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

Similaridade entre sequências genômicas aplicada na inferência de redes de regulação gênicas

Similarity between genomic sequences applied in the inference of gene regulation networks

Murilo Montaini Breve
murilo_breve@hotmail.com
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Fabrício Martins Lopes
fabricio@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

RESUMO

Este trabalho tem como intuito a produção de um método eficaz para a classificação de sequências de RNA, e determinar se são codificantes ou não-codificantes. O algoritmo facilita a análise destes dados geralmente brutos e em enorme quantidade.MÉTODOS: Para cada sequência de RNA foi gerado grafo com base em 2 parâmetros: passo e tamanho da palavra. Após a produção dos grafos, é realizada uma filtragem das arestas mais exclusivas de cada classe, e assim abstraír medidas que descrevem a rede e partir dessas medidas foram feitas as classificações com o algoritmos Random Forest. RESULTADOS: Após aplicado a metodologia deste trabalho, os resultados finais foram muito positivos, com média de 99,78% de acerto para o classificador Random Forest, e superiores aos métodos comparados. CONCLUSÕES: Os resultados indicaram uma alta distinção entre as classes de RNA devido a filtragem das arestas exclusivas. A aplicação e estudo mais aprofundado deste método pode levar a um melhor entendimento do RNA não codificante. **PALAVRAS-CHAVE:** Bioinformática. Reconhecimento de padrões. Grafos.

ABSTRACT

This paper presents an efficient method to classify RNA sequences and determine whether they are coding or non-coding RNA sequences. The algorithm facilitates the analysis of huge quantities of raw RNA data. METHODS: A graph was generated for each RNA sequence based on 2 parameters: Step and Word. A filter was applied to the generated graphs in order to select the most exclusive edges of each class, thus deriving measurements that describe the graph. These measurements are then fed into a Random Forest classification algorithm. RESULTS: The presented algorithm shows an average accuracy of 99.78 %, higher than with other traditionally used classification methods. CONCLUSIONS: The results indicate a high distinction between the classes of RNA after the filtering of the exclusive edges. The application and further study of this method can lead to a better understanding of non-coding RNA. **KEYWORDS:** Bioinformatics. Pattern Recognition. Graphs.

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

INTRODUÇÃO

Após mais de 150 anos da descoberta dos ácidos nucleicos, o interesse no estudo destas moléculas vem crescendo cada vez mais com o passar dos anos (DAHM, 2009). Devido a sua enorme participação na manutenção e criação da vida em nosso planeta, ganhou notoriedade em diversos campos do conhecimentos, como por exemplo na informática, que apesar de ser muito distinta em essência da área biológica, é de extrema importância na análise da crescente quantidade de dados produzidos.

Esta necessidade da aplicação da informática na biologia criou um novo termo, a Bioinformática, que começou a ser usado no começo dos anos 70 por Hogeweg e Hesper (HOGEWEG, 2011) para definir o estudo de processos computacionais nos sistemas bióticos. Desde então, a bioinformática tornou-se uma parte muito importante de muitas áreas da biologia molecular.

Novas técnicas de bioinformática, como imagem e processamento de sinais, fez possível a extração de resultados significativos a partir de grandes quantidades de dados brutos. No campo da genética e genômica, a bioinformática auxilia no sequenciamento e anotação de genomas e suas mutações observadas. Ela também desempenha um papel importante para o gerenciamento de dados na medicina moderna (JOAQUIM, 2010).

Desde de que Phage-X174 foi sequenciado em 1977 (SANGER, 1977), uma quantidade enorme de organismos vêm sendo sequenciados e armazenados em databases. As sequências 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.

Então, com a crescente quantidade de dados, há muito tempo se tornou impraticável analisar as sequências de DNA manualmente. Programas de computador como o BLAST, são usados rotineiramente para pesquisar sequências e a partir de 2008, mais de 260.000 organismos foram analisados, contendo mais de 190 bilhões de nucleotídeos (BENSON, 2008).

Uma importante parte dessa área do conhecimento é o RNA, um polímero linear composto por quatro tipos diferentes de subunidades nucleotídicas unidas entre si por ligações fosfodiéster, e têm como uma de suas funções servir de molde para a formação de proteínas. Há vários tipos de RNA, como os RNA mensageiros que são codificantes de proteínas e RNA não codificantes que são subdivididos em duas categorias, pequenos e longos (ALBERTS, 2009).

