MathFeature: Feature Extraction Package for Biological Sequences Based on Mathematical Descriptors

Robson P. Bonidia[1], Danilo S. Sanches[2], and André C.P.L.F. de Carvalho[1]

[1] Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos 13566-590, Brazil
[2] Department of Computer Science, Federal University of Technology - Paraná, UTFPR, Cornélio Procópio 86300-000, Brazil.

Abstract

Machine learning algorithms have been very successfully applied to extract new and relevant knowledge from biological sequences. However, the predictive performance of these algorithms is largely affected by how the sequences are represented. Thereby, the main challenge is how to numerically represent a biological sequence in a numeric vector with an efficient mathematical expression. Several feature extraction techniques have been proposed for biological sequences, where most of them are available in feature extraction packages. However, there are relevant approaches that are not available in existing packages, techniques based on mathematical descriptors, e.g., Fourier, entropy, and graphs. Therefore, this paper presents a new package, named MathFeature, which implements mathematical descriptors able to extract relevant information from biological sequences. MathFeature provides 20 approaches based on several studies found in the literature, e.g., multiple numeric mappings, genomic signal processing, chaos game theory, entropy, and complex networks. MathFeature also allows the extraction of alternative features, complementing the existing packages.

Availability and implementation: MathFeature is freely available at https://bonidia.github.io/MathFeature/ or https://github.com/Bonidia/MathFeature

Contact: bonidia@usp.br, rpbonidia@gmail.com

1 Background

In the last years, Machine learning (ML)-based tools have been developed for various genomics, transcriptomics, and proteomics problems [1]. Nevertheless, for the successful application of ML algorithms, relevant features need to be extracted, to represent the main aspects of the original sequence. In [2, 3], the authors address the relevance of using an appropriate mathematical expression to extract features from biological sequence data. Based on this, many techniques have been developed and investigated to extract numerical representative information from sequences [4, 5]. In which, the main challenge is how to numerically represent a biological sequence in a numeric vector. To deal with this challenge, several of these features were made available in public software packages, such as: PseKNC-General [5], PseKNC [4], Pse-in-One [3], DNAshapeR [6], repDNA [7], repRNA [8], Pse-in-One 2.0 [9], BioSeq-Analysis [10], iFeature [11], PyFeat [12], Seq2Feature [13], iLearn [14].

These previous studies have produced tools, packages, web servers, or toolkits. However, each of them was limited regarding the mathematical features made available (e.g., multiple numeric mappings, Fourier, chaos game theory, entropy, and complex networks). Therefore, in this work, we present an open-source Python package, named MathFeature, which provides in a single environment, many of the mathematical features previously proposed for feature extraction from biological sequences. MathFeature provides 20 mathematical approaches, here named descriptors, organized into five categories. To our best knowledge, MathFeature is the first package to provide such a large and comprehensive set of mathematical feature extraction descriptors for biological sequences.

2 Package Description

In MathFeature, the descriptors are applied to a sequence according to the pipelines illustrated by Figure 1. In Table 1, we organize the 20 descriptors into 5 groups, according to how they work. Furthermore, we have developed a user-friendly tool that covers several mathematical descriptors. Next, we also classified the main aspects of each group, as follow:

- **Numerical Mapping:** Several sequence analysis studies require converting a biological sequence to a numeric sequence. Previous studies have proposed different descriptors for such, which are able to represent important aspects of these sequences. This group contains 7 descriptors for numerical mapping: Voss [15] (known as binary mapping), Integer [16], Real [17], Z-curve [18], EIIP [19], Complex Numbers [20, 21] and Atomic Number [22, 23].

