BIOTEXT: EXPLOITING BIOLOGICAL-TEXT FORMAT FOR TEXT MINING

Diogo de Jesus Soares Machado [1,2†], Camilla Reginatto De Pierri[1,3 †], Letícia Graziela Costa Santos[4], Leonardo Scapin[1], Antonio Camilo da Silva Filho[5], Camila Pereira Perico[1,2], Fabio de Oliveira Pedrosa[3] and Roberto Tadeu Raittz[1,2 *]

[1] Laboratory of Artificial Intelligence applied to Bioinformatics (AIBIA), Curitiba, Brazil.
[2] Graduate Program in Bioinformatics, Federal University of Paraná, Curitiba, Brazil.
[3] Graduate Program in Sciences - Biochemistry, Federal University of Paraná, Curitiba, Brazil.
[4] Carlos Chagas Institute (FIOCRUZ), Curitiba, Brazil.
[5] Graduate Program in Pharmaceutical Sciences, Federal University of Paraná, Curitiba, Brazil.

[†]These authors contributed equally
[*]raittz@ufpr.br

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 and DNA. 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 users to perform text mining using bioinformatics tools. BIOTEXT is freely available at https://pypi.org/project/BIOTEXT/ (Python package) and https://sourceforge.net/projects/BIOTEXTtools/files/AMINOcode_GUI/ (Standalone tool).

1. INTRODUCTION

Texts are the most natural way of storing information and are responsible for the most extensive inventory of world scientific knowledge (TSHITOYAN et al., 2019). In sharp contrast to the analysis options, the mass of textual information in public databases demonstrates the need to develop tools to meet the growing demand for data manipulation and analysis (KOCH et al., 2020). For the omics area, whose primary focus is the analysis of complex biological events, Text Mining (TM) combined with bioinformatics techniques represents an advance for exploring and analyzing unstructured data (ZAFEIROPOULOS et al., 2022).

Today's TM tools address different methodologies, with potential for research in the most diverse scenarios (TABLE S1 – for details of review, see FILE S1). These approaches involve the main tasks such as Information Retrieval (IR), Named Entity Recognition (NER), and Information extraction (IE) (FIGURE S1), and a series of algorithms can be run depending on the type of analysis (TABLE S2) (FLEUREN; ALKEMA, 2015). However, we note that even with several aspects of TM well resolved in the literature, it is still possible to implement these processes advantageously. One of these ways is directly transposing Bioinformatics technologies to the TM area.

The exploration of biological sequences has a strong analogy with TM, since it works fundamentally with character string analysis (HARMSTON et al., 2010). From more traditional techniques such BLAST for sequence alignment (ALTSCHUL et al., 1990), to more recent ones such as the representation of biological sequences in vector spaces (ASGARI, MOFRAD, 2015; LEIMEISTER et al., 2018; DE PIERRI et al., 2020; RAITTZ et al., 2021) can be brought to the TM area, adding to the arsenal of existing text analysis tools. In addition, as in the field of Bioinformatics, vector representation of texts is also a trend in text processing (MIKOLOV et al., 2013; LEVY, GOLDENBERG, 2014; PENNINGTON et al., 2014), since TM vectoral technologies are equally promising (MA et al., 2015; LILLEBERG et al., 2015; DEVLIN et al., 2018; HASSANI et al., 2020).

However, the Bioinformatics tools available today do not process normal texts. It is necessary to use the Biological Sequence Format (BSF) to perform analysis. Thus, here we present an approach that adapts texts in natural language to be processed with several Bioinformatics tools, including the main resources for the comparison of sequences added to the agility and robustness of vectorization technologies. BIOTEXT is a package to perform text mining

strategies using bioinformatics tools, which provides AMINOcode and DNAbits features to encode text in BSF.

We explored tools for encoding arbitrary information present in the headers of biological sequences found in a typical BLAST search. We demonstrate that 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 level of physical information storage, as the use of DNA for this task allows the storage of information in high density in much smaller volumes (TSHITOYAN et al., 2019; KOCH et al., 2020).

2. METHODS

2.1 BIOTEXT Implementation

The Python package is available from the PyPI repository at https://pypi.org/project/BIOTEXT/. The case study is accessible through the script "fastaHeader2plot.py, available at https://github.com/diogomachado-bioinfo/BIOTEXT-examaples/tree/master/fastaHeader2plot. In addition, a complete list containing all BIOTEXT *functions is available in Table S3.*

2.2 BIOTEXT application

To demonstrate one of the applications of BIOTEXT using traditional methods in bioinformatics, we explored the local alignment for direct comparison between texts. We randomly selected the protein hypothetical "WP_011156533.1" of a set of sequences from a previous study (RUBEL et al., 2015) for BLAST research, using default parameters; the first 100 hits were considered (for details, see TABLE S4).

