Classificação de sequências biológicas usando Máxima Entropia

Classification of biological sequences using Maximum Entropy

Murilo Montanini Breve * Fabrício Martins Lopes †

11 de outubro de 2021

RESUMO

Nas últimas décadas, a quantidade de genes de RNAs sequenciados vêm aumentando drasticamente, as novas tecnologias de sequenciamento estão gerando milhares de RNAs não codificantes, cuja função e significado ainda não são completamente entendidos. Diante disto, um importante desafio é a necessidade de distinguir mRNAs, lncRNAs e sncRNAs de forma assertiva. A correta identificação dessas transcrições favorece a compreensão da expressão e da regulação da informação genética. Por outro lado, a utilização da Máxima Entropia foi aplicada com sucesso em muitos problemas do mundo real em diferentes contextos. Portanto, este trabalho apresenta uma metodologia eficiente e eficaz baseada na Máxima Entropia e na utilização de características extraídas de grafos para classificação de sequências de mRNA e ncRNA. Considerando dois conjuntos de dados, experimentos foram realizados para avaliar o método proposto em comparação a métodos importantes na literatura como CPC2, PLEK e BASiNET. Os resultados indicaram a adequação da metodologia proposta, alcançando acurácias superiores aos trabalhos CPC2 e PLEK, e semelhantes ao BASiNET. Além disso, o método proposto executou a classificação com menor tempo de processamento, indicando uma diminuição da complexidade mantendo a assertividade na classificação.

Palavras-chave: Classificação de RNA. Máxima Entropia. Grafos.

ABSTRACT

In recent decades, the amount of sequenced RNA genes has increased dramatically, new sequencing technologies has been generating thousands of non-coding transcripts, whose function and meaning are still not fully understood. Given this, an important challenge is the need to distinguish mRNAs, lncRNAs and sncRNAs in an assertive way. The correct identification of these transcripts favors the understanding of the expression and regulation of genetic information. On the other hand, the use of Maximum Entropy has been successfully applied to many real-world problems in different contexts. Therefore, this work presents an efficient and effective methodology based on Maximum Entropy for classification of mRNA and ncRNA sequences. With two datasets, experiments were performed to evaluate the proposed method in comparison with important works in the literature such as CPC2, PLEK and BASiNET. The results indicated the adequacy of the proposed methodology, reaching superior accuracies in comparison to CPC2 and PLEK works, and similar to BASiNET. Moreover, the proposed method performed the classification with less processing time, which indicates a complexity reduction while maintaining an assertiveness in the classification.

Keywords: RNA Classification. Maximum Entropy. Graphs.

1 INTRODUÇÃO

No século XIX, o bioquímico suíço Friedrich Miescher (1844-1895) isolou de uma célula, um ácido que continha fósforo e nitrogênio, e após 20 anos desta descoberta, seu discípulo, Richard Altmann estabeleceu o

* Engenharia de Controle e Automação; murilobreve@alunos.utfpr.edu.br; https://orcid.org/0000-0002-5781-048.
† Engenharia de Computação; fabricio@utfpr.edu.br; https://orcid.org/0000-0002-8786-3313.

nome de ácido nucleico a este composto, como conhecemos hoje (DAHM, 2005). E devido a estes avanços ciêntificos, em 1953, James D. Watson e Francis H. Crick, viriam a publicar "A Structure for Deoxyribose Nucleic Acid"(FEUGHELMAN et al., 1955) a primeira menção da estrutura DNA no campo cientifico, a qual seria considerada umas das mais importantes contribuições a Biologia.

Devido a sua indispensável participação na manutenção e criação da vida em nosso planeta, os ácidos nucleicos ganharam espaço e notoriedade em diversos campos do conhecimento, como por exemplo na bioinformática, já que a quantidade de informação presente nos aglomerados destas partículas é enorme, o que torna impraticável a análise dos dados de forma manual. No campo da genética e genômica, a bioinformática participa no sequenciamento e anotação de genes de um organismo, e suas mutações observadas (DAVIES, 2001).

