SEI-SICITE 2021
Pesquisa e Extensão para um mundo em transformação

XI Seminário de Extensão e Inovação
XXVI Seminário de Iniciação Científica e Tecnológica
08 a 12 de Novembro - Guarapuava/PR

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS GUARAPUAVA

Pipeline baseado em aprendizado de máquina para análises de RNAs circulares

Pipeline based on machine learning for analysis of circular RNAs

Alvaro Pedroso Queiroz (orientado) *, Danilo Sipoli Sanches (orientador)†

RESUMO

O avanço da pesquisa na área de bioinformática tem sido emergente. O tamanho dos dados acumulados em vários projetos de sequenciamento está aumentando exponencialmente, e assim, técnicas computacionais envolvendo algoritmos de classificação foram propostos para reduzir as dificuldades encontradas em métodos experimentais. Um dos campos que atrai uma quantidade crescente de atenção é o estudo de RNAs não codificantes, mais precisamente os RNAs circulares e os longos não codificantes. Foi confirmado que eles participam de muitos processos biológicos e as diferenças entre essas duas classes de RNAs não codificantes não foram totalmente descobertas. Logo, a diferenciação entre essas classes é uma tarefa complexa. O *pipeline* proposto utiliza 8 técnicas de extração de características diferentes, com descritores matemáticos e convencionais, aplicados em dados de sequências biológicas para alimentar um modelo de aprendizado de máquina, selecionado a partir de uma técnica de AutoML. Os experimentos preliminares apresentaram resultados promissores, com alta acurácia (0,9530), precisão (0,9416), revocação (0,9435), *f1-score* (0,9425) e AUC (0,9897) para sequências de RNAs circulares e longos não codificantes humanos. Também é apresentado a importância das características utilizadas no modelo criado e como a hibridização de técnicas matemáticas e convencionais apresentou resultados positivos para a predição.

Palavras-chave: *Pipeline*. Aprendizado de máquina. RNAs circulares.

ABSTRACT

The advance of research in the area of bioinformatics has been emerging. The size of data accumulated in various sequencing projects is increasing exponentially, and thus computational techniques involving classification algorithms have been proposed to reduce the difficulties encountered in experimental methods. One of the fields that attracts an increasing amount of attention is the study of non-coding RNAs, more precisely the circular RNAs and the long non-coding ones. It has been confirmed that they participate in many biological processes and the differences between these two classes of non-coding RNAs have not been fully discovered. So the differentiation between these classes is a complex task. The proposed pipeline uses 8 different feature extraction techniques, with mathematical and conventional descriptors, applied to biological sequence data to feed a machine learning model, selected using the AutoML technique. Preliminary experiments showed promising results, with high accuracy (0.9530), precision (0.9416), recall (0.9435), f1-score (0.9425) and AUC (0.9897) for human circular and long non-coding RNAs sequences. It is also presented the importance of the characteristics used in the created model and how the hybridization of mathematical and conventional techniques presented positive results for the prediction.

Keywords: Pipeline. Machine Learning. Circular RNAs.

* Engenharia de Computação, Universidade Tecnológica Federal do Paraná, Campus Cornélio Procópio, Paraná, Brasil; alvaroq@alunos.utfpr.edu.br.
† Universidade Tecnológica Federal do Paraná, Campus Cornélio Procópio, Paraná, Brasil; danilosanches@utfpr.edu.br.

1 INTRODUÇÃO

Nas últimas três décadas, o avanço da pesquisa na área de bioinformática tem sido emergente. A princípio, os objetivos finais eram armazenar e gerenciar dados biológicos para o desenvolvimento e análises de ferramentas computacionais que permitissem o aprimoramento de sua compreensão. Assim sendo, o tamanho de dados acumulados em vários projetos de sequenciamento está aumentando exponencialmente, apresentando dificuldades para realização de métodos experimentais. Com a finalidade de reduzir tal lacuna presente entre o sequenciamento e a anotação de suas funções, muitas técnicas computacionais envolvendo algoritmos de classificação e agrupamento foram propostos (IQBAL et al., 2014).

