

Extração de características e classificação de sequências de RNAs

Fabrício M. Lopes e Matheus H. Pimenta-Zanon

fabricao@utfpr.edu.br e matheus.pimenta@outlook.com

UTFPR-CP

Grupo de Pesquisa em Bioinformática e
Reconhecimento de Padrões

bioinfo-cp@utfpr.edu.br



EPB 2022
III ESCOLA PARANAENSE
DE BIOINFORMÁTICA



Escola Paranaense de Bioinformática (EPB)

09-AGO-2022

Agenda

- 1 SARS-CoV-2
- 2 Materials and Methods
- 3 Results
- 4 Conclusion

SARS-CoV-2

SARS-CoV-2 - Introduction

- At the end of 2019 an unexplained pneumonia emerged in Wuhan and quickly becoming a pandemic of **SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2)** causing **Corona Virus Disease 2019 (COVID-19)** [Worobey, 2021];
- The World Health Organization declared a coronavirus disease **pandemic 2019 (COVID-19) on March 11, 2020** [Zhou et al., 2020]. Since then, the number of patients infected with SARS-CoV-2 has increased rapidly and spread around the world.

SARS-CoV-2 - VOIs and VOCs

- Currently more than 480 million cases and 6 million deaths have been confirmed worldwide;
- SARS-CoV-2 has also evolved and due to this new variants are continuously identified and some of these variants are considered variants of interest (VOIs) or variants of concern (VOCs).

SARS-CoV-2 - VOIs and VOCs

- **Variant of interest (VOI):** has genetic variations that affect virus characteristics and circulating widely;
- **Variant of concern (VOC):** increases transmissibility or exhibits some prejudicial factor in the epidemiology of COVID-19 [WHO, 2022].

SARS-CoV-2 - Variants

- Until February 2022, World Health Organization [WHO, 2022] reported Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2) and Omicron (B.1.1.529) as **variants of concern** and Lambda (C.37) and Mu (B.1.621) as **variants of interest**;
- The number of SARS-CoV-2 sequences available for analysis in online repositories is enormous, there are millions of sequences deposited by thousands of researchers worldwide [Franceschi et al., 2021].

Sequence Analysis

- Conventional bioinformatics tools used for the analysis of biological sequences are alignment-based and cannot handle large numbers of data due to the high computational complexity involved in the alignment process [Perico et al., 2021];
- An alternative to using tools that depend on alignment is the use of machine learning, data mining, and pattern recognition techniques.

Sequence Analysis

- Alignment-free tools in sequence analysis are applied to classification problems for different types of biological sequences, such as classification of RNAs [Li et al., 2014, Ito et al., 2018, Kang et al., 2017, Zheng et al., 2021, Breve et al., 2022];
- Other computational methods are applied to inference of phylogenetic trees [De Pierri et al., 2020], taxonomic classification and other applications [Zielezinski et al., 2017].

Sequence Analysis - Proposed Approach

- In this context, we investigate the classification and analysis of the variants of concern relative to the SARS-CoV-2 reference sequence from BASiNETEntropy [Breve et al., 2022];
- BASiNETEntropy is an alignment-free feature extraction method that uses only biological sequences, which are mapped into complex networks.
- These networks are characterized by considering their topological measurements as features and, applied in the classification of the SARS-CoV-2 variants.

Materials

Materials

- The SARS-CoV-2 variants of concern corresponding to the genomes available from NCBI were adopted;
- The PANGO¹ [Rambaut et al., 2020] nomenclature was used for the selection of the variants;
- The variants Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2) and Omicron (B.1.1.529) were adopted, besides the genomic sequences of the SARS-CoV-2 variants, the Wuhan reference (NC_045512.2)² was also considered.

