1371/journal.pone.0249406
De Souza, E. B., Cload, S. T., Pendergrast, P. S., & Sah, D. W. Y. (2009). Novel therapeutic modalities to address nondrugable protein interaction targets. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *34*(1), 142–158. https://doi.org/10.1038/npp.2008.115
Demirci, M. D. S., & Allmer, J. (2017). Delineating the impact of machine learning elements in pre-microRNA detection. *PeerJ*, *5*, e3131.
Do, B. T., Golkov, V., Gürel, G. E., & Cremers, D. (2018). Precursor microRNA Identification Using Deep Convolutional Neural Networks. *bioRxiv*, 414656. https://doi.org/10.1101/414656
Do, D. N., Dudemaine, P.-L., Mathur, M., Suravajhala, P., Zhao, X., & Ibeagha-Awemu, E. M. (2021). MiRNA Regulatory Functions in Farm Animal Diseases, and Biomarker Potentials for Effective Therapies. *International Journal of Molecular Sciences*, *22*(6), 3080.
Douglass, S., Hsu, S. W., Cokus, S., Goldberg, R. B., Harada, J. J., & Pellegrini, M. (2016). A naïve Bayesian classifier for identifying plant microRNAs. In *The Plant journal : for cell and molecular biology* (Vol. 86, Issue 6, pp. 481–492). https://doi.org/10.1111/tpj.13180
Fan, D., Yao, Y., & Yi, M. (2021). Plantmirp2: An accurate, fast and easy-to-use program for plant pre-mirna and mirna prediction. In *Genes* (Vol. 12, Issue 8). https://doi.org/10.3390/genes12081280
Fu, X., Zhu, W., Cai, L., Liao, B., Peng, L., Chen, Y., & Yang, J. (2019). Improved pre-miRNAs identification through mutual information of pre-miRNA sequences and structures. *Frontiers in Genetics*, *10*(FEB), 1–12. https://doi.org/10.3389/fgene.2019.00119
Ghotbaldini, H., Mohammadabadi, M., Nezamabadi-pour, H., Babenko, O. I., Bushtruk, M. V., & Tkachenko, S. V. (2019). Predicting breeding value of body weight at 6-month age using Artificial Neural Networks in Kermani sheep breed. *Acta Scientiarum. Animal Sciences*, *41*.
Ghotbaldini H.R., Mohammadabadi M.R., & Nezamabadi Pour, H. (2017). Application of artificial intelligence for estimating breeding value of body weight in birth and 3 months age in Kermani sheep breed. *Journal of Modern Genetics*, *12*(3), 323–331.
Huang, Y., Zou, Q., Ren, H. T., & Sun, X. H. (2015). Prediction and characterization of microRNAs from eleven fish species by computational methods. *Saudi Journal of Biological Sciences*, *22*(4), 374–381. https://doi.org/10.1016/j.sjbs.2014.10.005
Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., & Lu, Z. (2007). MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids*

