Biotext: Exploiting Biological-like format for text mining

Diogo J. S. Machado[1†], Camilla R. De Pierri[2†], Letícia G. C. Santos[3], Fabio O. Pedrosa[1,2] and Roberto T. Raittz[1*]

[1]Graduate Program in Bioinformatics, SEPT, Federal University of Paraná, Curitiba, Paraná, Brazil. [2]Department of Biochemistry and Molecular Biology, Federal University of Paraná, Curitiba, Paraná, Brazil. [3]Graduate Program in Biosciences and Biotechnology, FIOCRUZ - Carlos Chagas Institute, Curitiba, Paraná, Brazil.

*raittz@ufpr.br

†These authors contributed equally.

ABSTRACT

The large amount of existing textual data justifies the development of new text mining tools. Bioinformatics tools can be brought to Text Mining, increasing the arsenal of resources. Here, we present Biotext, a package of strategies for converting natural language text into biological-like information data, providing a general protocol with standardized functions, allowing to share, encode and decode textual data for amino acid data. The package was used to encode the arbitrary information present in the headings of the biological sequences found in a BLAST survey. The protocol implemented in this study consists of 12 steps, which can be easily executed and/ or changed by the user, depending on the study area. Biotext empowers user to perform text mining using bioinformatics tools. Biotext is Freely available at https://pypi.org/project/biotext/ (Python package) and https://sourceforge.net/projects/biotext-tools/files/AMINOcode_GUI/ (Standalone tool).

1 BACKGROUND

Texts are the most natural way to store information and are responsible for storing most of the world's scientific knowledge (TSHITOYAN *et al* 2019). As textual information continues to accumulate in public databases, contrasting strongly with the analysis options of these currently existing documents, it is necessary to develop tools to deal with the growing task of manipulating and analyzing data (KOCH *et al*., 2020). Texts can be encoded into biological information using computational strategies based on Bioinformatics. The benefit of representing texts in the form of biological sequences can be extended to the physical information storing level, since the use of DNA for this task allows the storage of high density of information in much smaller volumes (CHURCH *et al*., 2012; KOCH *et al*., 2020). Here, we developed a

strategy that adapts texts in natural language to be processed with various available bioinformatics tools. We present Biotext: A package to execute text mining strategy using bioinformatics tools. Biotext provides the AMINOcode and DNAbits capabilities to encode text in biological sequences format BSF (Biological the texts analysis in bioinformatics tasks. As example of application of Biotext, we exploited the tools coding arbitrary information present in the headings of the biological sequences found in a typical BLAST survey.

2 RESULTS

2.1 Converting texts into biological sequence format (BSF)

The presented package includes two functions – and respective reverse functions - to code conventional texts to a valid biological format. AMINOcode consists of replacing text characters with letters used to represent amino acids, according to a coded list (See Supplementary Material Table S1 and Supplementary Material Table S2). DNAbits is based on performing the conversion of characters to binary form in the "American Standard Code for Information Interchange" (ASCII), followed by replacing the binary values two by two with "A", "C", " G "or" T", according to a predefined rule for us (See Supplementary Material Table S3).

2.2 Case study - Interpreting BLAST results

We randomly selected the hypothetical protein "WP_011156533.1" from a set of sequences from a previous study (RUBEL *et al*., 2016) for research on BLAST (ALTSCHUL *et al*., 1990), run with the default parameters; the first 100 hits were considered (see Supplementary Material Table 3). The Headers of the sequences were converted to BSF, using AMINOcode, resulting in a FASTA file containing the text information. For the analyses of the header's contents, the headers sequences were converted to vectors by the function SWeeP (DE PIERRI *et al*., 2020) - included in the package – preparing them for clustering. Sequence headers were then grouped, using agglomerative clustering (Figure 1a). Finally, we aligned separately the sequences from each cluster with Clustal Omega (SIEVERS & HIGGINS, 2018) and obtained the sequences consensus that were decoded through reverse AMINOCode, as shown in Figure 1b (To check the other clusters, see Supplementary Material Figure 1). The entire process consists of 12 steps (for details, see Supplementary Material Figure S2).