Isto posto, um dos campos que atrai uma quantidade crescente de atenção é o estudo de RNAs não codificantes, mais precisamente os RNAs circulares (circRNA) e os longos não codificantes (lncRNA). Foi confirmado que eles participam de muitos processos biológicos, incluindo papéis na regulação da transcrição, regulação de genes codificantes de proteínas e ligação a proteínas associadas ao RNA. Porém, as diferenças entre essas duas classes de RNAs não codificantes não foram totalmente descobertas, e a detecção de RNAs circulares de outros RNAs longos não codificantes é muito difícil usando técnicas simples (CHEN et al., 2017). Assim sendo, surge a questão: É possível utilizar métodos de aprendizado de máquina que possam realizar tal classificação?

Segundo Baldi et al. (2001), as abordagens de aprendizado de máquina são mais adequadas para áreas onde há uma abundância de dados, porém possui pouca teoria. Portanto, este é exatamente a situação encontrada na área de biologia molecular computacional. Embora os dados de sequências disponíveis estejam rapidamente em ascensão, o conhecimento atual da biologia constitui apenas em uma pequena fração descoberta.

Dessa forma, muitos trabalhos desenvolveram *pipelines* que permitem classificar as diferentes classes de RNAs supracitadas. No trabalho de Pan e Xiong (2015), foi desenvolvido o PredcircRNA, uma abordagem de aprendizado de máquina aplicando aprendizagem de múltiplos *kernels* para fusão de recursos heterogêneos, baseados nas características extraídas com as técnicas de *graph feature*, *conservation*, *component composition*, ALU *repeat*, *Open Reading Frame* (ORF) e *Single Nucleotide Polymorphism* (SNP) para a classificação de RNAs circulares. Utilizando validação cruzada, obtiveram uma precisão de 0,778.

Utilizando os mesmos recursos, Chen et al. (2017) construíram um modelo de classificação utilizado método de seleção de recursos incrementais e o algoritmo H-ELM, resultando em uma acurácia de 0,789.

Por fim, o software PCirc foi desenvolvido por Yin et al. (2021), utilizando aprendizado de máquina com a técnica Random Forest, com o intuito de prever RNAs circulares de plantas. Para extração de características, utilizaram-se os métodos ORF, k-mers e *splicing junction sequence coding* (SJSC). Nos métodos de avaliação, obteve-se uma pontuação de acurácia acima de 0,8.

Assim, ponderado o impacto e a crescente quantidade de estudos publicados sobre ferramentas de predição de classes de RNAs, este trabalho tem por objetivo desenvolver um *pipeline* baseado em aprendizado de máquina para classificação e análise de sequências de RNAs circulares.

2 MATERIAIS E MÉTODOS

Para o desenvolvimento do *pipeline* de classificação de RNAs circulares, utilizou-se a linguagem de programação Python, em conjunto com a plataforma Anaconda, onde possui 150 pacotes pré-instalados que possuem grande utilidade para a área da ciência de dados (KADIYALA; KUMAR, 2017).

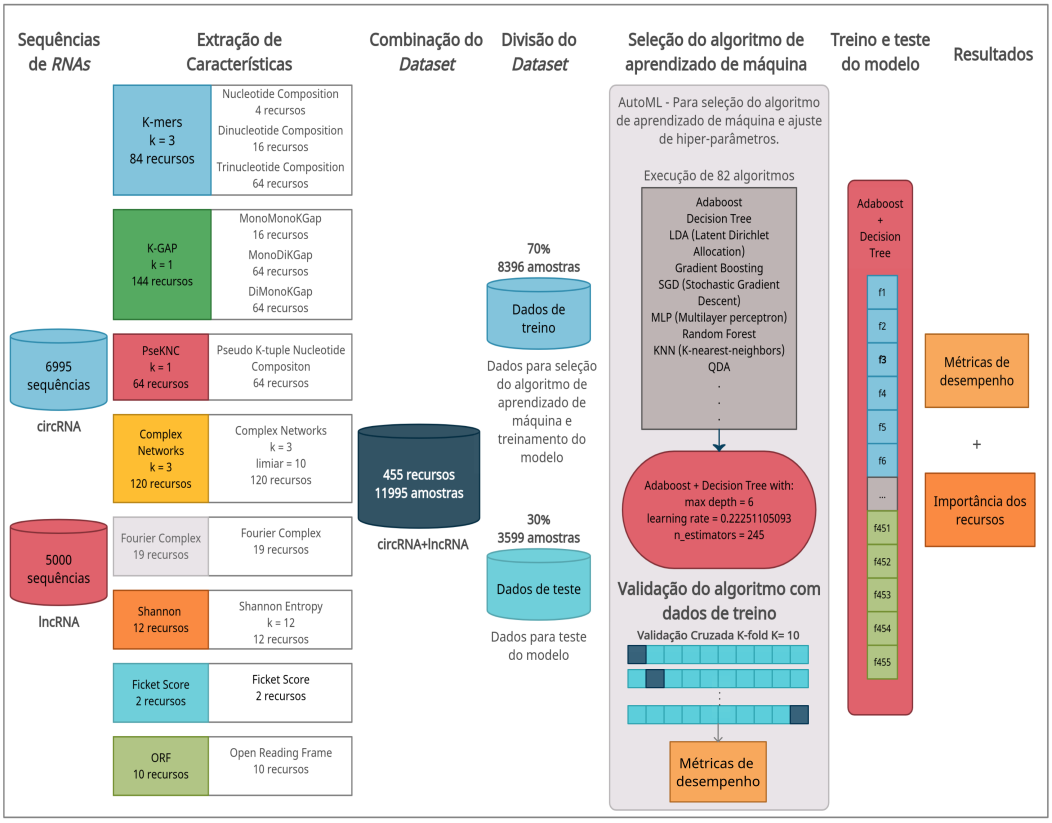
Para a extração de características dos dados, utilizou-se a ferramenta MathFeature, que implementa descritores

