Biologically oriented mud volcano database: muddy_db

Remizovschi Alexei[1] **and Rahela Carpa**[1]

[1]**Department of Molecular Biology and Biotechnology, Babeș-Bolyai University, Cluj-Napoca, Cluj, Romania**

Corresponding author:
Rahela Carpa[1]

Email address: rahela.carpa@ubbcluj.ro

ABSTRACT

Mud volcanoes (MVs) are naturally occurring hydrocarbon hotbeds with continuous methane discharge contributing to global warming. They host bacterial consortia adapted to hydrocarbon oxidation. Given their research value, MVs still represent a niche topic in microbiology. All the data regarding MVs is sporadic and decentralized. To mitigate this problem, we built a custom NLP pipeline (muddy_mine), and collected all the available MV data from open-access articles. Based on this data, we built the muddy_db database. The muddy_db represents the first biologically oriented database rendered as a user-friendly web app. This database includes all the relevant MV data, ranging from microbial taxonomy to hydrocarbon occurrence and geology. muddy_db is indefinitely available to everyone. It is licensed under the GPLv3.
muddy_db R Shiny web app: https://muddy-db.shinyapps.io/muddy/
muddy_db R package: https://github.com/TracyRage/muddy_db
muddy_mine Conda package: https://github.com/TracyRage/muddy_mine

INTRODUCTION

Mud volcanoes (MVs) represent hydrocarbon discharging landforms (Mazzini and Etiope, 2017). They are distributed worldwide in both marine and terrestrial environments (Milkov, 2000). The most distinctive feature of MVs is recurrent methane emission. Due to methane emissions, MVs contribute extensively to global warming (Etiope et al., 2009).

MV genesis is mainly caused by a naturally mediated process - kerogen maturation (Vandenbroucke and Largeau, 2007). Therefore, the surrounding area of MVs can provide valuable data regarding both aerobic and anaerobic hydrocarbon microbial oxidation (Cheng et al., 2012). Pristine oxidation is not influenced by anthropic factors.

Despite the evident research value, MV microbiology is still a niche topic. The biological data regarding MVs are sporadic and not centralized.

Mainstream biomedical fields have extensively employed natural language processing (NLP) techniques to mine meaningful data (Wang et al., 2020). Simultaneously, the number of databases related to biomedical fields is considerable (Luo et al., 2016). Niche environmental science fields have not caught up.

Fortunately, democratic NLP models and tools have been published over the last years. Some of them can be easily used by environmental scientists with limited computer science (CS) experience, for example, the spaCy library, ScispaCy models, and S2ORC database (Honnibal and Johnson, 2015; Neumann et al., 2019; Lo et al., 2020).

Cumulatively, the latest advancements in NLP can provide opportunities for consolidating and promoting niche environmental topics.

Given these factors, we aimed to build the first biologically oriented mud volcano database, muddy_db, a niche database that consolidates all the relevant biological data, which will be of great use for all the

microbiology researchers. Collaterally, our custom pipeline can serve as a methodological blueprint for research collectives interested in NLP and building their own specialized databases.

MATERIALS & METHODS

To collect all the available data regarding the biological aspects of MVs, we had to exclusively rely on open-access articles. Having these articles, we could freely mine all the biologically flavoured tokens, including taxonomy-, chemicals-, geology-, and MV-specific terms. Additionally, we had to build a custom mining pipeline - muddy_mine (Fig. 1). The scope of muddy_mine is to provide and enrich the muddy_db database with relevant MV data.

Data collection

We used the S2ORC database to collect open-access articles. S2ORC represents a centralized database that includes 12.7 million articles with a fully preserved paper structure. S2ORC is quite comprehensive and includes niche environment science articles (Lo et al., 2020). Given these facts, we extracted all the available MV-related titles (N=118) from the S2ORC.

Token extraction and muddy_mine pipeline

Having MV articles, we proceeded with token extraction using the muddy_mine pipeline.

Taxonomy extraction represented a difficult challenge due to the fact that we intended to collect as many tokens as possible. To overcome this problem, we used the spaCy library, ScispaCy NLP models and the most recent NCBI Taxonomy database (Honnibal and Johnson, 2015; Neumann et al., 2019; Schoch et al., 2020). First, we extracted all the taxon tokens using en_core_sci_sm ScispaCy model. Second, we checked those tokens against a local NCBI Taxonomy database. Thus, we managed to centralize MV-specific taxonomy on all the possible levels: phylum, class, order, family, and genus.

The other non-taxonomy tokens were also extracted with the abovementioned model. We extracted tokens related to the following categories: chemistry (inorganic ions, hydrocarbons), geology (geological periods, minerals), MV terminology (ANME, methanogenesis type), and experimental methods (PCR types, amplified genes, chromatography). The comprehensive list of categories can be consulted by visiting the muddy_db repository.