3. RESULTS AND DISCUSSION

3.1 Converting texts into biological sequence format (BSF)

The BIOTEXT 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. Table 1 shows the character substitution rules in AMINOcode. In addition, we present a minimal and an expanded encoding table that includes number and punctuation details representations. Whereas the 20 letters used to represent the amino acids are less than 26 needed to represent the whole alphabet, we define a conversion rule. For consonants, except for W and Y, the replacement is carried by a single letter of the amino acid representation (TABLE1). The coding replaces the vowels with two characters. The first is Y, a 'key' that defines the subsequent letter. Next, in the same perspective, all kinds of spaces in the texts are coded with YS. Finally, we propose the expanded version for digits and punctuation. In the expanded one, we use a key with two key characters, YP followed by the coding type. The minimal version can cause slight information, but it may be helpful in some instances. Also, we define a wildcard representation, YK, for any exception to Table 1.

TABLE 1. AMINOcode and DNAbits Dictionary

AMINOcode

Letters	Minimal/ Expanded
a	YA
b	E
c	C
d	D
e	YE
f	F
g	G
h	H
i	YI
j	I
k	K
l	L
m	M
n	N
o	YQ
p	P
q	Q
r	R
s	S
t	T
u	YV
v	V
x	W
z	A
w	YW
y	YY

Numbers	Expanded	Minimal
0	YDA	YD
1	YDQ	
2	YDT	
3	YDH	
4	YDF	
5	YDI	
6	YDS	
7	YDE	
8	YDG	
9	YDN	

Punctuation	Expanded	Minimal
.	YPE	YP
,	YPC	
;	YPS	
!	YPW	
?	YPQ	
:	YPT	

Other	Minimal/Expanded
Space	YS
Any different	YK

DNAbits

Bits	Nucleotide Representation
00	A
10	C
01	G
11	T

Legend: DNAbits dictionary is represented by the nucleotides A (adenine), C (cytosine) G (guanine) and T (thymine);

DNAbits performs the conversion of text characters splitting every byte into four pairs of bits. The codding maintains the information according to the "American Standard Code for

Information Interchange" (ASCII), replacing every couple of bits by A, C, G, and T according to the predefined rule (Table 1). In DNAbits, the characters are encoded from binary form by the function in Eq.(1), and the decoding of the DNAbits is translatable to the original text.

$$f(b) = \begin{cases} b = 00, \ \mathrm{A} \\ b = 10\,, \mathrm{C} \\ b = 01, \mathrm{G} \\ b = 11, \mathrm{T} \end{cases} \tag{1}$$

where b is any pair of bits, for example, the binary form for the character representing the letter 'a' is 10000110, resulting in CAGC after DNAbits encoding.

3.2 Interpreting BLAST results

We randomly selected the hypothetical protein WP_011156533.1" from a set of sequences from a previous study (RUBEL *et al.*, 2015) for research on BLAST (ALTSCHUL *et al.*, 1990) and run with the default parameters; the first 100 hits were considered (TABLE S4). Next, the headers of the sequences were converted to BSF using AMINOcode, resulting in a FASTA file containing the text information. For the header's contents analyses, the headers sequences were converted to vectors by the SWeeP function (DE PIERRI *et al.*, 2020) Then, the headers groups were defined by agglomerative clustering (FIGURE 1a). Finally, we aligned the sequences from each cluster separately with Clustal Omega (SIVIERS *et al.*, 2018) and obtained the sequence consensus decoded through reverse AMINOCode, as shown in Figure 1b. The entire process consists of 12 steps (see Figure S2):