matemáticos e convencionais capazes de extrair informações numéricas relevantes de sequências biológicas. (BONIDIA; SANCHES; CARVALHO, 2020).

Para seleção do algoritmo de *machine learning*, utilizou-se o Auto-Sklearn, uma abordagem de AutoML que permite a execução do aprendizado de máquina automatizado, realizando a escolha de um bom algoritmo e a definição de seus respectivos hiper-parâmetros de forma automática, aplicando um eficiente método Bayesiano de otimização (FEURER et al., 2015).

Posto a problemática da classificação de RNAs longos não codificantes e RNAs circulares, bem como as ferramentas e tecnologias utilizadas, desenvolveu-se o seguinte *pipeline* ilustrado na Figura 1.

Figura 1 – *Pipeline* proposto para realização da classificação de RNAs circulares.



Fonte: Autoria própria (2021).

O *pipeline* consiste primeiramente na obtenção dos dados, com 6995 sequências de RNAs circulares e 5000 sequências de RNAs longos não codificantes. Vale ressaltar que os dados presentes nas sequências estão na forma traduzida para DNAs, ou seja, é apresentado os nucleotídeos A, G, C e T.

Esses dados foram utilizados nas técnicas de extração de características apresentadas no *pipeline*, obtendo como base de dados final 11995 amostras e 445 características.

Em seguida, iniciou-se a divisão dos dados para o treinamento e teste. Sendo assim, os dados foram divididos em 70% (8396 amostras) destinadas para a seleção da técnica de aprendizado de máquina e o treinamento do modelo final, já os 30% restantes (3599 amostras) foram utilizadas para o teste de desempenho do modelo.

Na etapa de seleção do algoritmo de aprendizado de máquina, utilizou-se a técnica de AutoML, onde permitiu testar 82 algoritmos diferentes, obtendo como melhor resultado a combinação do método de aprendizado conjunto Adaboost combinado com árvores de decisão, utilizando os hiper-parâmetros definidos com 265 estimadores,

profundidade máxima igual a 6 e uma taxa de aprendizado igual a 0,22251105093.

A fim de entender como o modelo de *machine learning* é generalizado com dados desconhecidos, aplicou-se a técnica de validação cruzada *k-fold* (BROWNE, 2000) sobre os dados de treinamento, com *k* igual a 10.

Após a validação do algoritmo, iniciou-se o processo de modelagem final. O modelo foi treinado e estimado seu desempenho através das métricas de avaliação (acurácia, precisão, revocação, *f1-score* e AUC) (SKIENA, 2017) com os dados de teste até então desconhecidos. Além disso, investigou-se a importância das características para o modelo, e o quanto elas representam com base em suas técnicas de extração.

3 RESULTADOS

Na aplicação do *pipeline* proposto, realizou-se a técnica de validação cruzada sobre o algoritmo Adaboost combinado com árvores de decisão selecionadas pela técnica de AutoML. Os resultados estão expressos na Tabela 1, onde é demonstrado a média para cada métrica calculada. Conforme os resultados obtidos, a técnica foi utilizada para treinamento do modelo final, obtendo os resultados também expressados na Tabela 1. Para fins de comparação, também é apresentado os resultados obtidos utilizando o algoritmo Random Forest com configurações padrões como *baseline*.