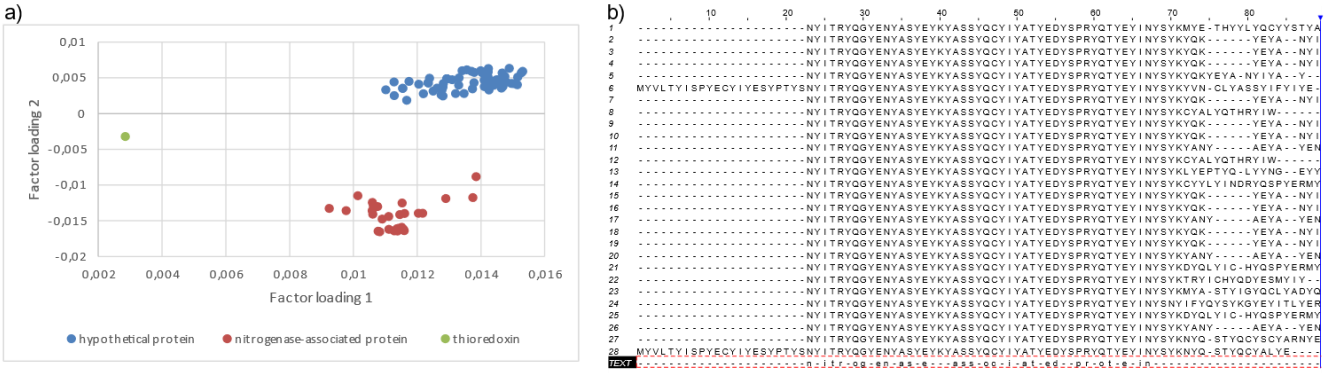


FIGURE 1. Biotext text mining process. **a.** Header text mining plot from protein sequence WP_011156533.1 (NCBI). Each point represents a header. The coordinates are factor loadings 1 and 2 resulting from PCA (principal component analysis), generated through the scikit-learn library, described by Pedregosa *et al*. (2011). The different colors represent different groups: Hypothetical protein (blue), Nitrogenase- associated protein (red), Thioredoxin (green). From the alignment of the encoded headers for each cluster, the decoded consensus was used in the legend. The alignment represented in figure "a" generated the legend "nitrogenase-associated protein" in figure “b”. **b.** Representations of alignments with coded texts of one cluster. The alignment positions are marked at the top. Cuts in the alignments were made to highlight the region of greatest importance. Headers are displayed on the right, with identification numbers for each sequence. The last, represented with the "TEXT" header, is the consensus text decoded, aligned with appropriate positions in coded texts.

3 DISCUSSION

The interpretation of a hypothetical protein is a challenge in Bioinformatics, and comparison strategies with several known sequences are often adopted to gain an understanding of them. Figure 1a shows the plot of the 2 principal components from a PCA of vectors representing the headers sequences. The clusters are labeled by the header consensus automatically generated for clusters: “Hypothetical protein”, “Protein associated with nitrogenase” and “Thioredoxin”. The distance between the points in the graph in Figure 1a is due to variations in the text in the headings. The more variations between the texts, the more distant their corresponding points appear on the graph. Even points of the same color (same cluster) are not in the same position, since the represented texts are not identical. The exploration of BSF proportionated by Biotext functions permitted automatically derive information from a hypothetical sequence from BLAST query using only common tools in bioinformatics to perform this task.

4 CONCLUSIONS

We propose Biotext and run an example in the interpretation a BLAST result. However, the library can be applied in a wide range of issues involving text processing and text mining common in bioinformatics, employing FASTA files and bioinformatics tools.

We encoded texts (FASTA headers) in BSF and thus it can be treating in a similar way to what would be done with conventional FASTA files. Through the case study, we showed that it is easy to

perform text mining tasks directly on BSF coded texts using bioinformatics tools and strategies, such as clustering, PCA, alignment (BLAST), consensus identification and even pattern recognition or machine learning.

5 IMPLEMENTATION AND AVAILABILITY

A Python package is available in the PyPI repository at <https://pypi.org/project/biotext/>. We also compiled a simple graphical interface to facilitate the use of AMINOcode (See Supplementary Material Figure 3), available at <https://sourceforge.net/projects/biotext- tools/files/AMINOcode_GUI/>. The method used in the case study in this paper can be performed using the script "fastaHeader2plot.py available at <https://github.com/diogomachado-bioinfo/biotext- examples/tree/master/fastaHeader2plot>. A complete list containing all the functions of Biotext is available in Supplementary Material Table 5.