¹<https://cov-lineages.org/>

²<https://www.ncbi.nlm.nih.gov/nuccore/1798174254>

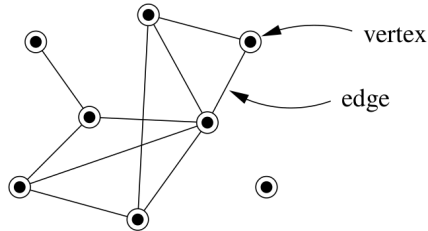
Materials - Data pre-processing

- All duplicated and incomplete sequences were removed by considering the SeqKit [Shen et al., 2016];
- The variant with the smallest number of samples was adopted as a lower bound in order to define the number of sequences for each class (variant);
- As a result, 220 sequences from each variant of the SARS-CoV-2 were considered for the analyses.

Methods

Complex Networks

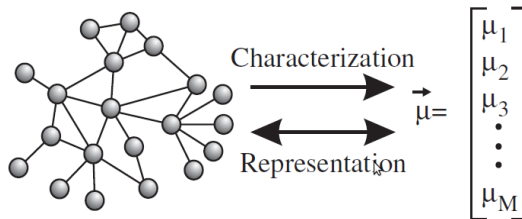
- A **network** is a graph with a set of vertices connected through edges;
- Complex network theory can be understood as the study of networks (structure and function), i.e. the relationships of their vertices [Newman, 2003].



[Newman, 2003]

Complex Networks - Characterization

- Complex networks have distinct topologies and well-defined properties [Boccaletti et al., 2006, Costa et al., 2007].
- Complex networks can be characterised in terms of topological measurements.



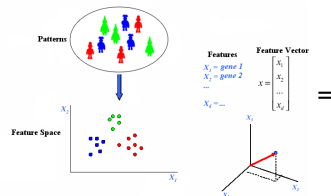
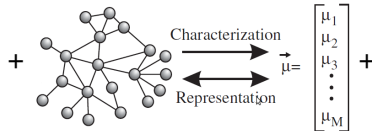
[Costa et al., 2007]

Complex Networks Measurements

- It is possible to extract measures that characterize the topology of the network, such as:
 - Average Betweenness Centrality
 - Cluster Coefficient
 - Average Path Length
 - Assortativity
 - Average, maximum, minimum degree
 - Frequency of motifs with size 3 and 4, etc.

BASiNET - Biological Sequences NETwork

ATGTATCCAGGTAGTGGACGTTACACCTACAAACGCTGGTGTAAATAATGGCTACCAA
 CGGCCATGGCTCCTCCACCTAACCAAGCATATGGACAGCAATATGGTCAGCAATATGAA
 CAGCAGTATGGACAGCAATATGGGCAACAAATGATCAGCAATTCAGTCAGCAATATGCT
 CCACCAACAGGCTCCTCCCTTATGGCTTATTAACAGGCTGTGTATCCCCCTCAATTC
 CAGCAGGAACAGGCAAGGCAATTAAGCAACGGCTACAAATCCTAATGTAAACGCA
 TCCAATATGTACGGTCCACCCAGAAATGTCACTTACCTCCAGCTCAAAACAGAACTAT
 CAAGGTACAGCAACCACTTATCAGTATCTCAATCTACTGGGCTAGAAAGGCTTGAAT
 ATCGGTATAAATCAGATAGTTCAAAAAATCAACTGCGTGGTGTATCAATGATGCTCAT
 AACATCTTCAACTTTTGAATAGGTGGTACAGTTCAGATGACATTTGTCATATTA
 ACTGATGATCAGAAAGATTTGGTCAGGTTCCCACTAGGGCTAAATATGATTAGGGCCATG
 CAATGGTTGGTCAAGGATCGCAACCAATGATTTGGTCTTCATATTCCTGGACAT
 GGTGGCCAAATCGAAGATTTGGATGGGCAAGATGGGATGGATGATTTATATAT
 CCGTCCGATTTCCGAACCTCAAGGCGCAATATTCGACGATGAATGACGATATATGGTGG
 AAGCCCTTACAAAGGTTAGACCTAACGAGCATTTGTGACTCTTGTCAATTCGGGTACA
 GTGTGGATCTTCCATATACCTATTTCTACTAAGGTTATATTAAGGAGCCCAATATTGG
 AAGGATGTTGGCCAGATGGCTCGCAAGCGCTATTTCATATGCCACAGAAACAGGGCT
 GCTTTGATTTGGTTCTTATAGTTCTATATTCAGACGTTAAAGGAGGTATGGSCAATAAT
 GTGATAGAGAACGCTGAGACAGATCAAAATCTCAGCAGCAGATGTTTATGTTATCA
 GGTTCGAAGGATAATCAAACTTCTGACAGTGTGTGAAAGTGGGCAAAATACAGGTGCA
 ATGTCCACCGCTTCATCAAGGTTATGACTTTACAAACAGCAATCATATTTATCTCT
 TCACAGAACATGAGGAAAGATTTGGTGGTAAATTTCTCAAAAAACCAATATCATCG
 TCACACCTTATGACGTAAATTCGCAATTTATATGAG