Os RNAs não codificadores estão envolvidos em muitos processos celulares de importância central na maior parte da vida celular. Acredita-se que os ncRNAs mais conservados sejam fósseis moleculares ou relíquias do último ancestral comum universal e do mundo do RNA, e seus papéis atuais permanecem principalmente na regulação do fluxo de informações do DNA para a proteína (JEFFARES, 1998).

O objetivo deste trabalho é implementar uma ferramenta de pré-processamento que processa as sequências pela exclusividade de cada classe de RNA, e, após extrair as medidas dos grafos filtrados, diminuir a quantidade de ruído presente, melhorando a classificação e a diferenciação das mesmas.

MATERIAIS E MÉTODOS

CONJUNTO DE DADOS

Para a realização dos testes foram selecionados dois tipos de RNA, o RNA mensageiro (mRNA) e o RNA não codificante (ncRNA). Isso porque, muitas pesquisas relacionadas com a predição de RNA não codificantes vem ganhando força nas últimas décadas, já que o número de genes RNA não codificante recentemente descobertos cresce exponencialmente. Evidências recentes indicam que esses ncRNAs desempenham papéis críticos em vários processos celulares importantes (AMBROS, 2001).

Os dados de RNAs testados foram os mesmos usados no CPC2 (KANG, 2017) que obteve RNAs de várias espécies, como é possível ser visto na Tabela 1.

Tabela 1 – Conjunto de dados utilizado

Espécies	Classe	Quantidade de Sequências
Arabidopsis	mRNA ncRNA	15931 3853
Human	mRNA ncRNA	6142 12015
Zebrafish	mRNA ncRNA	2344 1528
Fruitfly	mRNA ncRNA	3680 3556
Mouse	mRNA ncRNA	10638 12251
Worm	mRNA ncRNA	3551 9313

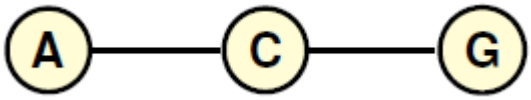
Fonte: Autoria própria (2020).

MÉTODOS

Para a realização da primeira etapa do trabalho, foram criados grafos (redes complexas) a partir de sequências de RNA armazenadas em arquivos do tipo FASTA, aonde sequências de RNA mensageiro e RNA longo não codificantes estão separados, isto é, já se encontram previamente classificadas.

Um grafo pode ser visto como um conjunto de pontos, chamados de vértices. Quando um vértice se junta a outro vértice é formada uma ligação, que é definida como aresta, como na Figura 1, onde A C e G são vértices ligados por arestas.

Figura 1 – Exemplo de um grafo gerado a partir da sequência: ACG.

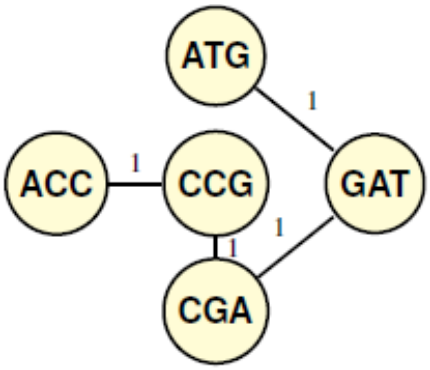


Fonte: Autoria própria (2020).

No desenvolvimento destes grafos era necessário a configuração de dois parâmetros, o Passo e a Palavra. A função do passo é definir a distância que será percorrida na sequência depois que uma ligação foi formada para a formação de uma nova ligação. Já a Palavra se refere a quantos nucleotídeos serão agrupados em cada vértice. Neste projeto, usaremos a mesma configuração do trabalho BASiNET (ITO, 2018), Passo igual a 1 e Palavra igual a 3, como exibido na Figura 2.

Para diminuir o ruído e melhorar a qualidade dos grafos, foi desenvolvida uma ferramenta que identifica as arestas mais exclusivas de cada classe de RNA, para que antes de fazer as medidas do grafo, faça uma filtragem nas arestas evitando, assim, o aparecimento de arestas que são exageradamente repetidas em ambas as classes, o que pode dificultar a diferenciação das classes.

Figura 2 – Exemplo de grafo gerado a partir da sequência: ACCGATG.