The raw output of the muddy_mine pipeline represents a set of csv tables with MV data.

Building muddy_db database

By obtaining muddy_mine raw output, we can advance to the next step - building a user-friendly database. To create this kind of database, we created a Shiny web app, entitled muddy_db. This app includes all the output generated by the muddy_mine pipeline. Not only can count MV-related tokens previously mined from the integral article bodies (N=57) be found there but also tokens extracted from the abstracts (N=115). Additionally, we added an annotated map, which displays the geographical distribution of MVs and their affiliated research metadata. The muddy_db app is indefinitely available for everyone. We intend to regularly update it over time.

RESULTS

The scope of the muddy_db is to gather all the available MV biologically relevant data and include it in a user-friendly database. First, we collected all the known taxa associated with MVs. The muddy_db includes data regarding archaeal and bacterial taxonomy on all the possible taxonomy levels. This particularity can facilitate the detection of microbial consortia patterns. Second, we gather information regarding metabolic pathways, geology, hydrocarbon availability and experimental methods performed on MV sediments. This information can guide specialists to implement appropriate research strategies. Database schema could be succincly described as follows:

1. Map (geographical location of mud volcanoes described in literature)
2. Articles (exhaustive list of mined open-access articles):
    - PMID, title, authors, years, journal, doi, mined_level
3. Bacteria and Archaea (bacterial / archaeal mined taxonomy)

- phylum, class, order, family, genus

4. Chemistry (mud volcano related chemical parameters)
    - hydrocarbons, inorganic_ions
5. Geology
    - minerals
6. Mud volcano (mud volcano biological & morphological data)
    - place, morphology, methane_type, metabolics, DAMO, type_methanogenesis, ANME
7. Methods (methods used in mud volcano research area)
    - genes, chromatography, microscopy etc.

DISCUSSION

MVs are considered to be the setting where early life evolved (Pons et al., 2011). They sustain a plethora of bacterial metabolic pathways, ranging from methane oxidation and synthesis to sulphate reduction (Kleindienst et al., 2014; Cheng et al., 2012). Given these facts, MVs should be the main focus of microbiology. Unfortunately, data regarding the biological aspects of MVs are scarce. Additionally, the data already gathered are not combined in a dedicated database. The lack of a specialized MV database determines mud volcano microbiology to be a niche and neglected topic.

Biomedical fields have always represented the cutting-edge subset of natural science. Indefinite accumulation of medical data over the years determined biomedicine to tightly intertwine with the big data term (Luo et al., 2016). However, the implementation of CS methods in niche environmental fields lags. To both apply CS methods in an environmental context and chronically mitigate the data deficient field of MV microbiology, we created a muddy_mine NLP pipeline and muddy_db database.

CONCLUSION

The muddy_db represents the first biologically oriented mud volcano database. It was designed to provide a comprehensive data corpus that can facilitate mud volcano research and shed light on the topic as a whole. The muddy_db contains data ranging from taxonomy to geology and experimental methods. Simultaneously, the muddy_mine NLP pipeline can serve as an example of accessible implementation of NLP techniques in environmental sciences.

REFERENCES

Cheng, T., Chang, Y., Tang, S., Tseng, C., Chiang, P., Chang, K., Sun, C., Chen, Y., Kuo, H., Wang, C., Chu, P., Song, S., Wang, P., and Lin, L. (2012). Metabolic stratification driven by surface and subsurface interactions in a terrestrial mud volcano. *ISME J*, 6(12):2280–2290.

Etiope, G., Feyzullayev, A., and Baciu, C. (2009). Terrestrial methane seeps and mud volcanoes: A global perspective of gas origin. *Mar Pet Geol*, 26(3):333–344.

Honnibal, M. and Johnson, M. (2015). Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics.

Kleindienst, S., Herbst, F., Stagars, M., von Netzer, F., von Bergen, M., Seifert, J., Peplies, J., Amann, R., Musat, F., Lueders, T., and Knittel, K. (2014). Diverse sulfate-reducing bacteria of the desulfosarcina/desulfococcus clade are the key alkane degraders at marine seeps. *ISME J*, 8(10):2029–2044.

Lo, K., Wang, L., Neumann, M., Kinney, R., and Weld, D. (2020). Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics.

Luo, J., Wu, M., Gopukumar, D., and Zhao, Y. (2016). Big data application in biomedical research and health care: A literature review. *Biomed Inform Insights*, 8:BII.S31559.

Mazzini, A. and Etiope, G. (2017). Mud volcanism: An updated review. *Earth Sci Rev*, 168:81–112.

Milkov, A. (2000). Worldwide distribution of submarine mud volcanoes and associated gas hydrates. *Mar Geol*, 167(1-2):29–42.

Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). Proceedings of the 18th bionlp workshop and shared task. Association for Computational Linguistics.