Em suma, um gene de RNA armazena informações genéticas por meio da combinação de quatro tipos de bases nitrogenadas (adenina (A), uracila (U), guanina (G) e citosina (C)), que irão formar moléculas de RNA distintas conforme a sequência. As informações gênicas coordenam o desenvolvimento e funcionamento de todas as formas de vidas conhecidas no planeta (VARKI, 2005). As sequências de RNA são analisadas para determinar quais genes codificam proteínas e, também, para comparar genes dentro de uma espécie ou entre espécies diferentes, o que pode mostrar semelhanças entre funções proteicas ou relações entre espécies. Há várias classes de RNA, como os RNAs mensageiros (mRNAs) que são codificantes de proteínas e os RNAs não codificantes (ncRNAs) que são subdivididos em duas categorias, pequenos e longos (ALBERTS, 2009).

Os RNAs longos não codificantes (lncRNAs) estão emergindo como novos participantes no paradigma do câncer, demonstrando papéis potenciais nas vias oncogênicas e supressoras de tumor (GIBB; BROWN; LAM, 2011). Os RNAs pequenos não codificantes (sncRNAs) fazem parte de oligo-nucleotídeos reguladores não codificantes com amplas funções fisiológicas e morfológicas. Eles controlam a programação genética das células e podem modular processos de diferenciação e morte (KLIMENKO, 2017).

De fato cada classe de RNA desempenha um papel ativo e distinto dentro das células, desde catalisar reações biológicas e até controlar a expressão gênica nos seres vivos (RINN; CHANG, 2012). Diante disto, diferenciar classes de RNA entre mRNA, lncRNA e sncRNA, pode contribuir para uma maior compreensão de suas funcionalidades e mecanismos.

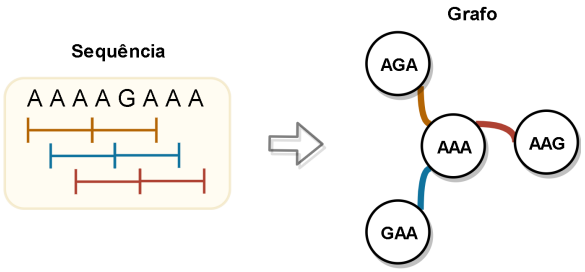
2 MÉTODO

Para o desenvolvimento deste trabalho foi necessário estudo e adaptação do código do BASiNET (ITO et al., 2018). Foram utilizados grafos complexos (redes complexas), criados a partir de sequências de RNA (contidas nos datasets do PLEK (LI; ZHANG; ZHOU, 2014) e do CPC2 (KANG et al., 2017)) armazenadas em arquivos do tipo $FASTA$. Para a produção destas redes complexas, foi necessário a configuração de dois parâmetros, o $passo$ e a $palavra$. A função do $passo$ é definir a distância de uma aresta à outra, e a $palavra$ se refere a quantos nucleotídeos são agrupados em cada vértice. Este processo é melhor descrito na Figura 1. Neste projeto, semelhante ao trabalho BASiNET, usamos $passo = 1$ e $palavra = 3$.

Por outro lado, o conceito de Máxima Entropia procura identificar a distribuição das probabilidades de um sistema. Sendo possível encontrar o ponto onde a entropia está maximizada, indicando a posição da separabilidade de duas partes distintas de um conjunto de dados ou de informações (JAYNES, 1957). Dessa forma, ao ser aplicado a um histograma de uma imagem equalizada, este método encontra o limiar que maximiza a soma das entropias de duas partes distintas (objeto e fundo) de uma imagem (LOPES, 2003). Análogo a este pensamento, para este trabalho, usaremos esta metodologia para a produção de um filtro de arestas, isto é, ao

analisar um espectro de 4096 possibilidades de arestas, resultante de uma matriz de 64x64 vértices, separar o que é importante (objeto), e o que não é importante (fundo) para a classificação de classes de RNA.

Figura 1 – Esquema explicativo da conversão de sequências de RNA em grafos.

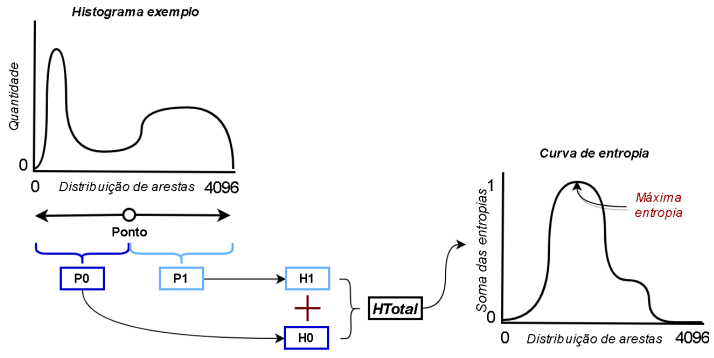


Fonte: Autoria própria (2021).