Fonte: Autoria própria (2020).

Esta ferramenta se baseia em produzir uma grande rede complexa para cada conjunto de sequências, por exemplo, um grafo referente ao mRNA e outro ao ncRNA. Com isso, é possível produzir uma matriz adjacente de frequências (vide Figura 3) para cada classe, sendo assim executável a separação das arestas mais frequentes de cada uma das classes de RNA.

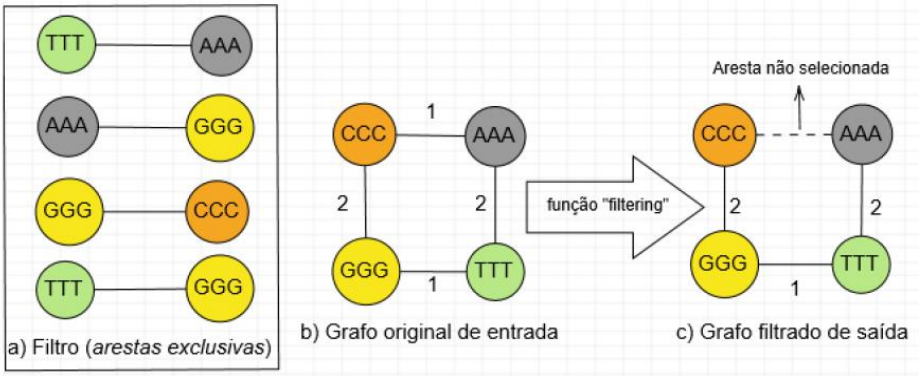
Figura 3 – Exemplo de matriz de adjacências de frequências.

	A	*C*	T	G
A	1	3	5	2
C	3	2	3	4
T	5	3	1	2
G	2	4	2	0

Fonte: Autoria própria (2020).

Com as arestas exclusivas de ambas as classes, é feita uma filtragem dos grafos encontrados, e sua finalidade é comparar todas as arestas entre si, e apenas selecionar aquelas que são exclusivas de cada classe.

Figura 4 – Exemplo de filtragem dos grafos.



Fonte: Autoria própria (2020).

Na Figura 4 a) temos um exemplo de um filtro, com as arestas exclusivas sendo TTT--AAA, AAA--GGG, GGG--CCC e TTT--GGG. Já na Figura 4 b), temos o grafo produzido por uma sequência sem passar por um filtro, comparando este grafo com as arestas do filtro, é possível observar que todas as arestas estão presentes em ambos com exceção da CCC--AAA, por isso esta aresta seria excluída pela função ‘‘filtering’’ como na Figura 4 c).

Como as redes complexas possuem topologias bem definidas que descrevem a estrutura da rede há medidas para abstrair tais características (BOCCALETTI, 2006). Para a abstração de características das redes complexas foram separadas 10 medidas usadas na literatura atual para que seja possível a classificação das sequências.

Com o resultado da filtragem, é possível produzir as medidas relacionadas aos grafos já filtrados e pré processados. Estas medidas serão agrupadas em um amplo dataframe, no qual será possível organizar estas medidas, de modo que, cada linha represente uma sequência, e cada coluna uma medida.

Por meio de um pré processamento, é feito um reajuste nos valores para que eles estejam sempre entre 0 e 1, permitindo uma classificação mais consistente. A Eq. (1) exemplifica como é reajustado os valores, sendo V referente ao valor, Min referente ao valor mínimo da classe e Max ao valor máximo da classe:

$$V(reajustado) = \frac{V - Min}{Max - Min} \quad (1)$$

Para a avaliação do método foi utilizado o pacote “randomForest” (LIAW, 2002) presente no software R (CORE TEAM.R, 2014), que possui ferramentas para classificações. Também foi utilizado o pacote “rfUtilities” (EVANS,2018) que possui uma função para realizar a Validação Cruzada, que é uma técnica de avaliação padrão, e consiste em executar divisões percentuais repetidas. Por exemplo, dividir um conjunto de dados em 10 pedaços, usar um pedaço para teste e avaliar os 9 restantes juntos, fazendo com que diminua a estimativa de erro da classificação. Para este projeto foi usado o numero de pedaços igual a 10.

