https://eventos.utfpr.edu.br//sicite/sicite2020

Inteligência artificial aplicado em dados de non-coding RNAs.

Artificial intelligence applied to data from non-coding RNAs.

Thaysla Fernanda Gomes da Cruz
thayslacruz@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil.

Alexandre Rossi Paschoal
paschoal@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil.

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

RESUMO

Os miRNAs são pequenos non-coding RNAs (ncRNAs) que possuem um papel fundamental na regulação dos genes, o qual diversos processos biológicos nas células animais/vegetais estão sob controle do micro RNA (miRNA). Para que o miRNA se torne maduro e apto a desempenhar seu papel regulatório, sua biogênese canônica é constituída de duas clivagens. Alguns estudos em organismos modelos identificaram uma subclasse de miRNAs, denominada Mirtrons, originadas de uma biogênese de via não canônica.

Os Mirtrons também participam de diversos processos regulatórios e são potenciais silenciadores de doenças. Apesar das semelhanças entre mirtrons e miRNAs canônicos, uma comparação entre suas diferenças estruturais permitem a compreensão quanto a forma como os processos biológicos são regulados. Além disso, devido as diferenças entre essas classes, preditores de miRNAS não são hábeis a realizar a predição de Mirtrons.

Nesse projeto foi realizada uma coleta de dados de mirtrons e microRNAs do mirtronDB e miRBase, bem como a extração de características identificadas pelo estado da arte como distintas entre as classes, a redução de um grande número de características pelo método SFS e por fim o treinamento de um classificador de aprendizado supervisionado denominado Random Forest afim de realizar a distinção das classes de forma automatizada.

PALAVRAS-CHAVE: Aprendizado de máquina. Bioinformática. Biologia molecular. Inteligência artificial.

ABSTRACT

microRNAs is short non-coding RNAs (ncRNAs) and one of the most explored classes of ncRNA as it acts in the cellular control of several biological processes. In order to become mature and qualified to execute their function, the miRNA perform two cleavages in their canonical biogenesis. Some studies in model organisms identified a miRNA subclass, called Mirtrons, originated from a non-canonical biogenesis pathway that omits Drosha cleavage. Thus, the identification of the differences between these two classes is important not only for mirtron prediction algorithms design but to improve the understanding of mirtron functioning too. In this project we developed a computational model dedicated to mirtron prediction and distinction of canonical miRNAs through the extraction and exploration of features identified by the state of art as distinct between these two classes.

KEYWORDS: Machine learning. Bioinformatics. Molecular biology. Artificial intelligence.

INTRODUÇÃO

Nos últimos anos, as pessoas têm prestado cada vez mais atenção ao papel do RNA não codificante (ncRNAs) como reguladores da expressão gênica. Antes vistos como lixo transcricional, dada a falta de tradução em proteínas, os RNAs não-codificantes são hoje conhecidos por desempenhar relevantes papéis biológicos. A classe dos microRNAs é uma das classes de ncRNAs mais exploradas, pois atua no controle celular de diversos processos biológicos (RIELLA, 2019).

Os miRNAs são ncRNAs de aproximadamente 22 nt (o maduro) que possuem um papel fundamental na regulação dos genes, o qual diversos processos biológicos nas células animais/vegetais estão sob controle do miRNA. Para que o miRNA se torne maduro e apto a desempenhar seu papel regulatório, sua biogênese canônica é constituída de duas clivagens. A desregulação ou disfunção dos miRNAs têm sido fortemente ligadas a aberrações no desenvolvimento, anomalias fisiológicas e comportamentais e câncer (FONSECA, DOMINGUES, PASCHOAL, 2019).

Mirtrons são uma subclasse de miRNAs originados de uma biogênese de via não-canônica, substituindo a primeira etapa de clivagem de Drosha por um mecanismo de *splicing* e posteriormente dando sequência ao processo pela via canônica. Os mirtrons também possuem função regulatória (FONSECA, DOMINGUES, PASCHOAL, 2019). Apesar das semelhanças entre mirtrons e miRNAs canônicos, uma comparação entre suas diferenças estruturais permitem a compreensão quanto a forma como os processos biológicos são regulados e fatores causadores de doenças.

Nesse projeto foi realizada uma coleta de dados de mirtrons e microRNAs, extração de características identificadas pelo estado da arte como distintas entre as classes e por fim o treinamento de um classificador de aprendizado supervisionado que realiza a distinção das classes de forma automatizada. Em suma, foi feito toda uma análise de dados via técnicas de aprendizado de máquina, bem como a análise de *features* em mirtrons.