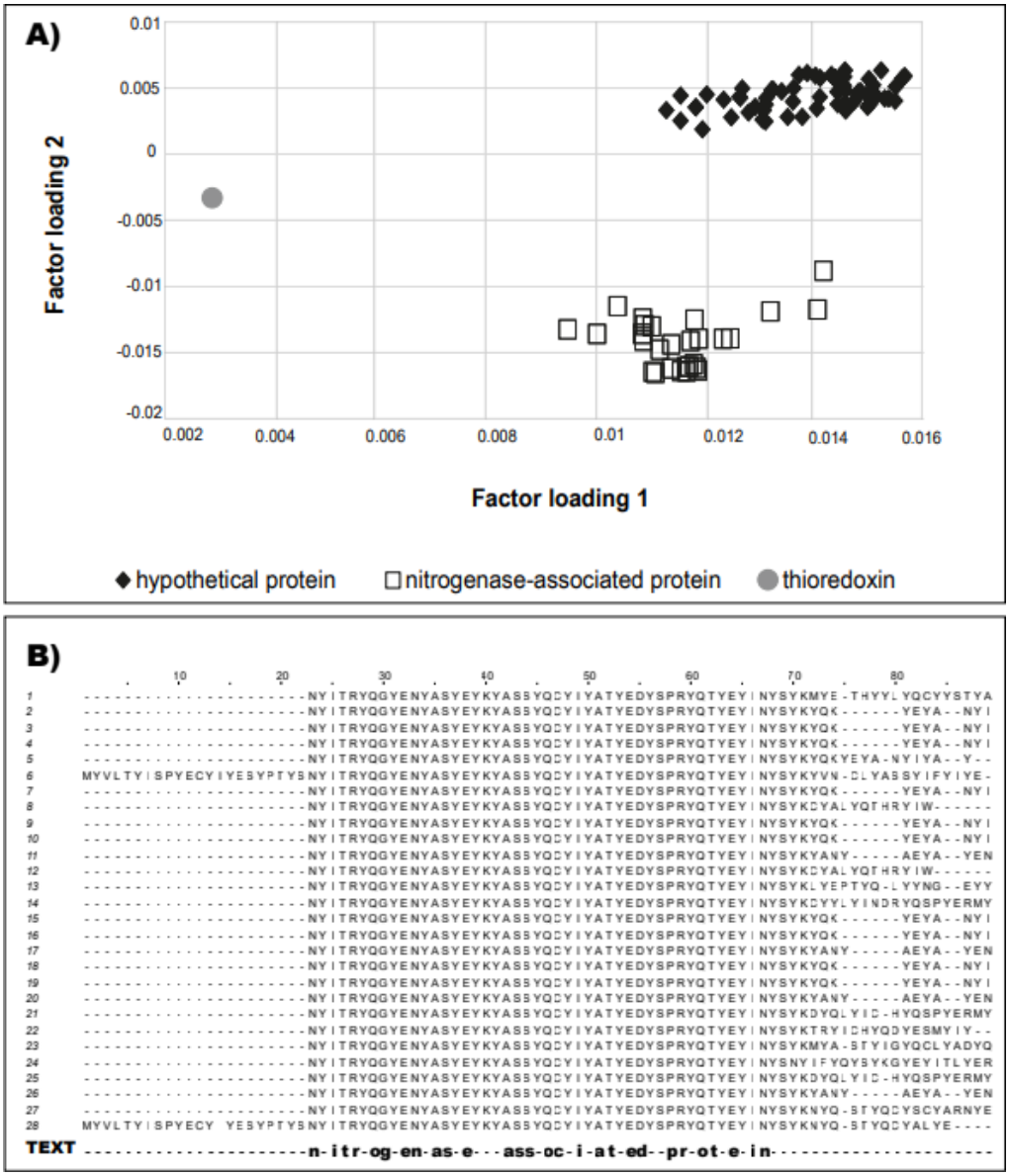
1. Extract headers from files;
2. Use expanded version with number and punctuation details (optional);
3. Find and remove patterns using regular expression for ID from all encoded texts;
4. Vectorize encoded header in a matrix with a line for each sequence;
5. Obtain the first two dimensions of loading of the PCA from SWeeP matrix;
6. Hierarquical clustering for two dimensions of loading, using the agglomerative clustering (scikit- learn Python library);
7. Agglomerative clustering, using the 0,015 threshould (optional);
8. Group the encoded texts according to the clusters;
9. Obtain the consensus of sequences in each cluster from Clustal Omega alignment;

10. Find and remove patterns using regular expression for species name from consensus strings;
11. Decode the profile strings in texts;
12. Plot a 2D graph with two loading dimensions with the equivalent texts derived from the profiles as a legend.

According to the exploited data, the use of expanded or minimal versions is optional. Steps 3 and 10 correspond to removing unique patterns from each header (ID and species), which do not provide helpful information for the clustering and analysis of the texts.

Figure 1a shows the graph of the two main components of a vector PCA that represent the header sequences. Clusters are labeled by the consensus header automatically generated for the clusters: "Hypothetical protein", "Nitrogenase-associated protein," and "Thioredoxin". Distances among points (graph in FIGURE 1a) result from variations in the headers' text. The more variations between texts, the further apart the dots appear on the chart. Even points of the same cluster are not in the same position, as the texts represented are not identical. The exploitation of the BSF provided by BIOTEXT functions allowed us to derive information from a hypothetical BLAST query automatically. Interpreting a hypothetical protein is a challenge in bioinformatics, and comparisons with various known sequences are often adopted to understand them (IJAQ *et al*., 2019). We seek to explore the relationship of information present in the header of a hypothetical protein using standard bioinformatics tools to accomplish this task.

FIGURE 1. Graphic representation of sequence headers



Reference: Authors, 2022.

Legend: a) Each dot represents a header obtained by the BLAST search using the WP_011156533.1 protein as a query. Consensus Headers are defined by: Hypothetical protein (lozangue), Nitrogenase-associated protein (square) and Thioredoxin (circle). The PCA was generated using the scikit-learn library (PEDREGOSA *et al.*, 2011). b) Representations of alignments with coded texts from a cluster. Alignment positions are shown at the top of the figure. Cuts were made in the alignments to highlight the region of greatest importance. Headers are displayed with identification numbers for each sequence (left). The last line, represented by the "TEXT" header is the decoded consensus text, properly aligned according to the coded texts.

3.3 BIOTEXT application experiment in large corpus

With the string "thioredoxin", we searched on PubMed using the Biopython package "Bio.Entrez" (COCK *et al*., 2009) (on March 18, 2022), resulting in 13,734 entries. We extract only the titles retrieved from the search and apply AMINOcode encoding using the expanded version (with detailed digits and punctuation). Then, the coded texts were vectorized and derived in a dendrogram with Euclidean distance based on the distance matrix - implementation of the SciPy library (VIRTANEN *et al*., 2020). We use the scikit-bio library to transform the dendrogram into Newick format. The resulting Newick file, input text, and script to re-run all processes are available at https://github.com/diogomachado-bioinfo/BIOTEXT-examples/tree/master/titleSet2tree.