Este espectro de 4096 possibilidades de arestas será convertido em um histograma, o qual será utilizado para o encontrar o limiar, ou seja, o ponto (aresta) com a máxima entropia no histograma. Para este propósito, é necessário o cálculo de P0 e P1, probabilidades das duas partes distintas, esquerda e direita respectivamente, de cada ponto no histograma (4096 pontos). Ao obter P0 e P1 de cada ponto dos 4096 presentes no histograma, são também calculadas as entropias referentes (H0 e H1). Deste modo, para cada ponto, uma entropia total chamada de HTotal (H0 + H1) é encontrada, a qual é comparada com todas as outras entropias, selecionando o ponto com a maior entropia no histograma (vide Figura 2).

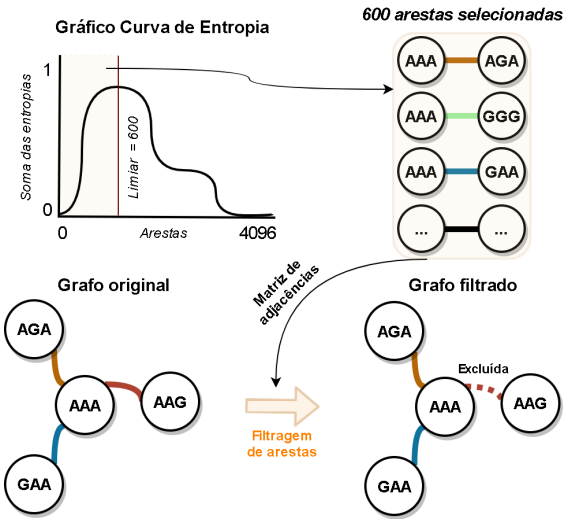
As arestas selecionadas pela máxima entropia são organizadas em uma lista de matrizes, isto é, uma matriz para cada classe de RNA. Com isso, é possível filtrar os grafos das sequências desejadas de RNA, este processo de filtragem está explicado na Figura 3. Com os grafos filtrados, são produzidas medidas topológicas que são organizadas em um dataframe, e pela ferramenta Random Forest (LIAW; WIENER, 2002) presente no software R (R CORE TEAM, 2021), é executada a classificação.

Figura 2 – Exemplo da produção da curva de entropia.



Fonte: Autoria própria (2021).

Figura 3 – Esquema explicativo da filtragem das arestas através da máxima entropia.



Fonte: Autoria própria (2021).

3 RESULTADOS

A Tabela 1 apresenta os resultados das classificações referentes aos mRNAs, lncRNAs e sncRNAs, a base de dados usada foi a mesma utilizada no trabalho PLEK (LI; ZHANG; ZHOU, 2014). Porém, devido a limitação dos outros métodos, as classes foram organizadas entre mRNA e ncRNA. Os resultados do método proposto foram comparados com os resultados dos classificadores PLEK (LI; ZHANG; ZHOU, 2014), CPC2 (KANG et al., 2017), BASiNET* (BASiNET sem threshold) e BASiNET (ITO et al., 2018).

Os resultados do método proposto em comparação aos classificadores CPC2, PLEK e BASiNET* foram satisfatórios, visto que, foram superiores em acurácia tanto por espécies, quanto na média geral e no desvio padrão (vide a Tabela 1). Entretanto, ao comparar com o BASiNET (ITO et al., 2018), os resultados foram semelhantes, com uma pequena vantagem para o método proposto na média geral e no desvio padrão na classe ncRNA, enquanto o BASiNET foi superior na classe mRNA. A Figura 4 apresenta um gráfico que melhor representa a comparação dos métodos em relação as espécies.

Com a Tabela 2 e com o gráfico contido na Figura 5, é possível observar que os resultados do método proposto foram superiores em acurácia na comparação entre os classificadores CPC2, PLEK e BASiNET*, na base de dados do CPC2. Contudo, ao comparar com o BASiNET (ITO et al., 2018), os resultados foram bastante semelhantes, com uma pequena vantagem para o BASiNET na média geral e no desvio padrão na classe mRNA, enquanto o método proposto foi superior na classe ncRNA.

Tabela 1 – Taxas de acerto na classificação das sequências de mRNA e ncRNA, dataset PLEK (LI; ZHANG; ZHOU, 2014).