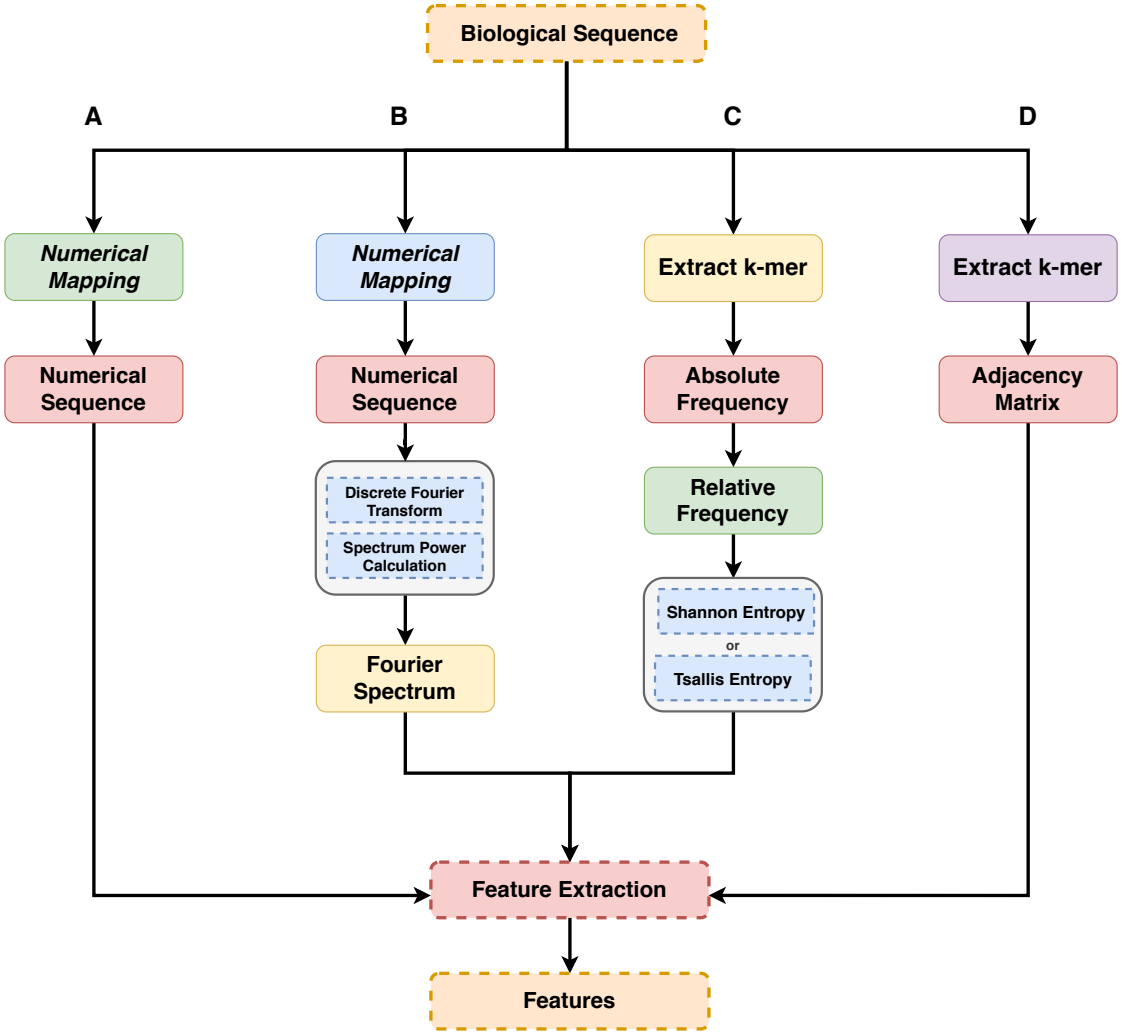

Figure 1: Pipeline of descriptors calculated by MathFeature. **A:** Numerical Mapping and Chaos Game Representation; **B:** Fourier Transform; **C:** Entropy; **D:** Complex Networks.

- **Chaos Game Representation (CGR):** This approach is also a mapping, but scale-independent and iterative for geometric representation of DNA sequences [24]. There are several variations of this group, in this package, we included classical CGR [24, 25], frequency CGR [26], and CGR signal with Fourier Transform (FT) [25].
- **Fourier Transform:** This group extracts sequence features based on Genomic Signal Processing, using FT, a widely applied approach in several biological sequence analysis problems [27, 25, 23, 28]. To implement GSP techniques, we use all numerical mappings. A mathematical exploration can be seen in [28].
- **Entropy:** Various studies have applied concepts from information theory for sequence feature extraction, mainly Shannon's entropy [29, 30]. However, according to [31], another entropy-based measure has been successfully explored in several studies, Tsallis entropy [32], proposed to generalize the Boltzmann/Gibbs's traditional entropy. This group includes these two descriptors [28].
- **Graphs:** This group has descriptors based on graph theory, a currently very active research area, with relevant results in biology sequences [33, 34]. In addition, the descriptors included in this group were proposed in [35] and also explored in [28].

3 Results

In this study, we presented MathFeature, a novel feature extraction package based on mathematical features. MathFeature provides 20 descriptors to numerically represent biological sequences, organized in 5 groups: multiple numeric mappings, Fourier transform, chaos game theory, entropy, and complex networks. Our main purpose is to complement

Table 1: Descriptors calculated by MathFeature for DNA, RNA, and Protein sequences.