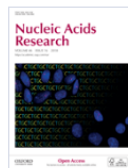
BASiNET—Biological Sequences NETwork: a case study on coding and non-coding RNAs identification

Eric Augusto Ito, Isaque Katahira, Fábio Fernandes da Rocha Vicente,
 Luiz Filipe Protasio Pereira, Fabrício Martins Lopes ✉

Nucleic Acids Research, Volume 46, Issue 16, 19 September 2018, Page e96,

<https://doi.org/10.1093/nar/gky462>

Published: 05 June 2018 Article history ▼

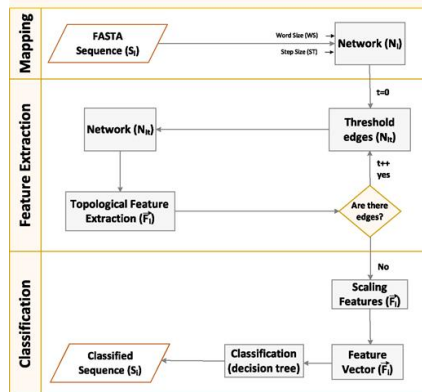


BASiNET - BiologicAI Sequences NETwork

- It is a feature extraction method for the classification of biological sequences based on the measurements of complex networks[Ito et al., 2018]:
 - Considers the neighbourhood of k-mers to build a network;
 - Extract topological measurements for network characterization and feature vector composition;
 - The feature vector is used in the classification of the sequences.

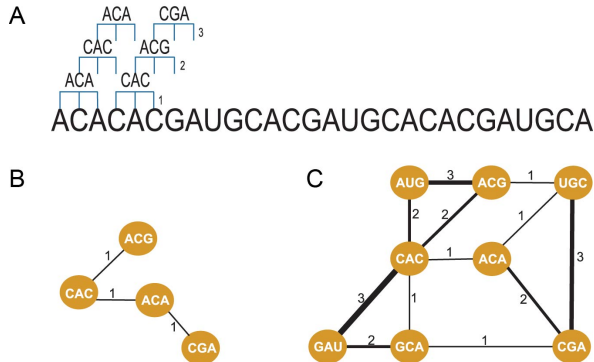
BASiNET - BiologicAI Sequences NETwork

- It explores existing patterns in the input sequences considering three main steps:



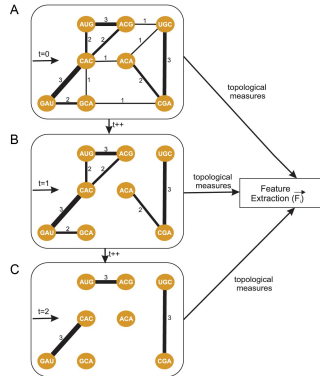
Mapping

- Mapping the sequence onto a network (graph) with weighted edges:



Feature extraction

- Feature extraction is iterative, considering the weights of edges for the extraction of topological measurements:



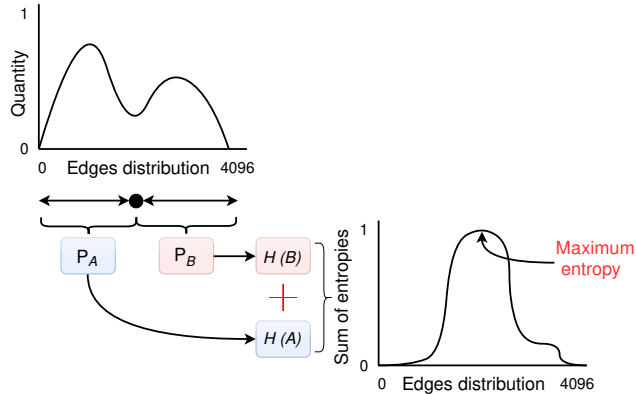
Improved Feature extraction

- Iterative feature extraction produces about 2,000 features and has a high computational cost;
- Then a filter step is proposed that “learns” which edges in the network are most informative for each class;
- This filter is based on the maximum entropy (ME) principle [Jaynes, 1957].

Maximum Entropy (ME)

- The ME principle can be applied to measure the amount of uncertainty contained in a probability distribution [Guiasu and Shenitzer, 1985];
- By considering the edges frequency (i.e., weights), a histogram is produced.
- The aim is to find the threshold (s) that maximizes the sum of the entropies of two distinct parts (informational and non-informational) edges of each class.

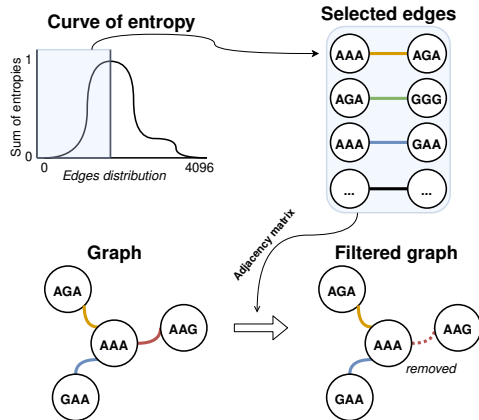
BASiNETEntropy - Entropy Maximization



BASiNETEntropy

- The BASiNETEntropy [Breve et al., 2022] proposes a filter for the network vertices in order to identify what edges are important and what are not important for each class.
- The ME principle is applied to select the most informational edges for each class, composing a single network.
- The BASiNETEntropy is freely available at <https://cran.r-project.org/web/packages/BASiNETEntropy/>.

BASiNETEntropy - Filtering



BASiNETEntropy - Feature Extraction

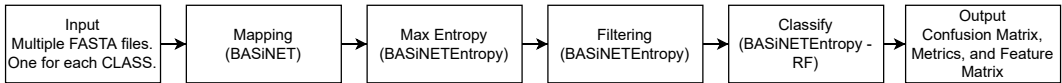
- A single filtered complex network is produced for each class and the measurements are extracted for its characterization;
- The iterative step is removed, reducing the amount of extracted features and the computational cost involved;
- Maintaining the information contained in the networks and removing noise (non-informational edges).

BASiNETEntropy - Feature Extraction

- The filtered complex networks are characterized by their topological measurements;
- 10 topological measurements commonly used in the literature are adopted: assortativity (ASS), average degree (DEG), maximum degree (MAX), minimum degree (MIN), average betweenness centrality (BET), clustering coefficient (CC), average short path length (ASPL), average standard deviation (SD), frequency of motifs with size 3 (MT3) and frequency of motifs with size 4 (MT4).