Pons, M., Quitte, G., Fujii, T., Rosing, M., Reynard, B., Moynier, F., Douchet, C., and Albarede, F. (2011). Early archean serpentine mud volcanoes at isua, greenland, as a niche for early life. *Proc. Natl. Acad. Sci. U.S.A.*, 108(43):17639–17643.

Schoch, C., Ciufo, S., Domrachev, M., Hotton, C., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J., Sun, L., Turner, S., and Karsch-Mizrachi, I. (2020). Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020.

Vandenbroucke, M. and Largeau, C. (2007). Kerogen origin, evolution and structure. *Org Geochem*, 38(5):719–833.

Wang, J., Deng, H., Liu, B., Hu, A., Liang, J., Fan, L., Zheng, X., Wang, T., and Lei, J. (2020). Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on pubmed. *J Med Internet Res*, 22(1):e16816.

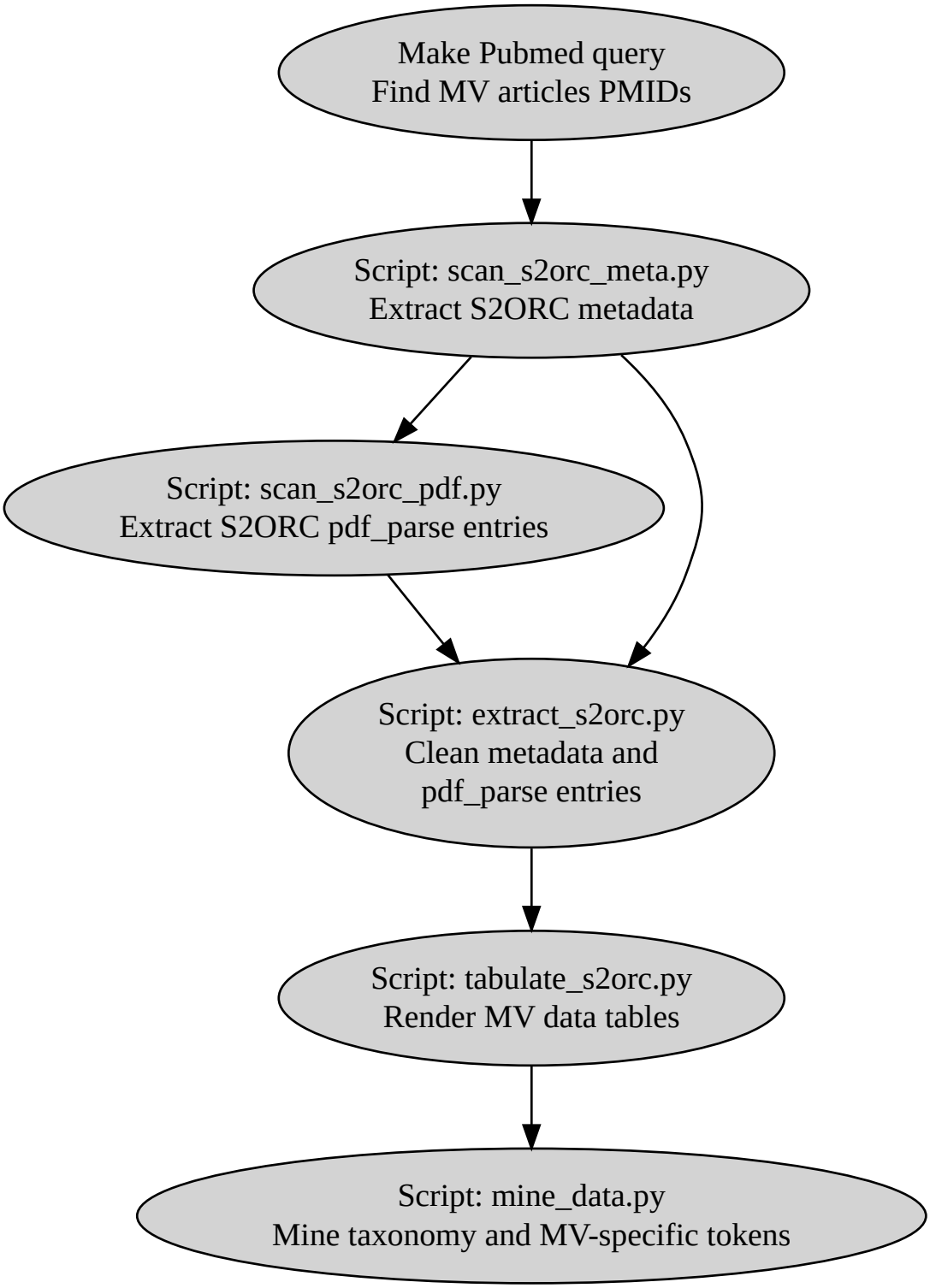

Figure 1. muddy_mine - pipeline used to build the muddy_db database. MV - mud volcano, PMIDs - Pubmed