Descriptor groups	Descriptor	Dimension	Biological Sequence
Numerical Mapping	Binary	$L \cdot 4$	DNA/RNA
	Z-curve	$L \cdot 3$	DNA/RNA
	Real	L	DNA/RNA
	Integer	L	DNA/RNA
	EIIP	L	DNA/RNA
	Complex Number	L	DNA/RNA
	Atomic Number	L	DNA/RNA
Chaos Game	Chaos Game Representation	$L \cdot 2$	DNA/RNA
	Frequency Chaos Game Representation	$L - k + 1$	DNA/RNA
	Chaos Game Signal (with Fourier)	19	DNA/RNA
Fourier Transform	Numerical Mapping + Fourier	19	DNA/RNA
Entropy	Shannon	k	DNA/RNA/Protein
	Tsallis	k	DNA/RNA/Protein
Graphs	Complex Networks (Codons)	$12 \cdot t$	DNA/RNA/Protein
	Complex Networks (different k-mers)	$12 \cdot t \cdot k$	DNA/RNA/Protein

L = length of the longest sequence, k = frequencies of k-mer, t = threshold - number of subgraphs.

the existing feature extraction packages, providing alternatives to extract new and relevant features using mathematical descriptors. These descriptors have been previously applied to biological sequences with relevant and robust results, e.g., [28] (performances (ACC) between 0.8901-0.9606), [25], [27], [30], and [35]. MathFeature is freely available at https://github.com/Bonidia/MathFeature and its documentation is provided at https://bonidia.github.io/MathFeature/.

Acknowledgements

The authors would like to thank USP and CAPES - Finance Code 001 and PROEX-11919694/D for the financial support for this research. *Conflict of Interest:* none declared.

References

[1] Wellison Jarles da Silva Diniz and Fernanda Canduri. Bioinformatics: an overview and its applications. *Genet Mol Res*, 16(1), 2017.

[2] Kuo-Chen Chou. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology*, 273(1):236–247, 2011.

[3] Bin Liu, Fule Liu, Xiaolong Wang, Junjie Chen, Longyun Fang, and Kuo-Chen Chou. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research*, 43(W1):W65–W71, 05 2015.

[4] Wei Chen, Tian-Yu Lei, Dian-Chuan Jin, Hao Lin, and Kuo-Chen Chou. Pseknc: A flexible web server for generating pseudo k-tuple nucleotide composition. *Analytical Biochemistry*, 456:53 – 60, 2014.

[5] Wei Chen, Xitong Zhang, Jordan Brooker, Hao Lin, Liqing Zhang, and Kuo-Chen Chou. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics*, 31(1):119–120, 09 2014.

[6] Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, 12 2015.

[7] Bin Liu, Fule Liu, Longyun Fang, Xiaolong Wang, and Kuo-Chen Chou. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, 31(8):1307–1309, 12 2014.

[8] Bin Liu, Fule Liu, Longyun Fang, Xiaolong Wang, and Kuo-Chen Chou. reprna: a web server for generating various feature vectors of rna sequences. *Molecular Genetics and Genomics*, 291(1):473–481, 2016.

[9] Bin Liu, Hao Wu, Kuo-Chen Chou, et al. Pse-in-one 2.0: an improved package of web servers for generating various modes of pseudo components of dna, rna, and protein sequences. *Natural Science*, 9(04):67, 2017.

[10] Bin Liu. Bioseq-analysis: a platform for dna, rna and protein sequence analysis based on machine learning approaches. *Briefings in bioinformatics*, 20(4):1280–1294, 2017.

[11] Zhen Chen, Pei Zhao, Fuyi Li, André Leier, Tatiana T Marquez-Lago, Yanan Wang, Geoffrey I Webb, A Ian Smith, Roger J Daly, Kuo-Chen Chou, and Jiangning Song. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 34(14):2499–2502, 03 2018.

[12] Rafsanjani Muhammod, Sajid Ahmed, Dewan Md Farid, Swakkhar Shatabda, Alok Sharma, and Abdollah Dehzangi. PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences. *Bioinformatics*, 35(19):3831–3833, 03 2019.

[13] Rahul Nikam and M Michael Gromiha. Seq2Feature: a comprehensive web-based feature extraction tool. *Bioinformatics*, 35(22):4797–4799, 05 2019.

[14] Zhen Chen, Pei Zhao, Fuyi Li, Tatiana T Marquez-Lago, André Leier, Jerico Revote, Yan Zhu, David R Powell, Tatsuya Akutsu, Geoffrey I Webb, Kuo-Chen Chou, A Ian Smith, Roger J Daly, Jian Li, and Jiangning Song. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings in Bioinformatics*, 21(3):1047–1057, 04 2019.