	Classe	PLEK	CPC2	BASiNET*	BASiNET	Método Proposto
Média por classe	mRNA	89.74	94.75	95.96	99.99	99.94
	ncRNA	97.58	99.13	86.02	99.59	99.94
Média geral	**–**	**93.66**	**96.94**	**91.20**	**99.79**	**99.94**
Desvio padrão	mRNA	4.81	1.45	6.21	0.03	0.09
	ncRNA	3.88	1.89	15.97	0.44	0.16

Fonte: Autoria própria (2021).

Tabela 2 – Taxas de acerto na classificação das sequências de mRNA e ncRNA, dataset CPC2 (KANG et al., 2017).

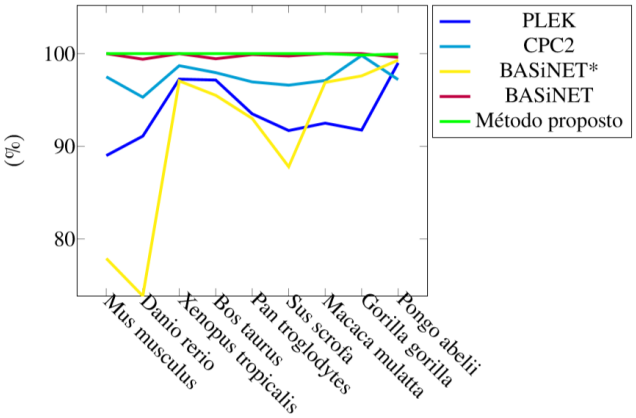
	Classe	PLEK	CPC2	BASiNET*	BASiNET	Método Proposto
Média por classe	mRNA	79.92	96.02	89.52	99.62	99.07
	ncRNA	96.17	96.92	83.09	99.45	99.49
Média geral	**–**	**88.05**	**96.47**	**86.31**	**99.54**	**99.28**
Desvio padrão	mRNA	17.92	2.03	8.15	0.58	0.68
	ncRNA	6.67	4.18	15.13	0.81	0.67

Fonte: Autoria própria (2021).

Quanto ao tempo de processamento, no dataset do CPC2, o método proposto executou a classificação das espécies em 246,6 minutos, em média. Uma redução de 5,3% em comparação ao BASiNET, que executou a classificação das mesmas sequências em 260,4 minutos, sendo que o BASiNET sem *thresholds* executou em 199,8 minutos. Isso se deve principalmente à relação do número de características que cada método utiliza, 10 em contraste com até 2000 características do BASiNET (geradas pelo uso de *thresholds*). Indicando que a maximização de entropia reduz a complexidade em termos da dimensionalidade das características, o que simplifica o problema e mantém uma alta assertividade na classificação. Logo, o método de máxima entropia foi

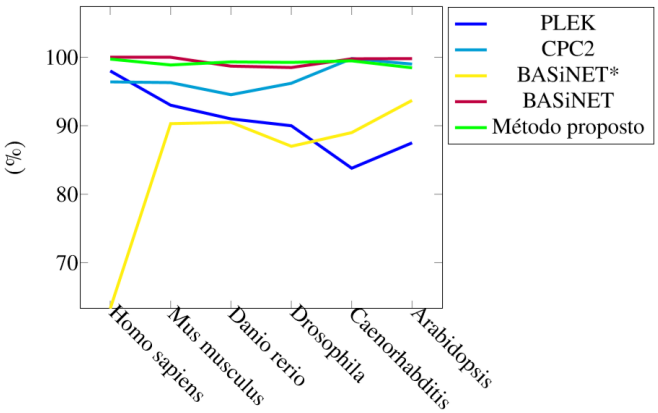
adequado para diminuir o tempo de processamento e manter as taxas de acurácias em comparação ao uso de *thresholds* pelo BASiNET.

Figura 4 – Gráfico para comparação dos classificadores por espécie (dataset PLEK).



Fonte: Autoria própria (2021).

Figura 5 – Gráfico para comparação dos classificadores por espécie (dataset CPC2).



Fonte: Autoria própria (2021).

4 CONCLUSÕES

Com os resultados observados nos testes dos dois datasets PLEK e CPC2, bem como a comparação com outros métodos com objetivos semelhantes, fica evidente que a metodologia adotada neste trabalho foi adequada para seu propósito. Ainda assim, o método do BASiNET teve taxas da acurácia semelhantes, porém com um tempo de processamento maior. Por isso, a adaptação do método de máxima entropia para selecionar as arestas com maior importância se provou eficaz e eficiente na classificação de diferentes classes de RNA em comparação ao uso de *thresholds*.