Todo o processo da metodologia descrita foi executado no software R, para esse fim foram desenvolvidas 6 funções que fazem desde a leitura dos arquivos no formato Fasta, como também a filtragem e a organização das medidas extraídas do grafo no dataframe.

RESULTADOS E DISCUSSÃO

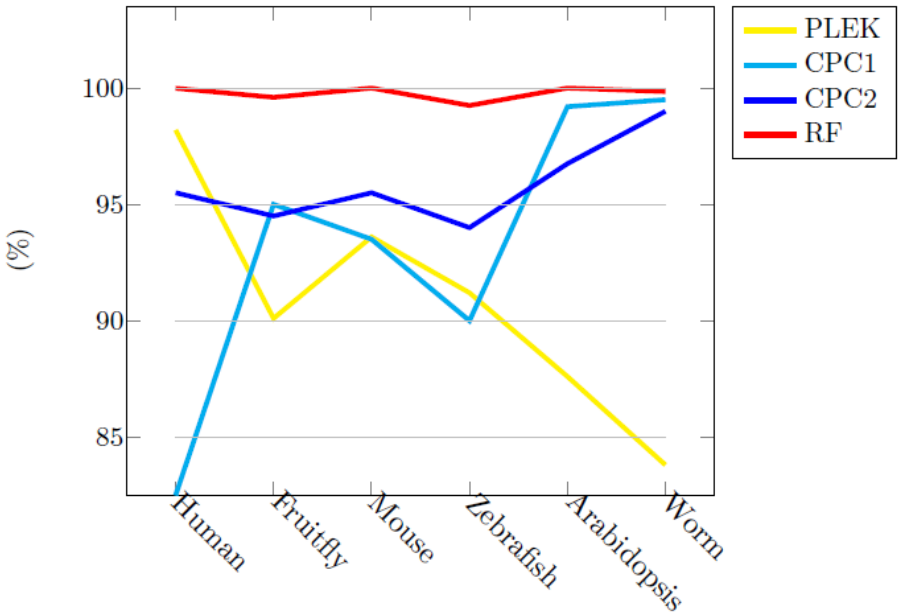
A Tabela 2 apresenta os resultados das classificações referentes aos mRNAs e os ncRNAs, analisados pela ferramenta Random Forest presente no pacote "randomForest" (LIAW, 2002) no R. Para ter um parâmetro de comparação os resultados foram comparados com os classificadores CPC1, CPC2 (KANG, 2017) e PLEK (LI, 2014) a base de dados usada foi a mesma utilizada no trabalho (KANG, 2017).

Tabela 2 – Taxas de acerto na classificação das sequências de mRNA e ncRNA por meio do método prosposto para diferentes espécies.

Espécies	Taxa de acerto (RF)	
	mRNA	ncRNA
Arabidopsis	6142 (100 %)	12005 (99.9%)
Human	3665 (99,5%)	3551 (99,7%)
Zebrafish	10638 (100%)	12251 (100%)
Fruitfly	2313 (98,7%)	1528 (100%)
Mouse	15930 (100%)	3834 (100%)
Worm	3523 (99,7%)	9312 (100%)

Fonte: Autoria própria (2020).

Figura 5 – Gráfico para comparação dos classificadores por espécie.



Fonte: Autoria própria (2020).

Na Tabela 2 é possível observar que todos os resultados das classificações apresentam acurácia maior que 98,6%. Na Tabela 3, nota-se que a classificação pelo método apresenta resultados superiores na classificação das duas classes (mRNA e ncRNA). A Figura 5, que apresenta a comparação dos classificadores CPC1, CPC2 e PLEK com este trabalho, demonstra que a metodologia de filtragem de arestas possui uma taxa de acurácia consideravelmente maior para todas as espécies analisadas, com excecão das espécies Arapidopsis e Worm onde a acurácia foi semelhante com os métodos de CPC1 e CPC2.

Tabela 3 – Comparação dos classificadores por classe e média total.

	PLEK	CPC1	CPC2	RF
mRNA	79,9%	99,50%	95,20%	99,65%
ncRNA	92.31%	87,30%	97,14%	99,92%
Média Total	**90.75%**	**93,20%**	**96,1%**	**99,78%**

Fonte: Autoria própria (2020).

Os resultados indicaram uma alta distinção entre as classes de RNA devido a filtragem das arestas exclusivas, já que em média e por espécie, os acertos deste método foram superiores em comparação com os classificadores CPC1, CPC2 e PLEK, como mostram a Tabela 3 e a Figura 5.