Tabela 1 – Resultado dos testes.

	Acurácia	Precisão	Revocação	F1-Score	AUC
Conjunto de treino - Validação cruzada *10-fold*	0,9579	0,9498	0,9492	0,9495	0,9903
Conjunto de teste - Adaboost + Decision Tree	0,9530	0,9416	0,9435	0,9425	0,9897
Conjunto de teste - *Baseline* (Random Forest)	0,8608	0,8903	0,7515	0,8151	0,8964

Fonte: Autoria própria (2021).

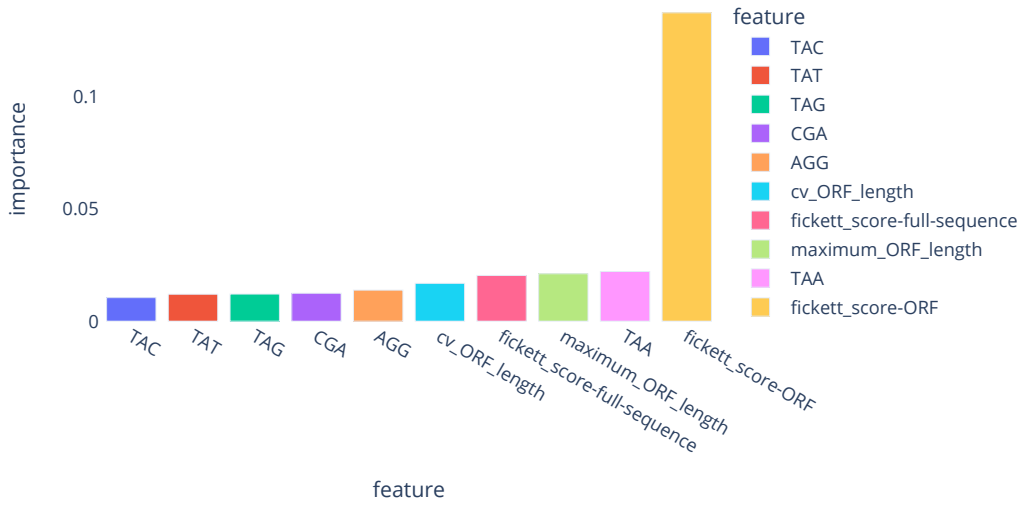
Os resultados obtidos do *baseline* já se mostraram muito promissores, obtendo indícios de que as características obtidas contribuem positivamente para a definição de RNAs circulares. Assim sendo, percebe-se que a utilização da técnica de AutoML pode proporcionar a escolha de uma combinação de algoritmos que permitem uma evolução na eficiência da classificação dos RNAs, visto que as métricas obtiveram melhores resultados próximos de 0,9500. Também vale destacar a proximidade dos resultados obtidos utilizando validação cruzada e o modelo final, apresentando indícios de que o modelo está atuando de forma generalizada.

Com base nisso, investigou-se quais características são mais importantes para a classificação de RNAs circulares e longos não codificantes, sendo apresentado na Figura 2 as 10 características principais segundo cálculos da técnica de "importância de Gini" de forma normalizada.

Conforme apresentado na Figura 2, o recurso "Ficket Score - ORF" é o mais importante na decisão do modelo, obtendo um valor bastante expressivo em relação aos demais recursos que estão representando os 10 principais. Também pode-se observar que às duas características extraídas pela técnica Ficket Score, estão presentes nas mais importantes, reforçando ser uma técnica muito determinante para este tipo de classificação. Outro ponto interessante é a quantidade de recursos extraídos da técnica Tri-nucleotide Composition (TNC), pois constituem grande parte dos recursos principais. Vale ressaltar que as informações obtidas dessa técnica podem ser investigadas por biólogos para tentar entender o significado dessas combinações de nucleotídeos, e como elas podem ser utilizadas para diferenciar RNA circulares e longos não codificantes. Partindo desse princípio, foram exploradas as técnicas que resultam em recursos que mais importam para o modelo, apresentadas na Tabela 2.

A importância das características extraídas pela técnica Tri-nucleotide composition (TNC) juntas representam

Figura 2 – Os 10 melhores recursos com base na Importância de Gini (normalizado).



Fonte: Autoria própria (2021).

Tabela 2 – Importância do recurso com base no descritor.

