Taxonomy Classification and Quinone Analysis from Scientific Publications: A Natural Language Processing Approach

Suraj Kanwar

Supervisors: Sophie Abby, Sophie-Carole Chobert, Fabien Pierrel

Abstract

Quinones are diverse organic compounds which play pivotal roles in various biological processes. This report presents a novel approach to understanding quinones' distribution and roles by applying natural language processing techniques to extract, analyze, and visualize data from scientific publications. Utilizing Python scripts, we mined data from the PubMed database, focusing on articles describing bacterial and archaeal species. Our methodology allowed us to extract species and genus names from titles and identify specific quinones and associated keywords in abstracts. Our project's ultimate objective is to compile a master table encapsulating metadata on formally described species and the nature of quinones they produce, establishing a foundation for future investigations and furthering our understanding of the vital function of quinones in a wide-ranging ecosystem. Future work aims to refine our methods for more accurate and comprehensive results, contributing valuable perspectives to the complex mechanisms governing living organisms.

Biological Context

Quinones and Their Biological Roles

Quinones are diverse organic compounds playing a pivotal role in various biological processes. Their functionality span across cellular respiration, energy synthesis, and metabolism essential elements for ATP production [1]. Apart from these roles, quinones also contribute to the synthesis of certain vitamins and hormones. Moreover, their antioxidant capabilities fortify organisms against oxidative damage [1]. Therefore, extracting quinone information from scientific literature such as abstracts can potentially unravel new insights into the mechanics of living organisms.

Quinones and Electron Transport Chains

The significance of quinones also extends to their indispensable role in the formation of electron transport chains which are crucial for the survival and growth of various organisms, including bacteria, fungi, and plants [2]. These chains serve as the foundation for energy conversion, allowing the transfer of energy into a usable form and facilitating essential metabolic processes required for life and growth [2]. Organisms rely on electron transport chains to generate the energy necessary to sustain cellular operations and physiological functions. It's important to note that not all organisms have electron transport chains, and some rely on alternative processes such as fermentation to produce cellular energy.

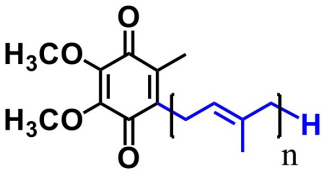
Quinone Structure

Quinone Distribution and Tail Lengths

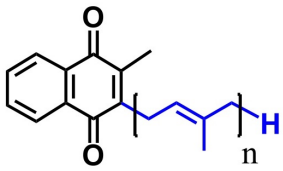
Quinones are made of a ring and tail, the tail and ring can vary and it changes the property of the quinone. Figure 1 represents the visual structure of tail lengths of quinones among different organisms. The distribution of quinones across different organisms is an area of significant biological importance. It provides valuable insights into various aspects of life forms [3]. The identification of new species is often associated with the determination of their quinone repertoire. However, the utility of quinone profiles as a

taxonomic classification tool depends on the level of diversity within a given taxonomic group. If there is too much diversity, quinones may not be effective for accurate classification. Taxonomy is organized into hierarchical ranks, including kingdom, phylum, class, order, family, genus, and species. [4]

By analyzing quinone profiles, researchers can differentiate between different species, and strains, and gain insights into their evolutionary relationships [4]. The tail lengths of quinones across different organisms bear substantial biological importance 1. They provide information concerning growth conditions, metabolic pathways, and taxonomic classifications of various life forms [3]. An analysis of the quinone profiles, including the variations in tail lengths, can aid researchers in differentiating species, strains, and discerning evolutionary relationships [4].



(a) Ubiquinone structure

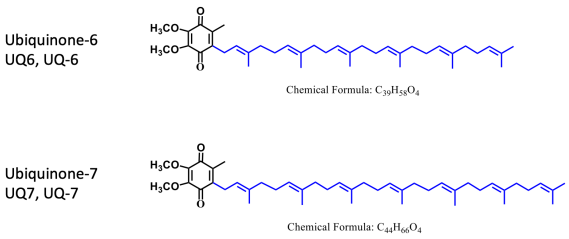


(b) Menaquinone structure

Figure 1: This illustration provides a visual representation of some quinone structures.

Conservation Implications

The chemical structure, as depicted in Figure 1, showing the basic chemical structures of ubiquinone and menaquinone. The figure 2 also depicts ubiquinones with different tail lengths, further illustrating the structural diversity of these compounds.



(a) Ubiquinone structure with six and seven carbon tails and nomenclature

Figure 2: This figure illustrates the varied nomenclature for quinone representation.