Research, *35*(SUPPL.2), W339–W344. https://doi.org/10.1093/nar/gkm368
Khan, A., Shah, S., Wahid, F., Khan, F. G., & Jabeen, S. (2017). Identification of microRNA precursors using reduced and hybrid features. *Molecular Biosystems*, *13*(8), 1640–1645.
Lertampaiporn, S., Thammarongtham, C., Nukoolkit, C., Kaewkamnerdpong, B., & Ruengjitchatchawalya, M. (2014). Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm. *Nucleic Acids Research*, *42*(11), e93–e93.
Li, L., Xu, J., Yang, D., Tan, X., & Wang, H. (2010). Computational approaches for microRNA studies: a review. *Mammalian Genome*, *21*(1-2), 1–12.
Liao, B., Jiang, Y., Liang, W., Zhu, W., Cai, L., & Cao, Z. (2014). Gene selection using locality sensitive Laplacian score. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(6), 1146–1156.
Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., & Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research*, *43*(W1), W65–W71.
Lu, Z. J., Yip, K. Y., Wang, G., Shou, C., Hillier, L. W., Khurana, E., Agarwal, A., Auerbach, R., Rozowsky, J., & Cheng, C. (2011). Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Research*, *21*(2), 276–285.
Magyar, L. (2018). A Review of the Utility of Bayesian Network Models. *The University of Akron. Honors Research Projects. 689.*
Moradian, H., Esmailizadeh Koshkoiyeh, A., Mohammadabadi, M., & Asadi Fozi, M. (2020). Whole genome detection of recent selection signatures in Sarabi cattle: a unique Iranian taurine breed. *Genes & Genomics*, *42*, 203–215.
ON Lopes, I. de, Schliep, A., & de LF de Carvalho, A. P. (2016). Automatic learning of pre-miRNAs from different species. *BMC Bioinformatics*, *17*(1), 1–18.
Pan, X., Chen, L., Feng, K. Y., Hu, X. H., Zhang, Y.-H., Kong, X. Y., Huang, T., & Cai, Y. D. (2019). Analysis of expression pattern of snoRNAs in different cancer types with machine learning algorithms. *International Journal of Molecular Sciences*, *20*(9), 2185.
Park, S., Min, S., Choi, H. S., & Yoon, S. (2017). Deep recurrent neural network-based identification of precursor micrornas. *Advances in Neural Information Processing Systems*, *30*.
Peng, L., Peng, M., Liao, B., Huang, G., Li, W., & Xie, D. (2018). The advances and challenges of deep learning application in biological big data processing. *Current Bioinformatics*, *13*(4), 352–359.
Pour Hamidi, S., Mohammadabadi, M. R., Asadi Foozi, M., & Nezamabadi-Pour, H. (2017). Prediction of breeding values for the milk production trait in Iranian Holstein cows applying artificial neural networks. *Journal of Livestock Science and Technologies*, *5*(2), 53–61.
Pritchard, C. C., Cheng, H. H., & Tewari, M. (2012). MicroRNA profiling: approaches and considerations. *Nature Reviews Genetics*, *13*(5), 358–369.
Ren, S., Liao, B., Zhu, W., Li, Z., Liu, W., & Li, K. (2018). The gradual resampling ensemble for mining imbalanced data streams with concept drift. *Neurocomputing*, *286*, 150–166.
Stegmayer, G., Yones, C., Kamenetzky, L., & Milone, D. H. (2016). High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *14*(6), 1316–1326.
Tran, V. D. T., Tempel, S., Zerath, B., Zehraoui, F., & Tahi, F. (2015). miRBoost: boosting support vector machines for microRNA precursor classification. *RNA (New York, N.Y.)*, *21*(5), 775–785. https://doi.org/10.1261/rna.043612.113
Tzelos, T., Ho, W., Charmana, V. I., Lee, S., & Donadeu, F. X. (2022). MiRNAs in milk can be used towards early prediction of mammary gland inflammation in cattle. *Scientific Reports*, *12*(1), 1–8. https://doi.org/10.1038/s41598-022-09214-9
Wang, H., Ma, Y., Dong, C., Li, C., Wang, J., & Liu, D. (2019). CL-PMI: A Precursor MicroRNA Identification Method Based on Convolutional and Long Short-Term Memory Networks. *Frontiers in Genetics*, *10*(October), 1–13. https://doi.org/10.3389/fgene.2019.00967
Wang, Y., Chen, X., Jiang, W., Li, L., Li, W., Yang, L., Liao, M., Lian, B., Lv, Y., Wang, S., Wang, S., & Li, X. (2011). Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM. *Genomics*, *98*(2), 73–78. https://doi.org/10.1016/j.ygeno.2011.04.011
Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., & Zou, Q. (2014). Improved and promising identificationof human microRNAs by incorporatinga high-quality negative set. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(1), 192–201. https://doi.org/10.1109/TCBB.2013.146
Wei, L., Tang, J., & Zou, Q. (2017). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Information Sciences*, *384*, 135–144.
Xue, C., Li, F., He, T., Liu, G.-P., Li, Y., & Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, *6*, 310. https://doi.org/10.1186/1471-2105-6-310
Yao, D., Zhan, X., & Kwoh, C. K. (2019). An improved random forest-based computational model for predicting novel miRNA-disease associations. In *BMC Bioinformatics*, 20(1).

https://doi.org/10.1186/s12859-019-3290-7

Yousef, M., Khalifa, W., Acar, İ. E., & Allmer, J. (2017). MicroRNA categorization using sequence motifs and k-mers. *BMC Bioinformatics*, *18*, 1–9.

Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L. C., & Showe, M. K. (2006). Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, *22*(11), 1325–1334.

Zhang, J., Hadj-Moussa, H., & Storey, K. B. (2016). Current progress of high-throughput microRNA differential expression analysis and random forest gene selection for model and non-model systems: an R implementation. *Journal of Integrative Bioinformatics*, *13*(5), 306.

Zhang, W., & Wang, S. L. (2017). An integrated framework for identifying mutated driver pathway and cancer progression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(2), 455–464.