It is essential to highlight that the execution was quickly on an ordinary computer (i5-3470 processor and 16 Gb of DDR3 RAM). The entire process took 317.64 seconds. In particular, encoding with AMINOcode took 1.12 seconds, vectoring with SWeeP 244.38 seconds and creating the dendrogram 72.14 seconds.

4. CONCLUSION

Here we present BIOTEXT and run an example in interpreting a result of BLAST. The library can be applied in various issues involving word processing and text mining, employing FASTA files and Bioinformatics tools. In addition, this study can contribute to research on the vector representation of texts.

We encode texts (FASTA headers) in BSF and, therefore, could be treated similarly to what would be with biological FASTA. The case study shows that it is easy to perform TM tasks directly in BSF coded texts using bioinformatics tools and strategies, such as Clustering, PCA, alignment, and BLAST for consensus identification and even pattern recognition or machine learning. However, the fact that the available tools are dependent on vocabulary makes text handling complex. BIOTEXT incorporates compact vector technology to streamline analysis and eliminate the need for predefined vocabulary.

BIOTEXT combined with SWeeP vectors technology provided handling of large amounts of text, with fast execution time even on a personal computer. Moreover, our proposal is versatile, allowing the use of the entire arsenal of algorithms/methods in bioinformatics for analyzes in which texts are transformed into BSF. BIOTEXT is a general purpose tool for text mining that allows one to use different bioinformatics algorithms for multipurpose analyses and perform the main text mining tasks (IR, NER, and IE). In short, our approach can be handy for bioinformatics, given the analysts' experience with these features.

ACKNOWLEDGMENT

The authors thank the group of Artificial Intelligence Applied to Bioinformatics of the Federal University of Paraná and the Coordination of Improvement of Higher Education Personnel – CAPES for the support.

REFERENCES

AKELLA, L. M., NORTON, C. N., MILLER, H. NetiNeti: discovery of scientific names from text using machine learning methods. BMC bioinformatics, 13, 211, 2012.

ALLOT, A., PENG, Y., WEI, C. H., LEE, K., PHAN, L., LU, Z. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic acids research, v.46(W1), p. W530–W536, 2018.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410, 1990.

ASGARI, E. & MOFRAD, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One 10, 1–15, 2015.

BARBOSA-SILVA, A., FONTAINE, J. F., DONNARD, E. R., STUSSI, F., ORTEGA, J. M. ANDRADE- NAVARRO, M. A. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted fromPubMed queries. BMC bioinformatics, v.12, n.435, 2011.

BARBOSA-SILVA, A., SOLDATOS, T.G., MAGALHÃES, I.L. et al. LAITOR - Literature Assistant forIdentification of Terms co-Occurrences and Relationships. BMC Bioinformatics v.11, n. 70, 2010.

BECKER, K. G., HOSACK, D. A., DENNIS, G., JR, LEMPICKI, R. A., BRIGHT, T. J. CHEADLE, C., ENGEL, J. PubMatrix: a tool for multiplex literature mining. BMC bioinformatics, v.4, n.61, 2003.

BRANCOTTE, B., BITON, A., BERNARD-PIERROT, I., RADVANYI, F., REYAL, F., COHEN-BOULAKIA, SGene List significance at-a-glance with GeneValorization. Bioinformatics (Oxford, England), v.27, n.8,p.1187–1189, 2011.

CAÑADA, A., CAPELLA-GUTIERREZ, S., RABAL, O., OYARZABAL, J., VALENCIA, A., KRALLINGER, M. LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. Nucleic acids research, v.45(W1), p. W484–W489, 2017.

CEJUELA, J. M., BOJCHEVSKI, A., UHLIG, C., BEKMUKHAMETOV, R., KUMAR KARN, S. MAHMUTI, S. BAGHUDANA, A., DUBEY, A., SATAGOPAM, V. P., Rost, B. nala: text mining natural language mutationmentions. Bioinformatics (Oxford, England), v.33, n.12, p.1852–1858, 2017.

CHAIX, E., DELÉGER, L., BOSSY, R. NÉDELLEC, C. Text mining tools for extracting information about microbial biodiversity in food. Food microbiology, v.81, p.63–75, 2019.

CHEN, C. C., HO, C. L. PubstractHelper: A Web-based Text-Mining Tool for Marking Sentences in Abstracts from PubMed Using Multiple User-Defined Keywords. Bioinformation, v.10, n.11, p.708–710, 2014.

CHEN, H., SHARP, B. M. Content-rich biological network constructed by mining PubMed abstracts. BMC bioinformatics, v.5, n.147, 2004.

CHENG, D., KNOX, C., YOUNG, N., STOTHARD, P., DAMARAJU, S., WISHART, D. S. PolySearch: a web- based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic acids research, n.36, W399–W405, 2008.