State of the Art

Significance of Work

Our work seeks to fill a knowledge gap since the seminal 1981 paper titled "Distribution of Isoprenoid Quinone Structural Types in Bacteria and Their Taxonomic Implications" by Collins and Jones [3]. The paper provided a comprehensive analysis of the distribution of isoprenoid quinone structural types in bacteria. However, there have been considerable advancements in the field since then [4]. By automatically extracting previously uncollected quinone data where new species are described in scientific journals, we aim to contribute significantly to updating and enriching our knowledge base.

Undiscovered Information

Scientific publications are a large and wide repository of knowledge that holds a lot of previously unexplored material about the interesting and intricate world of quinones. These papers contain a wealth of data, research, and insights waiting to be found and investigated, ranging from the most current discoveries on the chemical characteristics and behavior of quinones to the most recent discoveries on their possible applications in many industries [4]. By carefully studying these sources, we can get a comprehensive grasp of the intricate mechanisms that underpin the behavior and capabilities of quinones, as well as discover fresh applications for them [4]. These findings

can then be used to inform and shape future research attempts, resulting in a more thorough and nuanced understanding of the subject at issue.

TIMC Labs Motivations

Exploration of Quinones

The TrEE team at the TIMC lab is motivated by various factors, including the exploration of complex quinone production processes, respiratory chains, and substrate preferences across diverse organisms. This thorough analysis sheds light on the crucial enzymes and genes involved in the complex process of quinone production. An in-depth investigation of these enzymes and genes provides valuable insights into their physiological roles and the intricate metabolic pathways that support quinone synthesis, revealing the complex interplay between different components and advancing our understanding of quinones. This research strives to gain a nuanced comprehension of the intricate biological processes governing quinone synthesis, leading to the development of innovative approaches in the fascinating field of biochemistry.

Investigations

One of the goals of the TrEE team is to perform an in-depth examination of the growth and colonization circumstances of diverse species. This is accomplished by thoroughly studying their quinone profiles, which provide insights into their distinct traits and properties. The team can obtain a better grasp of the exact conditions required for these organisms to live and grow by delving further into their quinone profiles. This knowledge can then be utilized to create more successful ways of cultivating and studying these organisms, leading to important advances in the discipline of microbiology. Examining the composition of quinones can yield valuable insights into the adaptive strategies, ecological interactions, and survival mechanisms of microbial communities inhabiting various environments.

Computational Approach

We relied on Pythons' vast library sets such as NLTK (Natural Language Toolkit) and ETE3 (Evolutive Tree Explorer) to streamline our analysis process. NLTK provided us with a suite of text processing and analysis tools that were instrumental in keyword extraction, text tokenization, and contextual analysis. ETE3, on the other hand, offered an efficient means to examine and manipulate taxonomy profiles, thereby helping us understand and represent the relationship among various species and genera.

Literature Review

Information Collection

We observed that the abstracts of journals contained valuable information about quinones, which motivated us to gather this data. While the titles primarily contained species and genus information, the abstracts provided specific details about the quinones. We started by gathering the data from the PubMed database, a free database of references and abstracts on life sciences and biomedical topics maintained by the National Center for Biotechnology Information (NCBI). We specifically focus on two journals: Antonie van Leeuwenhoek and the International Journal of Systematic and Evolutionary Microbiology. These journals were chosen in particular for their extensive coverage of articles describing the quinone profiles of bacterial and archaeal species. To streamline the data analysis process, we developed an efficient Python script that converts XML files from the March 2023 version of the PubMed database using the available FTP download provided by NCBI. Our script automated the data extraction process, saving significant time and resources compared to manual extraction. The panda's DataFrame layout played a crucial role in data management and analysis, offering a structured and versatile framework for various transformation and analysis tasks. Overall, we continually developed the script to allow us to convert valuable information into an organized table.

Examining the Journals

The data mining process we developed involved parsing numerous XML files, which greatly differed in their size and complexity. Computation time for this task was influenced by several factors, including the number of files processed, the complexity of the files, and the efficiency of our data processing code to look through the 1166 PubMed XML files took 174.9 minutes. We have tried to optimize the data processing techniques to ensure swift and efficient data mining, thereby minimizing computation time while maximizing output which can further be parallelly implemented given more time. Figure 3 visually depicts the distribution of articles before we initialize the search.

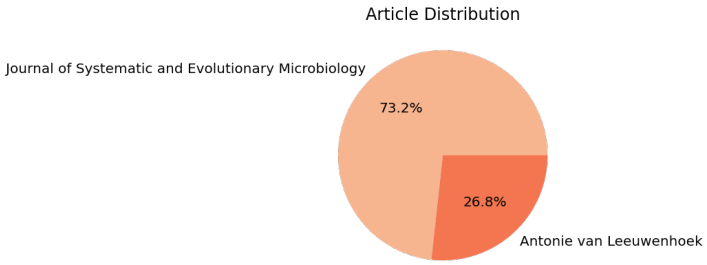


Figure 3: The pie chart shows that the International Journal of Systematic and Evolutionary Microbiology has more articles when compared to Antonie van Leeuwenhoek.