Descritor	Importância
Tri-nucleotide composition (TNC)	0,375193
Ficket Score	0,157546
Complex Network	0,117040
Open Reading Frame (ORF)	0,081057
MonoDiKGap	0,073297
DiMonoKGap	0,072506
Di-nucleotide composition (DNC)	0,054250
Fourier Transform with Complex Number	0,020534
Pseudo K-tuple Nucleotide Composition: *type* 2	0,018784
MonoMonoKGAP	0,011886
Shannon Entropy	0,009029
Nucleic acid composition (NAC)	0,008878

Fonte: Autoria própria (2021).

37,7% da importância total na decisão do modelo. Também vale destacar a divisão de importância entre as demais técnicas, destacando as características do Ficket Score e da Complex Network, indicando ser benéfico a hibridização com abordagens de extração de características baseadas em conceitos matemáticos.

4 CONCLUSÕES

Portanto, após a realização da análise dos resultados obtidos, constatou-se que este trabalho de pesquisa conseguiu alcançar o objetivo proposto, ao realizar um *pipeline* que permite a classificação de RNAs circulares e longos não codificantes, apresentando resultados promissores em relação a sequências genéticas humanas.

O trabalho também permite acrescentar a relevância de uma proposta de *pipeline* baseada em características híbridas, ou seja, utilizando tanto técnicas de extração de características conservadoras quanto técnicas baseadas

em conceitos matemáticos.

Por fim, o presente trabalho abre caminho para aplicação do *pipeline* e a verificação de seu desempenho em diferentes bases de dados e diferentes organismos (eucariotos e procariotos), podendo abordar novas técnicas de extração e seleção de características presentes na literatura.

AGRADECIMENTOS

Agradecimento à UTFPR pela bolsa de iniciação científica, que contribuiu para o desenvolvimento desse trabalho e ao professor e orientador Danilo Sipoli Sanches pela oportunidade.

REFERÊNCIAS

BALDI, P. et al. **Bioinformatics: The Machine Learning Approach**. USA: Bradford, 2001. (A Bradford book). ISBN 9780262025065.

BONIDIA, Robson P.; SANCHES, Danilo S.; CARVALHO, André C.P.L.F. de. MathFeature: Feature Extraction Package for Biological Sequences Based on Mathematical Descriptors. Cold Spring Harbor Laboratory, dez. 2020. DOI: 10.1101/2020.12.19.423610. Disponível em: .

BROWNE, Michael W. Cross-Validation Methods. **Journal of Mathematical Psychology**, Elsevier BV, v. 44, n. 1, p. 108–132, mar. 2000. DOI: 10.1006/jmps.1999.1279. Disponível em: .

CHEN, Lei et al. Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. **Molecular Genetics and Genomics**, Springer Science e Business Media LLC, v. 293, n. 1, p. 137–149, set. 2017. DOI: 10.1007/s00438-017-1372-7. Disponível em: .

FEURER, Matthias et al. Efficient and Robust Automated Machine Learning. In: PROCEEDINGS of the 28th International Conference on Neural Information Processing Systems - Volume 2. Montreal, Canada: MIT Press, 2015. (NIPS'15), p. 2755–2763. Disponível em: .

IQBAL, Muhammad Javed et al. Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics. **The Scientific World Journal**, Hindawi Limited, v. 2014, p. 1–12, 2014. DOI: 10.1155/2014/173869. Disponível em: .

KADIYALA, Akhil; KUMAR, Ashok. Applications of Python to evaluate environmental data science problems. **Environmental Progress & Sustainable Energy**, Wiley, v. 36, n. 6, p. 1580–1586, out. 2017. DOI: 10.1002/ep.12786. Disponível em: .

PAN, Xiaoyong; XIONG, Kai. PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features. **Molecular BioSystems**, Royal Society of Chemistry (RSC), v. 11, n. 8, p. 2219–2226, 2015. DOI: 10.1039/c5mb00214a. Disponível em: .

SKIENA, Steven S. **The Data Science Design Manual**. 1st. Switzerland: Springer Publishing Company, Incorporated, 2017. ISBN 3319554433.

YIN, Shuwei et al. PCirc: random forest-based plant circRNA identification software. **BMC Bioinformatics**, Springer Science e Business Media LLC, v. 22, n. 1, jan. 2021. DOI: 10.1186/s12859-020-03944-1. Disponível em: .