[15] Richard F Voss. Evolution of long-range fractal correlations and 1/f noise in dna base sequences. *Physical review letters*, 68(25):3805, 1992.

[16] Paul Dan Cristea. Conversion of nucleotides sequences into genomic signals. *Journal of cellular and molecular medicine*, 6(2):279–303, 2002.

[17] Niranjan Chakravarthy, Andreas Spanias, Leonidas D Iasemidis, and Kostas Tsakalis. Autoregressive modeling and feature analysis of dna sequences. *EURASIP Journal on Applied Signal Processing*, 2004:13–28, 2004.

[18] Ren Zhang and Chun-Ting Zhang. Z curves, an intutive tool for visualizing and analyzing the dna sequences. *Journal of Biomolecular Structure and Dynamics*, 11(4):767–782, 1994.

[19] Achuthsankar S Nair and Sivarama Pillai Sreenadhan. A coding measure scheme employing electron-ion interaction pseudopotential (eiip). *Bioinformation*, 1(6):197, 2006.

[20] D. Anastassiou. Genomic signal processing. *IEEE Signal Processing Magazine*, 18(4):8–20, July 2001.

[21] Ning Yu, Zhihua Li, and Zeng Yu. Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning. *Big Data Mining and Analytics*, 1(3):191–210, 2018.

[22] Todd Holden, R Subramaniam, R Sullivan, E Cheung, C Schneider, G Tremberger Jr, A Flamholz, DH Lieberman, and TD Cheung. Atcg nucleotide fluctuation of deinococcus radiodurans radiation genes. In *Instruments, Methods, and Missions for Astrobiology X*, volume 6694, page 669417. International Society for Optics and Photonics, 2007.

[23] Gerardo Mendizabal-Ruiz, Israel Román-Godínez, Sulema Torres-Ramos, Ricardo A Salido-Ruiz, and J Alejandro Morales. On dna numerical representations for genomic similarity computation. *PloS one*, 12(3):e0173288, 2017.

[24] H Joel Jeffrey. Chaos game representation of gene structure. *Nucleic acids research*, 18(8):2163–2170, 1990.

[25] Tung Hoang, Changchuan Yin, and Stephen S-T Yau. Numerical encoding of dna sequences by chaos game representation with application in similarity comparison. *Genomics*, 108(3-4):134–142, 2016.

[26] Jonas S Almeida, Joao A Carrico, Antonio Maretzek, Peter A Noble, and Madilyn Fletcher. Analysis of genomic sequences by chaos game representation. *Bioinformatics*, 17(5):429–437, 2001.

[27] Changchuan Yin, Ying Chen, and Stephen S-T Yau. A measure of dna sequence similarity by fourier transform with applications on hierarchical clustering. *Journal of theoretical biology*, 359:18–28, 2014.

[28] Robson Parmezan Bonidia, Lucas Dias Hiera Sampaio, Fabrício Martins Lopes, André Carlos Ponce de Leon Ferreira de Carvalho, and Danilo Sipoli Sanches. Feature extraction approaches for biological sequences: A comparative study of mathematical models. *bioRxiv*, 2020.

[29] Sajia Akhter, Barbara A Bailey, Peter Salamon, Ramy K Aziz, and Robert A Edwards. Applying shannon's information theory to bacterial and phage genomes and metagenomes. *Scientific reports*, 3:1033, 2013.

[30] JA Tenreiro Machado, António C Costa, and Maria Dulce Quelhas. Shannon, rényie and tsallis entropy analysis of dna using phase plane. *Nonlinear Analysis: Real World Applications*, 12(6):3135–3144, 2011.

[31] Takuya Yamano. Information theory based on nonadditive information content. *Physical Review E*, 63(4):046105, 2001.

[32] Constantino Tsallis, RenioS Mendes, and Anel R Plastino. The role of constraints within generalized nonextensive statistics. *Physica A: Statistical Mechanics and its Applications*, 261(3-4):534–554, 1998.

[33] Georgios A. Pavlopoulos, Maria Secrier, Charalampos N. Moschopoulos, Theodoros G. Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G. Bagos. Using graph theory to analyze biological networks. *BioData Min*, 4(1), 2011.

[34] Tero Aittokallio and Benno Schwikowski. Graph-based methods for analysing networks in cell biology. *Brief Bioinformatics*, 7(3):243–255, sep 2006.

[35] Eric Augusto Ito, Isaque Katahira, Fábio Fernandes da Rocha Vicente, Luiz Filipe Protasio Pereira, and Fabrício Martins Lopes. Basinet—biological sequences network: a case study on coding and non-coding rnas identification. *Nucleic acids research*, 2018.