Methodology

Furthermore, our exploration extended to alternative approaches for data examination. Particularly, when extracting species information from titles, we experimented with different techniques, such as identifying non-dictionary words and detecting Latin suffixes and prefixes. Eventually, considering computational complexity, we adopted a regex pattern that captures capitalized words followed by lowercase words. Although this approach occasionally included negative matches, such as capitalized nouns in the title, we mitigated this issue during later stages by utilizing the NCBI taxonomy API.

Extraction Process

Extraction of Quinones

Throughout the internship, we optimized the script to perform essential operations accurately and rapidly. The titles and abstracts are frequently laden with critical keywords and contextual hints that might assist researchers in better understanding the complicated relationships that exist between various biological bacteria. Throughout the internship, we designed and experimented with implementing various data analysis methods that focused on extracting as much meaningful and relevant information from the titles and abstracts.

Matching Quinones Patterns

Using a strategic approach combining word tokenization and regular expression (regex) patterns, we effectively identified quinones in the content. Regex patterns were crucial for handling complex quinone nomenclature, allowing a flexible and accurate match. This method resulted in precise quinone identification within the dataset. To facilitate the process of capturing the quinone keywords, a robust dictionary was created of regex patterns corresponding to various quinone naming conventions. This dictionary comprises various regex patterns, each corresponding to the distinct naming conventions of different types of quinones. The dictionary accounts for a broad spectrum of nomenclature and notation practices, including alternate names, abbreviations, and specific terminologies associated with quinones. The strength of this method lies in its ability to not only recognize a wide array of quinone names but also discern their presence within the titles and abstracts of relevant literature. Manual studying of data and attention to detail ensure that as much of the pertinent information is appropriately considered and analyzed.

Improvements to Dictionary

The continual refinement process of the dictionary used to match quinones and their related keywords proved to be an important component throughout our

investigation. This procedure entailed a thorough evaluation of the dictionary, intending to refine and enhance its correctness over time. We were able to obtain a better level of precision and reliability in our findings after plenty of manual attention to the negative results. We painstakingly added extra keywords to our project over numerous revisions, ensuring that we accounted for any conceivable spelling changes. We also paid close attention to synonyms and alternative terminology, ensuring that our final result was as complete and accurate as possible. This procedure required meticulous attention to detail and a solid understanding of the subject matter, working closely with the TIMC team helped to understand the biological aspects of what we were extracting. **A sample of the final dictionary used for the quinone search is presented below:**

"q": ["quinone", "quinone−{}", "quinone {}", "q{}", "q−{}", "{},{}−benzoquinone", "{},{}−bq", "{},{}−dimethyl−{},{}−benzoquinone", "{},{}−dimethyl−{},{}−bq", "{},{}−d−{},{}−benzoquinone", "{},{}−d−{},{}−bq, etc..."]
"mk": ["menaquinone", "menaquinone {}", "MK{}", "MK{}H{}", "menaquinone−{}", "menaquinone−{}(H{})", "mk−{}", "mk{}H{}", etc...]
"quinone...": ["nomenclature.."]

The dictionary works by mapping variations of the same quinone nomenclature. The placeholder "{}" used in the dictionary stands for a variable part of the keyword, which can represent different parts of the nomenclature associated with the chemical compounds.

Extraction of Keywords

In our comprehensive analysis of the abstracts, we successfully extracted context-specific keywords to gain a deeper understanding of quinone discovery. By extracting relevant keywords, we obtained valuable contextual information and insights into the main findings of the publications. Additionally, we gathered data on quinone types, keyword sentences, quinone tail lengths, and, when present, the hydrogen saturation of quinone bonds. As a result, our findings led to nuanced conclusions and informed decisions for subsequent data analysis.

Results

Taxonomic Classification

One of the significant accomplishments during our internship was gaining a deeper understanding of the taxonomy and quinone profiles of various species and genera. To achieve this, we developed a script that utilizes the ETE3 NCBI python library to search for the most recent names of the species and genera extracted. Before classification, the script updates the taxonomy to ensure we have the most up-to-date information. The resulting taxonomy provides valuable insights into the distribution of quinones across different taxonomic ranks, allowing us to examine the diversity and prevalence of quinone types among various organisms. Figure 4 illustrates the abundance of quinones at the "class" level of the taxonomy, with the x-axis representing the taxonomy level and the y-axis indicating the count of quinones.