CONCLUSÃO

Fica evidente devido aos resultados observados e a comparação realizada com outros métodos com objetivos semelhantes, que o método utilizado neste trabalho se mostrou adequado e com resultados superiores. A classificação de RNA mensageiros e RNA não codificantes usando a filtragem das arestas exclusivas e a produção e extração de medidas dos grafos, fez possível observar que a implementação da ferramenta de pré-processamento foi eficaz para maior diferenciação entre as classes.

A aplicação e estudo mais aprofundado deste método pode levar a um melhor entendimento do RNA não codificante. Assim como este método funciona para RNA, supostamente também pode ser aplicado a outras sequências biológicas como as de DNA, que possuem uma estrutura semelhante com os RNA.

AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa de Iniciação Científica (Projeto de Pesquisa 406099 /2016-2) concedida ao acadêmico Murilo Montanini Breve.

REFERÊNCIAS

DAHM, RALF. **“Friedrich Miescher and the Discovery of DNA.”** Developmental Biology, vol. 278, no. 2, Feb. 2005, pp. 274–88. DOI.org (Crossref), doi:10.1016/j.ydbio.2004.11.028.

HOGEWEG, P. **The Roots of Bioinformatics in Theoretical Biology.**PLOSComputational Biology, Public Library of Science, v. 7, n. 3, p. 1–5, mar. 2011. DOI:10.1371/journal.pcbi.1002021.

JOAQUIM, L. M.; EL-HANI, C. N**. A genética em transformação: crise e revisão doconceito de gene.** pt.Scientiae Studia, scielo, v. 8, p. 93–128, mar. 2010. ISSN 1678-3166.

SANGER, F. et al. **Nucleotide sequence of bacteriophage ΦX174 DNA.** Nature, v. 265,n. 5596, p. 687–695, fev. 1977. ISSN 1476-4687. DOI:10.1038/265687a0.

JEFFARES, D. C.; POOLE, A. M.; PENNY, D. **Relics from the RNA World**. Journal of Molecular Evolution, v. 46, n. 1, p. 18–36, jan. 1998. ISSN 1432-1432. DOI:10.1007/PL00006280.

BENSON, D. A. et al. **GenBank.** eng. Nucleic acids research, Oxford University Press,v. 36, Database issue, p. d25–d30, jan. 2008. gkm929[PII]. ISSN 1362-4962. DOI:10.1093/nar/gkm929.

WANG, Z.; GERSTEIN, M.; SNYDER, M**. RNA-Seq: a revolutionary tool fortranscriptomics**.Nature Reviews Genetics, v. 10, n. 1, p. 57–63, jan. 2009. ISSN1471-0064. DOI:10.1038/nrg2484.

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

AMBROS, V. **microRNAs: Tiny Regulators with Great Potential**.Cell, v. 107, n. 7,p. 823–826, 2001. ISSN 0092-8674. DOI: https://doi.org/10.1016/S0092-8674(01)00616-X.

KANG, Y.-J. et al. **CPC2: a fast and accurate coding potential calculator based onsequence 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.

ITO, E. A. et al. BASiNET—**BiologicAl Sequences NETwork: a case study on coding andnon-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.

LIAW, A.; WIENER, M. **Classification and Regression by randomForest.** R News, v. 2,n. 3, p. 18–22, 2002.

EVANS, J. S.; MURPHY, M. A. **rfUtilities.** [S.l.], 2018. R package version 2.1-3.

LI, A., Zhang, J., Zhou, Z.: **Plek: a tool for predicting long non-coding rnas and messenger rnas based on an improved k-mer scheme.** BMC Bioinformatics 15(1), 311 (Sep 2014). https://doi.org/10.1186/1471-2105-15-311, https://doi.org/10.1186/1471-2105-15-311.

BOCCALETTI, S. et al. **Complex networks: Structure and dynamics.**Physics Reports,v. 424, n. 4, p. 175–308, 2006. ISSN 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2005.10.009.

LIAW, A.; WIENER, M. **Classification and Regression by randomForest.**R News, v. 2,n. 3, p. 18–22, 2002.

CORE TEAM.R: **A Language and Environment for Statistical Computing.** Vienna,Austria, 2014.