ACKNOWLEDGEMENT

The authors thank the group of Artificial Intelligence Applied to Bioinformatics of Federal University of Paraná for the support.

FUNDING

This work has been supported by the Coordination of improvement of Higher Education Personnel – CAPES and Araucaria Foundation (n. 058/2019).

Conflict of Interest: none declared.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-generation digital information storage in DNA. *Science (New York, N.Y.)*, *337*(6102), 1628. https://doi.org/10.1126/science.1226355.

De Pierri, C. R., Voyceik, R., Santos de Mattos, L., 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., & Raittz, R. T. (2020). SWeeP: representing large biological sequences datasets in compact vectors. *Scientific reports*, *10*(1), 91. https://doi.org/10.1038/s41598-019-55627-4.

Koch, J., Gantenbein, S., Masania, K., Stark, W. J., Erlich, Y., & Grass, R. N. (2020). A DNA-of-things storage architecture to create materials with embedded memory. *Nature biotechnology*, *38*(1), 39–43. https://doi.org/10.1038/s41587-019-0356-z

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHAEL, V., THIRION, B (2011). Scikit-learn:

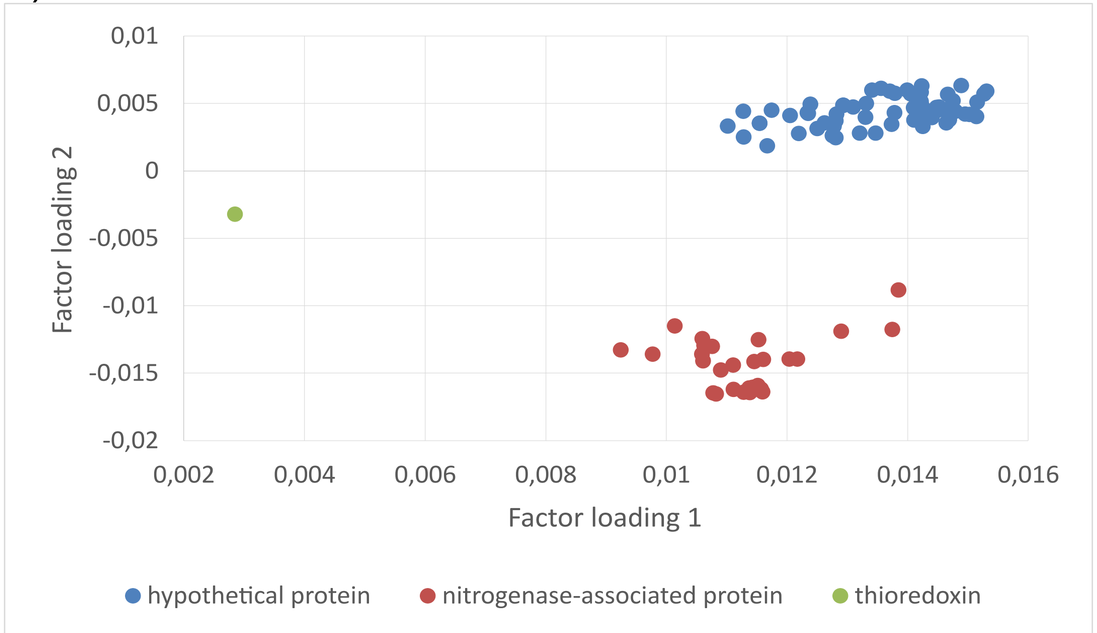
Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830.

Rubel, E. T., Raittz, R. T., Coimbra, N. A., Gehlen, M. A., & Pedrosa, F. O. (2016). ProClaT, a new bioinformatics tool for in silico protein reclassification: case study of DraB, a protein coded from the draTGB operon in Azospirillum brasilense. *BMC bioinformatics*, *17*(Suppl 18), 455. https://doi.org/10.1186/s12859-016-1338-5