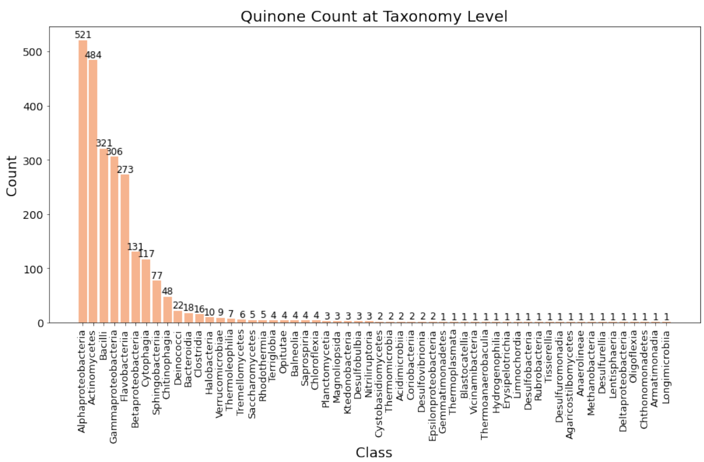


Figure 4: Distribution of the number of abstracts with quinones at the "class" level of the taxonomy. The x-axis denotes the taxonomy level and the y-axis represents the abstracts with information on quinones.

Discussion

Since the publication by Collins and Jones [3] on this topic in 1981, no additional research has been conducted in this area. Our automated analysis, founded

by the manual study of the language and nomenclature used in the research papers, revealed that we were able to extract previously untouched information. In our investigation, we successfully developed an automated method to create a comprehensive metatable using only the titles and abstracts from the selected journals. By extracting species and genus information from the titles and utilizing ETE3, we constructed taxonomies for each mentioned species and genus. The extracted table has the title, the abstract, the species and genus, the type of quinone, quinone, the keyword sentence, the length(s) of the tail(s), the hydrogen saturation (if present), and the taxonomy of the species and genera. These final results provided relevant insights from the research articles and automating this process, it has enabled us to contribute valuable insights to the field.

The following is a formatted output of our extraction:

Journal: International journal of systematic and evolutionary microbiology
PubMedID: 18398207
Species and Genus: Aeromicrobium ponti
Taxonomy: [{'class': 'Actinomycetes', 'genus': ' Aeromicrobium', 'order': 'Propionibacteriales', ' family': 'Nocardioidaceae', 'phylum': ' Actinomycetota', 'species': 'Aeromicrobium ponti '}]
Quinone Found: 'mk': (menaquinone, MK—9)
Keyword Sentence: The predominant menaquinone was MK—9(H4).
Quinone Tail: [9]
Hydrogen Saturation: ['H4']

Difficulties and Limitations

Throughout our internship, we faced many difficult problems that required the use of our problem-solving and technical skills. One of the challenges we faced was a lack of consistency in the way abstracts are written, specifically the way the quinone nomenclature was written, which took a significant amount of time and effort to extract. We also had issues improving the matching algorithms, which required a systematic and detailed approach to ensure accuracy and precision. The technical challenge consisted of finding, diagnosing, and fixing many coding errors. This assignment required a lot of detail and a manual scan of positive and negative results to improve the search.

Conclusion and Future Work

In conclusion, the metatable generated through our research holds valuable information that can be utilized for cross-validation with the TrEE Team. We aim to compare the extracted quinone data with the predictions made by the TrEE Team, allowing us to identify any missing quinones and validate our results. The computational data analysis techniques employed, including data mining, analysis, and visualization, have been instrumental in extracting and interpreting critical information from vast databases, enhancing accuracy, and improving our understanding of complex systems. Throughout the internship, we strategically analyzed and addressed various challenges associated with the data. The team's dedication and aid have yielded favorable outcomes despite obstacles. Furthermore, the development of a metatable has provided additional insights into the intricate relationships between species and their associated quinone profiles. Overall, our examination of quinones and the computational methodologies employed have contributed valuable perspectives on the complex mechanisms governing living organisms, enriching our understanding of the natural world. In future work, we plan to collaborate closely with the TrEE Team and continue refining our methods for even more accurate and comprehensive results.

References

[1] Monks TJ, Jones DC, The effects of quinones on cellular redox status. Free Radic. Biol. Med. 12(5): 463–490, 1992.

[2] Chenuil A, Adrien F, Agüera A, Bejaoui N, Boudouresque CF, Denis JL, et al. (2010) The fate of marine biodiversity in the Mediterranean Sea: Setting priorities for conservation actions. PLoS ONE 5(9): e12179. https://doi.org/10.1371/journal.pone.0012179

[3] Collins MD, Jones D (1981) Distribution of Isoprenoid Quinone Structural Types in Bacteria and Their Taxonomic Implications. Microbiol Rev 45(2): 316-354.

[4] Hiraishi A, Ueda Y, Ishihara J (2002) Quinone profiling of bacterial communities in natural and synthetic sewage activated sludge for enhanced phosphate removal. Water Res 36(17): 4227-4238.