MATERIAL E MÉTODOS

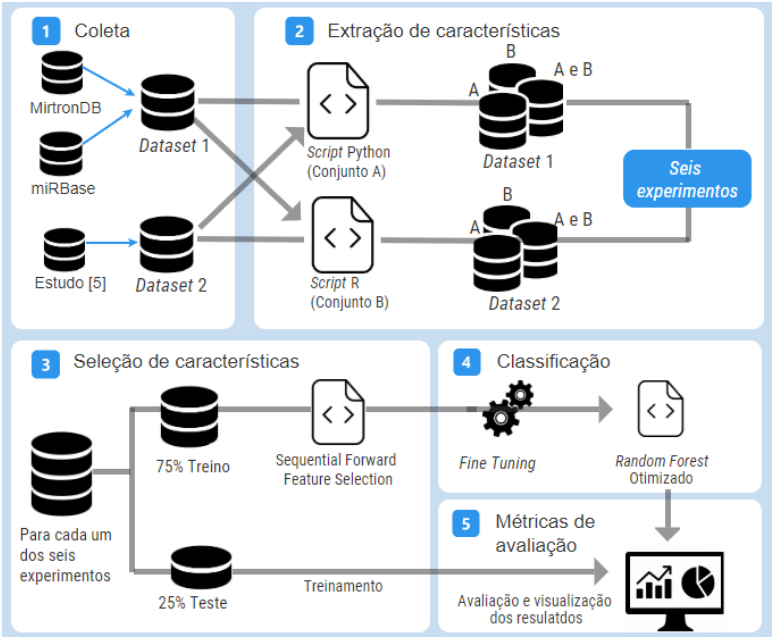
Para implementação do algoritmo e classificação das sequências de miRNA há uma série de processos anteriores que devem ser executados, dentre eles a coleta e tratamento dos dados, a extração e seleção de características e por fim um *Fine Tuning* (ajuste fino) ao classificador para que seja aplicado de forma otimizada. No fluxograma da Figura 1 estão representadas de forma geral cada uma dessas etapas, que serão explicadas de forma detalhada na sequência.

Os experimentos foram realizados utilizando dois *datasets,* o primeiro deles (*Dataset* 1) é composto por 455 mirtrons de humanos fornecidos em um arquivo fasta pelo repositório MirtronDB e 455 miRNAs obtidos do miRBase (versão 22.1), totalizando 910 amostras. Apesar do miRBase disponibilizar 1917 miRNAs de humanos, através de um *script* em *python* foram selecionadas 455 amostras de forma aleatória, para que fosse utilizado um *dataset* balanceado.

O *Dataset* 2 foi extraído do estudo de Rorbach et al. (2018), que consiste em 216 mirtrons e 707 miRNAs canônicos advindos do miRBase (versão 21) e um

conjunto de mirtrons putativo, com 201 novos locis de mirtron anotados no estudo de Wen et al. Totalizando 1124 amostras, estas foram disponibilizadas em dois arquivos no formato CSV e então convertidas para arquivos fasta através de um *script* em *python*.

Figura 1 – Etapas do projeto.



Fonte: Autoria própria (2020).

Foram definidos dois conjuntos de características para descreverem as sequências coletadas. O conjunto A possui 5464 características, sendo elas: Comprimento da sequência, GC *ratio*, GC *content*, Energia mínima livre (MFE) e k-mer de 1 a 6 nt. Em um script desenvolvido em python e através da biblioteca *Bio.seq* disponibilizada pelo *Scikit Learn* as sequências foram obtidas do arquivo fasta e foram extraídas as características de cada uma delas. A MFE foi extraída utilizando a aplicação RNA Fold disponibilizada pelo pacote ViennaRNA.

O conjunto B de características se trata das mesmas características utilizadas no estudo de Rorbach et al. (2018). São 25 características que descrevem uma sequência de forma completa e também dividida em três regiões: *interarm*, *mature* 3p e *mature* 5p. A extração dessas 25 características foi realizada através de um *script* em R desenvolvido pelos autores do estudo e disponibilizado no site https://github.com/ror94/Mirtrons.

No total ambos os conjuntos resultam em 5483 características, uma vez que a MFE, comprimento da sequência e *Hairpin* A, C G e U estavam presente em ambos. Os dois conjuntos de características foram analisadas nesse projeto de forma conjunta e separada.

Considerando a disposição de dois *datasets* e dois conjuntos de características distintos, a análise e classificação de cada um dos *datasets* foi realizada de forma a utilizar cada conjunto de características de forma separada e por fim em conjunto. Resultando em seis diferentes abordagens, descritas no Quadro 1.

Quadro 1 – Seis diferentes abordagens de classificação. Sendo o Dataset 1 e o Conjunto A extraídos nesse projeto e Dataset 2 e Conjunto B extraídos do estudo de Rorbach et al, (2018).