COMEAU, D. C., BATISTA-NAVARRO, R. T., DAI, H. J., DOĞAN, R. I., YEPES, A. J., KHARE, R., LU, Z., MARQUES, H., MATTINGLY, C. J., NEVES, M., PENG, Y. et al. BioC interoperability track overview. Database: the journal of biological databases and curation, bau053, 2014.

CRISAN, A., MUNZNER, T., GARDY, J. L. Adjutant: an R-based tool to support topic discovery for systematicand literature reviews. Bioinformatics (Oxford, England), v.35, n.6, p.1070–1072. 2019.

DAVID J. STATES, ALEX S. ADE, ZACHARY C. WRIGHT, AARON V. BOOKVICH, BRIAN D. Athey, MiSearch adaptive pubMed search tool, Bioinformatics, v. 25, Issue 7, p. 974–976, 2009.

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, ORTAGEA, J.M; PEDROSA, F.O. RAITTZ, R.T. SWeeP: Representing large biological sequences datasets in compact vectors. Sci. Rep, 10, 91, 2020.

DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

DING, R., BOUTET, E., LIEBERHERR, D., SCHNEIDER, M., TOGNOLLI, M., WU, C. H., VIJAY-SHANKER, K., ARIGHI, C. N. eGenPub, a text mining system for extending computationally mapped bibliography for UniProt Knowledgebase by capturing centrality. Database: the journal of biological databases and curation, bax081, 2017.

DOUGHTY, E.; KERTESZ-FARKAS, A.; BODENREIDER, O.; THOMPSON, G.; ADADEY, A.; PETERSON, T.; KANN, M.G. Toward an automatic method for extracting cancer- and other disease-related point mutationsfrom the biomedical literature, Bioinformatics, v.27, n. 3, 408–415, 2011.

ELSWORTH, B., GAUNT, T. R. MELODI Presto: a fast and agile tool to explore semantic triples derived from biomedical literature. Bioinformatics (Oxford, England), v.37, n.4, p. 583–585, 2021

FONTAINE, J. F., BARBOSA-SILVA, A., SCHAEFER, M., HUSKA, M. R., MURO, E. M., ANDRADE- NAVARRO, M. A. MedlineRanker: flexible ranking of biomedical literature. Nucleic acids research, v.37 (Web Server issue), p. W141–W146, 2009.

FURRER, L., JANCSO, A., COLIC, N., RINALDI, F. OGER++: hybrid multi-type entity recognition. Journal of cheminformatics, n.11(1), v.7, 2019.

GERNER, M., NENADIC, G., BERGMAN, C. M. LINNAEUS: a species name identification system for biomedical literature. BMC bioinformatics, v.11, n.85., 2010

GERNER, M., SARAFRAZ, F., BERGMAN, C. M., NENADIC, G. BioContext: an integrated text mining system for large- scale extraction and contextualization of biomolecular events. Bioinformatics (Oxford, England), v.28, n.16, p.2154–2161, 2012.

GIGLIA E. Quertle and KNALIJ: searching PubMed has never been so easy and effective. European journal of physical and rehabilitation medicine, v.47, n.4, p.687–690, 2011.

GIJÓN-CORREAS, J. A., ANDRADE-NAVARRO, M. A., FONTAINE, J. FAlkemio: association of chemicals with biomedical topics by text and data mining. Nucleic acids research, v.42(Web Server issue), p.W422– W429, 2014.

GUPTA, S., ROSS, K. E., TUDOR, C. O., WU, C. H., SCHMIDT, C. J., VIJAY-SHANKER, K. miRiaD: A Text Mining Tool for Detecting Associations of microRNAs with Diseases. Journal of biomedical semantics, v.7(1),n.9, 2016.

HARMSTON, N.; FILSELL, W.; STUMPF, M.P.H. What the papers say: Text mining for genomics and system biology. Human Genomics, v.5, n.1, p.17-29, 2010.

HASSANI, H.; BENEKI, C.; UNGER, S.; MAZINANI, M. T.; YEGANEGI, M.R. Text mining in big data analytics. Big Data and Cognitive Computing, v. 4, n. 1, p. 1, 2020.
HE, X., LI, Y., KHETANI, R., SANDERS, B., LU, Y., LING, X., ZHAI, C., SCHATZ, B. BSQA: integrated text mining using entity relation semantics extracted from biological literature of insects. Nucleic acids research, n.38 (Web Server issue), p. W175–W181, 2010.

HIRSCHMAN, L., YEH, A., BLASCHKE, C., VALENCIA, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC bioinformatics, v.6. Suppl 1, 2005.

IJAQ, J.; MALIK, G.; KUMAR, A.; DAS, P. S.; MEENA, N.; BETHI, N.; SURAVAJHALA, P. A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. BMC bioinformatics, v. 20, n. 1, p. 1-8, 2019.

JESSOP, D. M., ADAMS, S. E., WILLIGHAGEN, E. L., HAWIZY, L., MURRAY-RUST, P. OSCAR4: a flexible architecture for chemical text-mining. Journal of cheminformatics, v.3(1), n.41, 2011.

KOCH J., GANTENBEIN S., MASANIA K., STARK W.J., ERLICH Y., GRASS R.N. A DNA-of-things storage architecture to create materials with embedded memory. Nat Biotechnol 38, 39–43, 2020.