BASiNETEntropy - Classification Overview

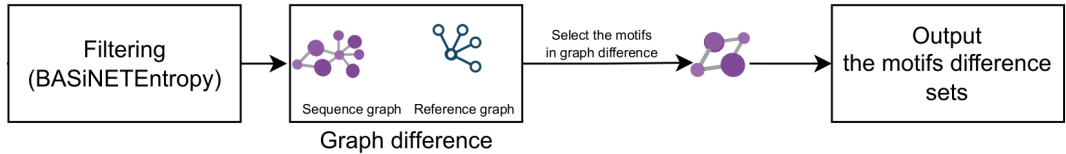
Classify



BASiNETEntropy - Motif Analysis

- As an additional analysis, the networks generated by the SARS-CoV-2 variants were compared with the Wuhan network as a reference;
- The difference of graphs was performed by considering the difference between the adjacency matrices of the reference (G_{ref}) and variant (G_{seq}).

BASiNETEntropy - Motif Analysis



Results

Classification - Data

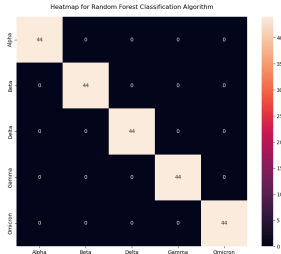
- The sequences from NCBI were adopted considering 220 genomic sequences from each of the SARS-CoV-2 variants of concern: Alpha, Beta, Gamma, Delta and Omicron;
- In order to avoid influences or biased behaviors in the classification methods a re-scaling step is applied, since the extracted topological measures have different numerical scales in relation to each other, defining a closed interval $[0, 1]$.

Classification - Algorithm

- The Random Forest [Breiman, 2001] classification algorithm was adopted in order to evaluate the behavior of the features;
- The parameter values used in the Random Forest algorithm were set as defaults. The 10-fold cross-validation method was also adopted.

Classification - First Experiment

- The first experiment was performed in order to evaluate the classification of the SARS-CoV-2 variants of concern. An optimal classification of all variants was achieved.

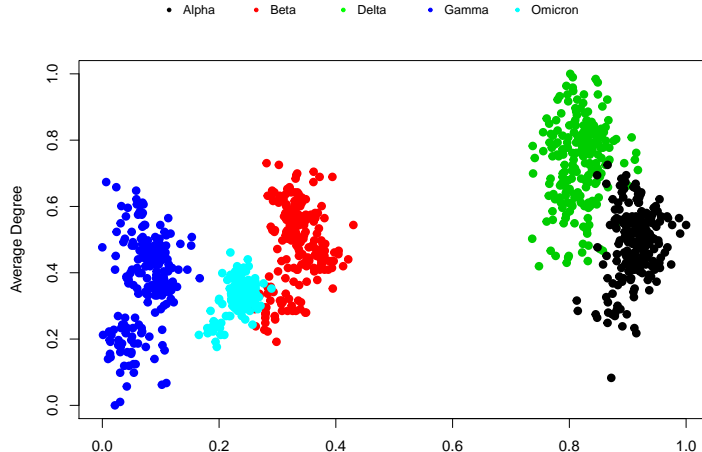


Method	Accuracy	Recall	Precision	F1-score
BASiNETEntropy	1	1	1	1

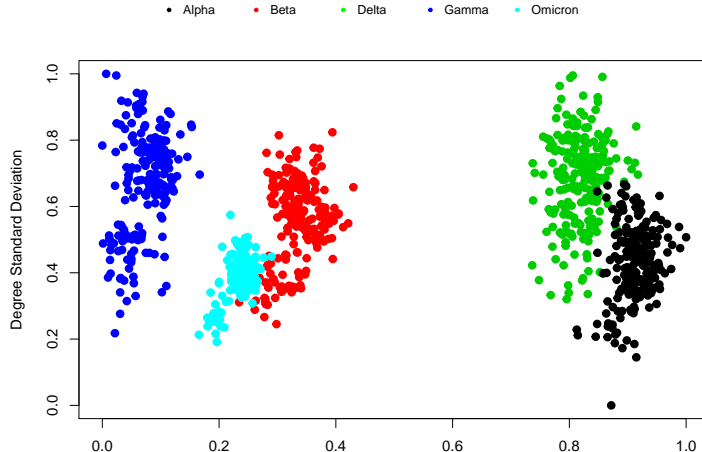
Classification - Features

- Considering the cross-validation, it was analysed which features produced a suitable feature space for classification;
- 3 features were identified: Assortativity, Average Degree and Standard Deviation of degrees.