Abordagens	Dataset	Conjunto de características
1	1	A
2	1	B
3	1	A e B
4	2	A
5	2	B
6	2	A e B

Fonte: Autoria própria (2019).

Para cada um dos 6 experimentos foi utilizado um algoritmo de seleção sequencial de características denominado *Sequential Forward Selection* para a seleção de 15 características consideradas como mais relevantes para a classificação daquela abordagem. Com isso, há uma melhoria na eficiência computacional e uma redução no erro de generalização do classificador através da remoção de características ou ruídos irrelevantes.

Na etapa de classificação os *datasets* foram divididos entre conjunto de treino e teste, numa proporção de 75% das amostras sendo utilizadas na etapa de treino e os 25% restantes para teste. Sendo assim, no *Dataset* 1 foram utilizadas 682 amostras pra treino e 228 pra teste e no *Dataset* 2, 843 pra treino e 281 pra teste.

A classificação foi realizada através do algoritmo classificador de aprendizado supervisionado *Random Forest*. Trata-se de um algoritmo que realiza seleções aleatórias de subconjuntos de características e com elas monta mini árvores de decisão, substituindo a necessidade de uma grande árvore que abranja todas as características de uma só vez. Para obtenção do resultado ele realiza uma votação baseada nos valores retornados por cada uma das árvores.

Além disso, o algoritmo utiliza determinados parâmetros que permitem a customização da classificação para um dado conjunto de dados ou situação problema, então foi realizado o *Fine Tuning*, que por meio da realização de diversos testes com os possíveis parâmetros define uma solução ótima. Para a validação da classificação dos dados foram utilizadas 5 métricas de avaliação, são elas: acurácia, precisão, Recall, F1-Score e especificidade.

RESULTADOS E DISCUSSÃO

Para cada um dos seis experimentos especificados anteriormente, foi realizada a seleção de 15 características identificadas pelo algoritmo *Sequential Step Forward* como mais relevantes e posteriormente a classificação. Na Tabela 1 estão os resultados do treino e teste da classificação de ambos os *datasets* considerando apenas as 15 características selecionadas e na Tabela 2 estão os resultados obtidos quando utilizado os conjuntos de características completos.

Se comparados os resultados do classificador utilizando os conjuntos de características completos com o uso de apenas 15 características pode-se observar que o desempenho foi tão bom quanto ou até mesmo melhor, sem que haja a necessidade de lidar com uma quantidade tão grande de características. Através dos gráficos nas Figuras 2 e 3 é possível visualizar uma comparação entre os

resultados do classificador para o *Dataset* 1 e 2, quando consideradas somente as 15 características selecionadas.

Tabela 1 – Métricas: Etapa de treino e teste para as 15 características selecionadas

Etapa	Dataset	Conjunto	Acurácia	Precisão	Recall	F1 Score	Especificidade
Treino	1	A	0.735	0.802	0.608	0.692	0.856
		B	0.845	0.935	0.734	0.822	0.951
		A e B	0.745	0.808	0.629	0.707	0.856
	2	A	0.877	0.871	0.974	0.920	0.626
		B	0.987	0.992	0.987	0.989	0.987
		A e B	0.984	0.987	0.991	0.989	0.966
Teste	1	A	0.732	0.819	0.636	0.716	0.841
		B	0.759	0.875	0.636	0.806	0.925
		A e B	0.737	0.796	0.678	0.732	0.804
	2	A	0.816	0.824	0.949	0.883	0.455
		B	0.950	0.972	0.951	0.961	0.949
		A e B	0.902	0.923	0.944	0.933	0.788

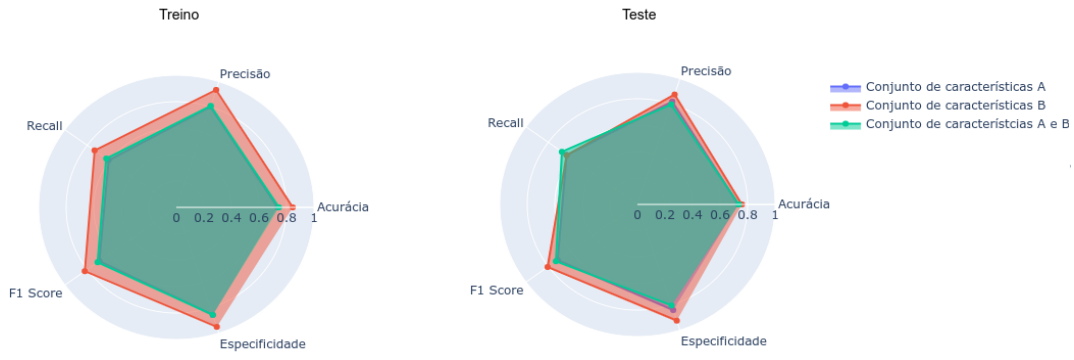
Fonte: Autoria própria (2020).