KONING, D., SARKAR, I. AND MORITZ, T. TaxonGrab: Extracting taxonomic names from text', BiodiversityInform. v. 2, pp. 79–82, 2005.

KRALLINGER, M., RABAL, O., LOURENÇO, A., OYARZABAL, J., VALENCIA, A. A. Information retrieval and text mining technologies for chemistry. Chem. Rev. v.117, p. 7673–7761, 2017.

KRALLINGER, M., RODRIGUEZ-PENAGOS, C., TENDULKAR, A., VALENCIA, A. PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction. Nucleic acids research, v.37(WebServer issue), p.W160–W165, 2009.

KUHN, M., VON MERING, C., CAMPILLOS, M., JENSEN, L. J., BORK, P. STITCH: interaction networks of chemicals and proteins. Nucleic acids research, v.36(Database issue), D684–D688, 2008.

KWON, D., KIM, S., WEI, C. H., LEAMAN, R., LU, Z. ezTag: tagging biomedical concepts via interactivelearning. Nucleic acids research, v.46(W1), p.W523–W529, 2018.

LE, D. H., TRAN, T. RWRMTN: a tool for predicting disease-associated microRNAs based on a microRNA-target gene network. BMC bioinformatics, v.21(1), n.244, 2020.

LEARY, P. R., REMSEN D. P, NORTON, C.N., PATTERSON, D.J. SARKAR, I.N. uBioRSS: Tracking taxonomic literature using RSS, Bioinformatics, v.23, n.11, June, 1p.434–1436 2007.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. Nature v.521, p,436–444, 2015.

LEE, H., YI, G. S., PARK, J. C. E3Miner: a text mining tool for ubiquitin-protein ligases. Nucleic acids research, v.36, W416–W422, 2008.

LEIMEISTER CA, SCHELLHORN J, DÖRRER S, GERTH M, BLEIDORN C, MORGENSTERN B. Prot-SpaM: fasta alignment-free phylogeny reconstruction based on whole-proteome sequences. Gigascience 8, 1–14, 2018.ASGARI

LEVY, O.; GOLDBERG, Y. Linguistic regularities in sparse and explicit word representations. In: Proceedings of the eighteenth conference on computational natural language learning. 2014. p. 171-180.

LI, G., ROSS, K. E., ARIGHI, C. N., PENG, Y., WU, C. H., VIJAY-SHANKER, K. miRTex: A Text Mining System for miRNA-Gene Relation Extraction. PLoS computational biology, v.11, n.9, e1004391, 2015.

LI, M., HE, Q., MA, J., HE, F., ZHU, Y., CHANG, C., CHEN, T. PPICurator: A Tool for Extracting Comprehensive Protein-Protein Interaction Information. Proteomics, v.19(4), e1800291, 2019.

LIM, K. M. K.; LI, C.; CHING, K. R.; NAGARAJAN, N. @MInter: Automated text-mining of microbial interactions. Bioinformatics 32, 2981–2987, 2016.

LILLEBERG, J.; ZHU, Y.; ZHANG, Y. Support vector machines and word2vec for text classification with semantic features. In: 2015 IEEE 14th International

LIU, W., LAULEDERKIND, S. J., HAYMAN, G. T., WANG, S. J., NIGAM, R., SMITH, J. R., DE PONS, J., DWINELL, M. R., SHIMOYAMA, M. OntoMate: a text-mining tool aiding curation at the Rat GenomeDatabase. Database: the journal of biological databases and curation, bau129, 2015.

LOURENÇO, A., CARREIRA, R., CARNEIRO, S., MAIA, P., GLEZ-PEÑA, D., FDEZ-RIVEROLA, F., FERREIRA, E. C., ROCHA, I., ROCHA, M. @Note: a workbench for biomedical text mining. Journal of biomedical informatics, v.42, n.4, p.710–720, 2009.

MA, Long; ZHANG, Yanqing. Using Word2Vec to process big text data. In: 2015 IEEE International Conference on Big Data (Big Data). IEEE, 2015. p. 2895-2897.

MAHMOOD, A., RAO, S., MCGARVEY, P., WU, C., MADHAVAN, S., VIJAY-SHANKER, K.. eGARD: Extracting associations between genomic anomalies and drug responses from text. PloS one, v.12, n.12,e0189663, 2017.

MCDONALD, R., PEREIRA, F. Identifying gene and protein mentions in text using conditional randomfields. BMC Bioinformatics v.6, S6, 2005.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

MÜLLER, H. M., KENNY, E. E., STERNBERG, P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS biology, v. 2, n.11, e309, 2004.

NADERI, N., KAPPLER, T., BAKER, C. J., WITTE, R. OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. Bioinformatics (Oxford, England), v.27, .19), p.2721–2729, 2011.

NUNES, T., CAMPOS, D., MATOS, S., OLIVEIRA, J. L. BeCAS: biomedical concept recognition services and visualization. Bioinformatics (Oxford, England), v.29, n.15, p.1915–1916, 2013.