Classification - Features ASS and DEG



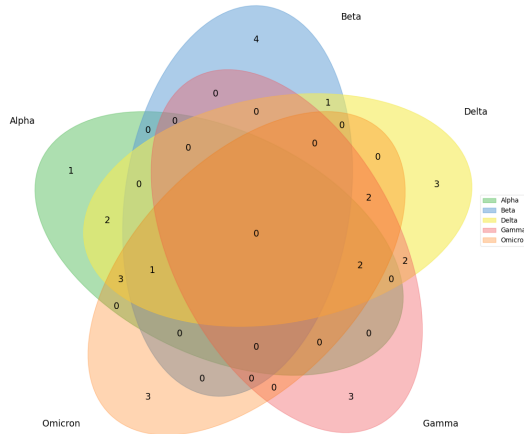
Classification - Features ASS and SD



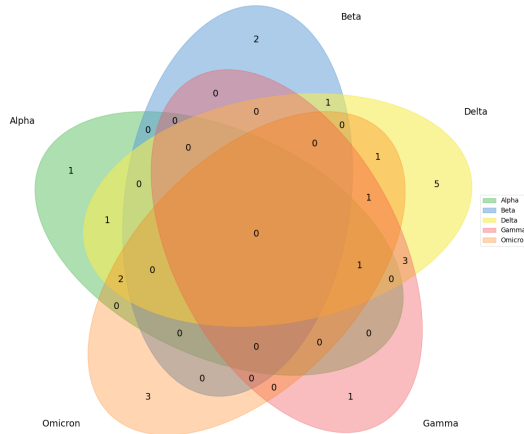
Motif Analysis

- The identification of motifs considering each variant of concern and the SARS-CoV-2 reference (Wuhan) were performed;
- Thus, a difference graph was produced by considering each variant of concern and the SARS-CoV-2 reference;
- As a result, motifs of size 3 and 4 that are unique to each of the variants are obtained.

Motif Analysis - Size 3



Motif Analysis - Size 4

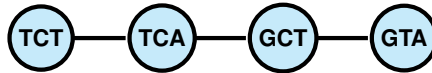


Motif Analysis - Unique

- It were identified unique and shared motifs patterns among SARS-CoV-2 variants;
- For instance, Alpha variant presents only one unique motif of size 3 (a) and one of size 4 (b):



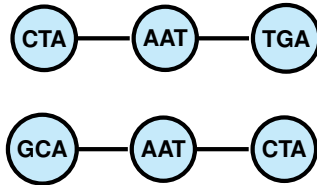
(a)



(b)

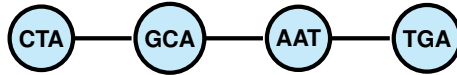
Motif Analysis - Shared - Size 3

- No motifs shared with the 5 variants were identified;
- 2 motifs of size 3 were identified between the Alpha and Delta and Gamma and Omicron variants:



Motif Analysis - Shared - Size 4

- One motif of size 4 between the Alpha and Delta and Gamma and Omicron variants:



Conclusion

- Classification of SARS-CoV-2 variants is a computationally arduous task because of the high computational complexity involved in the alignment algorithms.
- In this work, BASiNETEntropy, an alignment-free method for the classification of variants of concern has been considered;
- The results show an optimal classification of all variants of concern. Besides, it was possible to identify patterns of k-mers that are unique and shared among variants.




Directions

- The mapping and analysis of the k-mers patterns in the genomic sequences, attempting to map the location of the identified patterns in the SARS-CoV-2 sequences.
- The analysis of patterns regarding its semantics and biological functionality is suggested as further work.
- BASINETEntropy is open source and it is freely available at:
<https://cran.r-project.org/web/packages/BASiNETEntropy/>.