Tabela 2 – Métricas: Etapa de treino e teste para todas as características.

Etapa	Dataset	Conjunto	Acurácia	Precisão	Recall	F1 Score	Especificidade
Treino	1	A	0.658	0.733	0.746	0.577	0.833
		B	0.845	0.952	0.719	0.819	0.966
		A e B	0.811	0.896	0.695	0.782	0.922
	2	A	0.799	0.783	1.000	0.878	0.276
		B	0.989	0.922	0.990	0.991	0.987
		A e B	0.817	0.798	1.000	0.888	0.340
Teste	1	A	0.632	0.740	0.471	0.576	0.813
		B	0.771	0.879	0.661	0.755	0.897
		A e B	0.728	0.817	0.628	0.710	0.841
	2	A	0.779	0.770	0.994	0.868	0.197
		B	0.936	0.961	0.940	0.950	0.929
		A e B	0.803	0.788	1.000	0.881	0.273

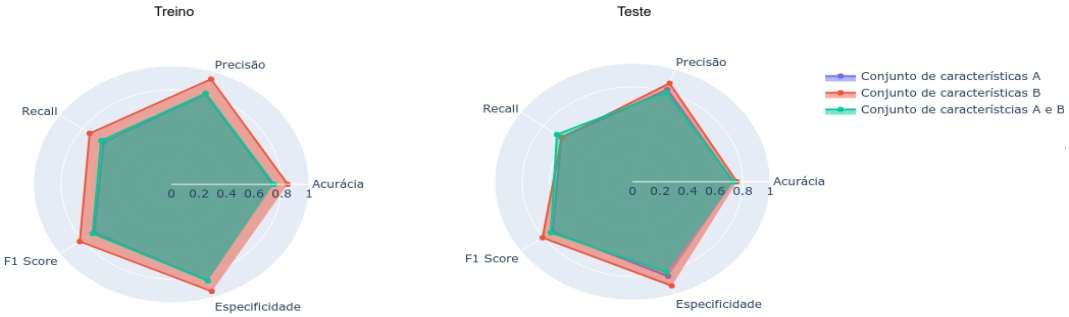
Fonte: Autoria própria (2020).

Figura 2 – Radar plot: classificação do *Dataset* 1 com 15 características.



Fonte: Autoria própria (2020).

Figura 3 – Radar plot: classificação do *Dataset* 2 com 15 características.



Fonte: Autoria própria (2020).

CONCLUSÃO

Apesar de miRNAs e mirtrons possuírem a mesma origem, apresentam distinções em etapas biogênicas e características estruturais e uma comparação entre suas diferenças estruturais permitem a compreensão da regulação de processos biológicos e fatores causadores de doenças. O classificador desenvolvido permite a identificação de características relevantes para distinção de miRNAs e mirtrons e através dos resultados obtidos é possível afirmar que a aplicação desses conjuntos de características apresenta potencial para a realização dessa distinção.

Para o *Dataset* 1, os valores de acurácia da etapa de teste quando avaliadas apenas 15 características foram muito próximos aos resultados para todas as características. Já para o *Dataset* 2, a acurácia foi mais alta na etapa de teste com 15 características do que para todo o conjunto, comprovando que foi possível otimizar o tempo de execução do algoritmo. De modo geral, as melhores métricas na etapa de teste foram obtidas para o *Dataset* 2 com o conjunto de características B. Além disso, tanto para o *Dataset* 1 quanto para o *Dataset* 2 o conjunto de características B apresentou a melhor acurácia.

AGRADECIMENTOS

Meus agradecimentos a Fundação Araucária pela bolsa de Iniciação Científica, ao meu orientador Alexandre Rossi Paschoal e a Universidade Tecnológica Federal (UTFPR-CP) pela oportunidade de realizar esse projeto de pesquisa.

REFERÊNCIAS

A DA FONSECA, B. H. R.; DOMINGUES, D. S.; PASCHOAL, A. R. **mirtronDB: a mirtron knowledge base**. Bioinformatics, v. 35, n. 19, p. 3873–3874, mar. 2019.

RIELLA, C. V. **Small non-coding RNAs: from trash to treasure**. Brazilian Journal of Nephrology, scielo, v. 41, p. 168–169, jun. 2019. ISSN 0101-2800.

RORBACH, G.; UNOLD, O.; KONOPKA, B. **Distinguishing mirtrons from canonical miRNAs with data exploration and machine learning methods**. Scientific Reports, v. 8, dez. 2018.