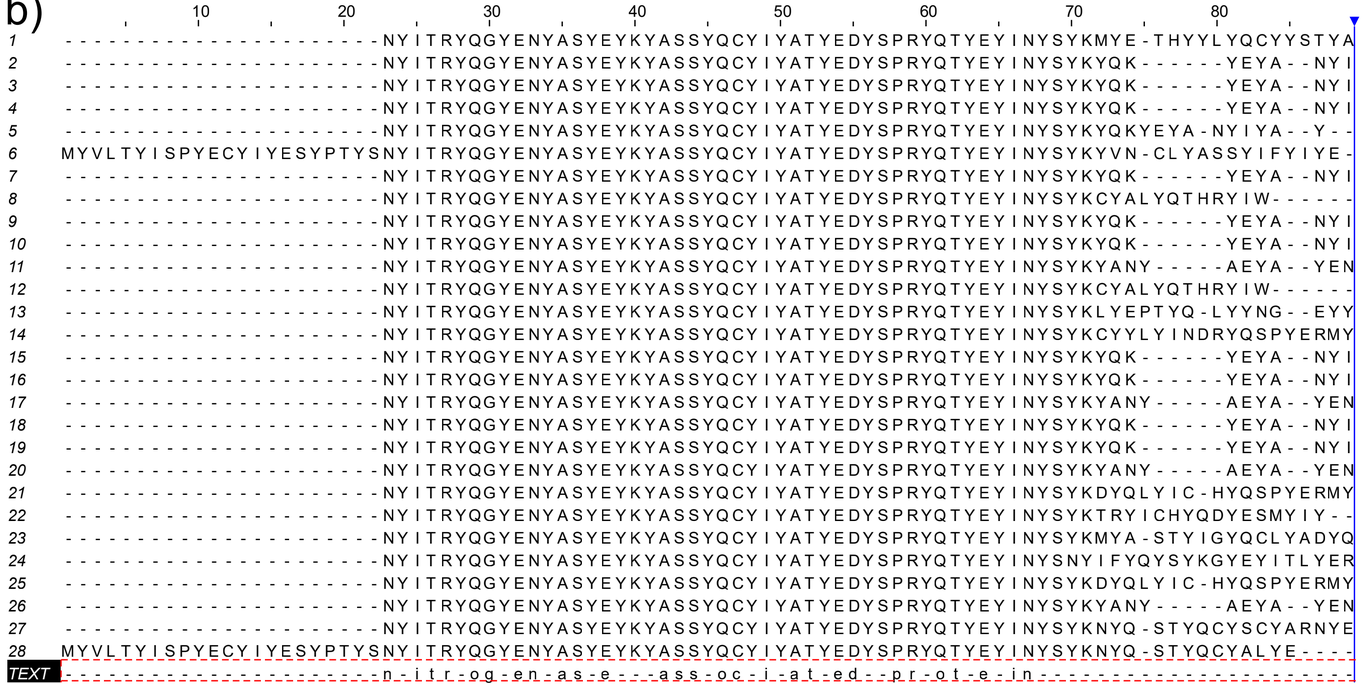
Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein science: a publication of the Protein Society*, *27*(1), 135–145. https://doi.org/10.1002/pro.3290

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571*(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8

a)


0,01
0,005
0
-0,005
-0,01
-0,015
-0,02
Factor loading 2
0,002
0,004
0,006
0,008
0,01
0,012
0,014
0,016
Factor loading 1
hypothetical protein
nitrogenase-associated protein
thioredoxin

b)


10
20
30
40
50
60
70
80
1 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKMYE-THYYLYQCYYSTYA
2 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
3 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
4 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
5 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQKYEYA-NYIYA--Y--
6 MYVLTYISPYECYIYESYPTYSNYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYVN-CLYASSYIFYIYE-
7 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
8 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKCYALYQTHRYIW------
9 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
10 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
11 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYANY-----AEYA--YEN
12 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKCYALYQTHRYIW------
13 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKLYEPTYQ-LYYNG--EYY
14 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKCYYLYINDRYQSPYERMY
15 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
16 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
17 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYANY-----AEYA--YEN
18 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
19 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYQK------YEYA--NYI
20 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYANY-----AEYA--YEN
21 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKDYQLYIC-HYQSPYERMY
22 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKTRYICHYQDYESMYIY--
23 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKMYA-STYIGYQCLYADYQ
24 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSNYIFYQYSYKGYEYITLYER
25 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKDYQLYIC-HYQSPYERMY
26 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKYANY-----AEYA--YEN
27 ----------NYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKNYQ-STYQCYSCYARNYE
28 MYVLTYISPYECYIYESYPTYSNYITRYQGYENYASYEYKYASSYQCYIYATYEDYSPRYQTYEYINYSYKNYQ-STYQCYALYE----
TEXT ----------n-itr-og-en-as-e---ass-oc-i-at-ed--pr-ot-e-in----------