Thank you!

`fabricio@utfpr.edu.br` e `matheus.pimenta@outlook.com`



References I

-  Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D. U. (2006).
Complex networks: Structure and dynamics.
Physics Reports 424, 175–308.
-  Breiman, L. (2001).
Random Forests.
Machine Learning 45, 5–32.
-  Breve, M. M., Pimenta-Zanon, M. H. and Lopes, F. M. (2022).
BASiNETEntropy: an alignment-free method for classification of biological sequences through complex networks and entropy maximization.



References II

-  Costa, L. d. F., Rodrigues, F. A., Travieso, G. and Villas-Boas, P. R. (2007). Characterization of complex networks: a survey of measurements. *Advances in Physics* 56, 167–242.
-  De Pierri, C. R., Voyceik, R., Santos de Mattos, L. G. C., Kulik, M. G., Camargo, J. O., Repula de Oliveira, A. M., de Lima Nichio, B. T., Marchaukoski, J. N., da Silva Filho, A. C., Guizelini, D., Ortega, J. M., Pedrosa, F. O. and Raittz, R. T. (2020). SWeeP: representing large biological sequences datasets in compact vectors. *Scientific Reports* 10, 91.



References III

-  Franceschi, V. B., Ferrareze, P. A. G., Zimerman, R. A., Cybis, G. B. and Thompson, C. E. (2021).
Mutation hotspots and spatiotemporal distribution of SARS-CoV-2 lineages in Brazil, February 2020-2021.
Virus Research 304, 198532.
-  Guiasu, S. and Shenitzer, A. (1985).
The principle of maximum entropy.
The mathematical intelligencer 7, 42–48.



References IV

-  Ito, E. A., Katahira, I., Vicente, F. F., Pereira, L. P. and Lopes, F. M. (2018). BASiNET - BiologicAI Sequences NETwork: a case study on coding and non-coding RNAs identification. *Nucleic Acids Research* , gky462.
-  Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Phys. Rev.* *106*, 620–630.




References V

-  Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L. and Gao, G. (2017).
CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features.
Nucleic Acids Research 45, W12–W16.
-  Li, A., Zhang, J. and Zhou, Z. (2014).
PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme.
BMC Bioinformatics 15, 311.




References VI

-  Newman, M. E. J. (2003).
The Structure and Function of Complex Networks.
SIAM Review 45, 167–256.
-  Perico, C. P., De Pierri, C. R., Neto, G. P., Fernandes, D. R., Pedrosa, F. O., de Souza, E. M. and Raittz, R. T. (2021).
Genomic landscape of SARS-CoV-2 pandemic in Brazil suggests an external P.1 variant origin.
preprint Epidemiology.


References VII

-  Rambaut, A., Holmes, E. C., O'Toole, A., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L. and Pybus, O. G. (2020).
A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology.
Nat. Microbiol. *5*, 1403–1407.
-  Shen, W., Le, S., Li, Y. and Hu, F. (2016).
SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.
PLOS ONE *11*, e0163962.
-  WHO (2022).
SARS-CoV-2 variants of concern and variants of interest.
Technical report World Health Organization.

References VIII

-  Worobey, M. (2021).
Dissecting the early COVID-19 cases in Wuhan.
Science 374, 1202–1204.
-  Zheng, H., Talukder, A., Li, X. and Hu, H. (2021).
A systematic evaluation of the computational tools for lncRNA identification.
Briefings in Bioinformatics 22, bbab285.
-  Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E. C., Hughes, A. C., Bi, Y. and Shi, W. (2020).
A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein.
Current Biology 30, 2196–2203.e3.

References IX

-  Zielezinski, A., Vinga, S., Almeida, J. and Karlowski, W. M. (2017).
Alignment-free sequence comparison: benefits, applications, and tools.
Genome Biology 18, 186.