Do mesmo modo, a extração de medidas topológicas de uma rede complexa gerada por uma sequência, pode contribuir para um melhor entendimento e diferenciação dos RNAs. Além disso, este método pode ser aplicado a outras sequências biológicas como as de DNAs, que possuem uma estrutura semelhante aos RNAs, auxiliando assim, a análise dos crescentes dados gerados nos últimos anos.

AGRADECIMENTOS

Os autores agradecem a Universidade Tecnológica Federal do Paraná (UTFPR), pela bolsa de Iniciação Científica (PIBIC 2020/2021), concedida ao acadêmico Murilo Montanini Breve.

REFERÊNCIAS

ALBERTS. **Biologia molecular da célula**. 6. ed. [S.l.]: Artmed Editora, 2009.

DAHM, Ralf. Friedrich Miescher and the discovery of DNA. **Developmental Biology**, v. 278, n. 2, p. 274–288, 2005. ISSN 0012-1606. DOI: https://doi.org/10.1016/j.ydbio.2004.11.028. Disponível em: .

DAVIES, Kevin. **Decifrando o Genoma**. [S.l.]: Companhia das Letras, 2001.

FEUGHELMAN, M. et al. Molecular Structure of Deoxyribose Nucleic Acid and Nucleoprotein. **Nature**, v. 175, n. 4463, p. 834–838, mai. 1955. ISSN 1476-4687. DOI: 10.1038/175834a0. Disponível em: .

GIBB, Ewan A.; BROWN, Carolyn J.; LAM, Wan L. The functional role of long non-coding RNA in human carcinomas. **Molecular Cancer**, v. 10, n. 1, p. 38, abr. 2011. ISSN 1476-4598. DOI: 10.1186/1476-4598-10-38. Disponível em: .

ITO, Eric Augusto et al. BASiNET—BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification. **Nucleic Acids Research**, v. 46, n. 16, e96–e96, jun. 2018. ISSN 0305-1048. DOI: 10.1093/nar/gky462. eprint: https://academic.oup.com/nar/article-pdf/46/16/e96/25802486/gky462.pdf. Disponível em: .

JAYNES, E. T. Information Theory and Statistical Mechanics. **Phys. Rev.**, American Physical Society, v. 106, p. 620–630, 4 mai. 1957. DOI: 10.1103/PhysRev.106.620. Disponível em: .

KANG, Yu-Jian et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. **Nucleic Acids Research**, v. 45, W1, w12–w16, mai. 2017. ISSN 0305-1048. DOI: 10.1093/nar/gkx428. eprint: https://academic.oup.com/nar/article-pdf/45/W1/W12/23741208/gkx428.pdf. Disponível em: .

KLIMENKO, Oxana V. Small non-coding RNAs as regulators of structural evolution and carcinogenesis. **Non-coding RNA Research**, v. 2, n. 2, p. 88–92, 2017. ISSN 2468-0540. DOI: https://doi.org/10.1016/j.ncrna.2017.06.002. Disponível em: .

LI, Aimin; ZHANG, Junying; ZHOU, Zhongyin. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. **BMC Bioinformatics**, v. 15, n. 1, p. 311, set. 2014. ISSN 1471-2105. DOI: 10.1186/1471-2105-15-311. Disponível em: .

LIAW, Andy; WIENER, Matthew. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002. Disponível em: .

LOPES, FABRÍCIO MARTINS. UM MODELO PERCEPTIVO DE LIMIARIZAÇÃO DE IMAGENS DIGITAIS. **Universidade Estadual de Maringá**, 2003.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Disponível em: .

RINN, John L.; CHANG, Howard Y. Genome regulation by long noncoding RNAs. eng. **Annual review of biochemistry**, v. 81, p. 145–166, 2012. PMC3858397[pmcid]. ISSN 1545-4509. DOI: 10.1146/annurev-biochem-051410-092902. Disponível em: .

VARKI, A. Comparing the human and chimpanzee genomes: Searching for needles in a haystack. **Genome Research**, Cold Spring Harbor Laboratory, v. 15, n. 12, p. 1746–1758, dez. 2005. DOI: 10.1101/gr.3737405. Disponível em: .