PAFILIS, E., FRANKILD, S. P., FANINI, L., FAULWETTER, S., PAVLOUDI, C., VASILEIADOU, A., ARVANITIDIS, C., JENSEN, L. J. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. PloS one, v.8(6), e65390, 2013.

PAPANIKOLAOU, N., PAVLOPOULOS, G. A., THEODOSIOU, T., ILIOPOULOS, I. Protein-protein interaction predictions using text mining methods. Methods v.74, p.47–53, 2015.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHAEL, V., THIRION, B. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830, 2011.

PIERECK, B., OLIVEIRA-LIMA, M., BENKO-ISEPPON, A. M., DIEHL, S., SCHNEIDER, R., BRASILEIRO-VIDAL, A. C., BARBOSA-SILVA, A. LAITOR4HPC: A text mining pipeline based on HPC for building interaction networks. BMC bioinformatics, v.21(1), n.365, 2020.

PLAKE, C., ROYER, L., WINNENBURG, R., HAKENBERG, J., SCHROEDER, M. GoGene: gene annotation in the fast lane. Nucleic acids research, v.37 (Web Server issue), p.W300–W304, 2009.

PLAKE, C., SCHIEMANN, T., PANKALLA, M., HAKENBERG, J., LESER, U. AliBaba: PubMed as a graph. Bioinformatics (Oxford, England), v.22, n.19, p. 2444–2445, 2006.

PLETSCHER-FRANKILD, S., PALLEJÀ, A., TSAFOU, K., BINDER, J. X., JENSEN, L. J. DISEASES: Text mining and data integration of disease-gene associations. Methods v.74, p.83–89, 2015.

PRICE, M. N.; ARKIN, A. P. PaperBLAST: Text Mining Papers for Information about Homologs. Systems 2,1–10, 2017.

RAVIKUMAR, K. E., RASTEGAR-MOJARAD, M., LIU, H. BELMiner: adapting a rule-based relation extractionsystem to extract biological expression language statements from bio-medical literature evidence sentences. Database: the journal of biological databases and curation, baw156, 2017.

REBHOLZ-SCHUHMANN, D., ARREGUI, M., GAUDAN, S., KIRSCH, H., JIMENO, A. Text processing through Web services: calling Whatizit. Bioinformatics (Oxford, England), v.24, n.2, p.296–298, 2008.

REN, J., LI, G., ROSS, K., ARIGHI, C., MCGARVEY, P., RAO, S., COWART, J., MADHAVAN, S., VIJAY- SHANKER, K., WU, C. H. iTextMine: integrated text-mining system for large-scale knowledge extraction fromthe literature. Database: the journal of biological databases and curation, bay128, 2018.

RINALDI, F., CLEMATIDE, S., MARQUES, H., ELLENDORFF, T., ROMACKER, M., RODRIGUEZ-ESTEBAN, R. OntoGene web services for biomedical text mining. BMC bioiformatics, v.15 Suppl 14(Suppl 14), S6, 2014.

RUBEL E.T., RAITTZ RT, COIMBRA N.A., GEHLEN M.A., PEDROSA F.O. ProClaT, a new bioinformatics toolfor in silico protein reclassification: case study of DraB, a protein coded from the draTGB operon in Azospirillum brasilense. BMC Bioinformatics 17, 455, 2016.

SALGADO, D., KRALLINGER, M., DEPAULE, M., DRULA, E., TENDULKAR, A. V., LEITNER, F., VALENCIA, A., MARCELLE, C. MyMiner: a web application for computer-assisted biocuration and text annotation. Bioinformatics (Oxford, England), v.28, n.17, 2285–2287, 2012.

SEN SARMA, M., ARCOLEO, D., KHETANI, R. S., CHEE, B., LING, X., HE, X., JIANG, J., MEI, Q., ZHAI, C., SCHATZ, A. BeeSpace Navigator: exploratory analysis of gene function using semantic indexing of biologicalliterature. Nucleic acids research, v.39(Web Server issue), p.W462–W469. 2011.

SETTLES B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names intext. Bioinformatics (Oxford, England), v.21, n.14, p.3191–3192, 2005.

SHAO, W., ADAMS, C. E., COHEN, A. M., DAVIS, J. M., MCDONAGH, M. S., THAKURTA, S., YU, P. S. SMALHEISER, N. R. Aggregator: A machine learning approach to identifying MEDLINE articles that derive from the same underlying clinical trial. Methods v.74, p.65–70, 2015.

SHATKAY, H.; BRADY, S.; WONG, A. Text as data: Using text-based features for protein representation and for computational prediction of their characteristics. Methods 74, 54–64, 2015.

SIEVERS F., HIGGINS D.G. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci; 27:135–145, 2018.

SIMON, C., DAVIDSEN, K., HANSEN, C., SEYMOUR, E., BARNKOB, M. B., OLSEN, L. R. BioReader: a text mining
tool for performing classification of biomedical literature. BMC bioinformatics, v.19(Suppl 13), n.57,2019.

SINGHAL, A.; SIMMONS, M.; LU, Z. Text Mining Genotype-Phenotype Relationships from BiomedicalLiterature for Database Curation and Precision Medicine. PLoS Comput. Biol. 12, 1–19, 2016.

SMALHEISER, N. R., ZHOU, W., TORVIK, V. I. Anne O'Tate: A tool to support user-driven summarization, drill down and browsing of PubMed search results. Journal of biomedical discovery and collaboration, v.3,n.2, 2008.

SNEL, B., LEHMANN, G., BORK, P., HUYNEN, M. A. STRING: a web-server to retrieve and display therepeatedly occurring neighbourhood of a gene. Nucleic acids research, v.28, n.18, p. 3442–3444, 2000.

SRINIVASAN P. MeSHmap: a text mining tool for MEDLINE. Proceedings. AMIA Symposium, p.642–646, 2001.

STEFFEN, P., WU, J., HARIHARAN, S., VOSS, H., RAGHUNATH, V., MOLLOY, M. P., SCHLÜTER, H. OmixLitMiner: A Bioinformatics Tool for Prioritizing Biological Leads from 'Omics Data Using Literature Retrieval and Data Mining. International journal of molecular sciences, v.21(4), 1374, 2020.

TANABE, L., SCHERF, U.; SMITH, L. H.; LEE, J. K.; HUNTER, L.; WEINSTEIN, J. N. MedMiner: An Internet text- mining tool for biomedical information, with application to gene expression profiling. BioTechniques, v. 27, N. 6. pp. 1210-1217, 1999.

TORII, M., ARIGHI, C. N., LI, G., WANG, Q., WU, C. H., VIJAY-SHANKER, K. RLIMS-P 2.0: A Generalizable Rule- Based Information Extraction System for Literature Mining of Protein Phosphorylation Information. IEEE/ACM transactions on computational biology and bioinformatics, v.12, n.1, p.17–29, 2015.

TSEYTLIN, E., MITCHELL, K., LEGOWSKI, E., CORRIGAN, J., CHAVAN, G., JACOBSON, R. S. NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. BMC bioinformatics, v.17, n.32, 2016.

TSHITOYAN V, DAGDELEN J, WESTON L, DUNN A, RONG Z, KONONOVA O, PERSSON KA, CEDER G, JAIN A. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95–98, 2019.

TSURUOKA, Y., MIWA, M., HAMAMOTO, K., TSUJII, J., ANANIADOU, S. Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics (Oxford, England), v.27, n.13, p.i111–i119, 2011.

TURINA, P., FARISELLI, P., CAPRIOTTI, E. ThermoScan: Semi-automatic Identification of Protein StabilityData from PubMed. Frontiers in molecular biosciences, v.8, 620475, 2021.

URAMOTO, N., MATSUZAKA, H., NAGANO, T., MURAKAMI, A., TAKEUCHI, H., TAKEDA, K. text-mining system for knowledge discovery from biomedical documents. IBM Syst. J., v.43, p. 516–533, 2010.

VAN LANDEGHEM, S., BJÖRNE, J., WEI, C. H., HAKALA, K., PYYSALO, S., ANANIADOU, S., KAO, H. Y., LU, Z., SALAKOSKI, T., VAN DE PEER, Y., Ginter, F. Large-scale event extraction from literature with multi-level gene normalization. PloS one, v.8(4), e55814, 2013.

VIRTANEN, Pauli et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods, v. 17, n. 3, p. 261-272, 2020.

WEI, C. H., KAO, H. Y., LU, Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic acids research, v.41(Web Server issue), p.W518–W522, 2013.

WOLFSON, M., WALLACE, S. E., MASCA, N., ROWE, G., SHEEHAN, N. A., FERRETTI, V., LAFLAMME, P., TOBIN, M. D., MACLEOD, J., LITTLE, J. et al. RDataSHIELD: resolving a conflict in contemporary bioscience-performing a pooled analysis of individual-level data without sharing the data. International journal ofepidemiology, v.39, n.5, p. 1372–1382, 2010.

YE, Z., TAFTI, A. P., HE, K. Y., WANG, K., HE, M. M. SparkText: Biomedical Text Mining on Big Data Framework. PloS one, v.11(9), e0162721, 2016.

YU, H., KIM, T., OH, J., KO, I., KIM, S., HAN, W. S. Enabling multi-level relevance feedback on PubMed byintegrating rank learning into DBMS. BMC bioinformatics, v.11 Suppl 2, S6. 2010.

ZAFEIROPOULOS, H.; PARAGKAMIAN, S.; NINIDAKIS, S.; PAVLOPOULOS, G. A.; JENSEN, L. J.; PAFILIS, E. PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. Microorganisms, v. 10, n. 2, p. 293, 2022.

ZHANG, L., BERLEANT, D., DING, J., CAO, T., SYRKIN WURTELE, E. PathBinder--text empirics and automatic extraction of biomolecular interactions. BMC bioinformatics, v.10 Suppl 11, S18, 2009.

ZHU, F., PATUMCHAROENPOL, P., ZHANG, C., YANG, Y., CHAN, J., MEECHAI, A., VONGSANGNAK, W. SHEN, B. Biomedical text mining and its applications in cancer research. J. Biomed. Inform., v.46, p. 200–211, 2013.