engineering proceedings

MDPI

Proceeding Paper

Computational Representation of Cellular Lines: A Text Mining Approach †

Ivan Carrera *, Henry Guanoluisa and Alexis Miranda

Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito 170525, Ecuador; henry.guanoluisa@epn.edu.ec (H.G.); alexis.miranda@epn.edu.ec (A.M.)

* Correspondence: ivan.carrera@epn.edu.ec

† Presented at the XXXI Conference on Electrical and Electronic Engineering, Quito, Ecuador, 29 November–1 December 2023.

Abstract: In the rapidly evolving landscape of cancer drug research, cellular lines serve as invaluable tools for understanding drug-sensitive and drug-resistant tumors. The computational representation of cellular lines is usually based on genomic profiling, even though this method cannot be applied in a large scale. This study introduces a novel approach to the computational representation of cellular lines using text mining techniques. By meticulously extracting and analyzing textual data from the scientific literature, we developed a computational representation of these cellular lines. Our methodology encompassed advanced Natural Language Processing (NLP) for text extraction and machine learning models for predictive analysis. We achieved a comprehensive description of each cellular line. To validate our findings, we generated a distance matrix for all cellular lines, leading to the construction of a dendrogram representing cellular line relationships. This dendrogram shows a resemblance with the established cell line ontology from CLO. Our results bridge the gap between cellular line representation and text mining, offering a robust computational model that can significantly impact cancer drug research.

Keywords: text mining; natural language processing; predictive modeling; machine learning; cellular line representation; drug response prediction; personalized medicine

check for updates

Citation: Carrera, I.; Guanoluisa, H.; Miranda, A. Computational Representation of Cellular Lines: A Text Mining Approach. *Eng. Proc.* **2023**, *47*, 13. https://doi.org/10.3390/engproc2023047013

Academic Editor: Soraya Lucia Sinche Maita

Published: 4 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cellular lines are indispensable tools in cancer research. These in-vitro models offer a controlled environment to study the biology of cancer cells, test potential therapeutic agents, and understand drug resistance mechanisms. Cellular lines have been instrumental in elucidating the molecular pathways of carcinogenesis and have paved the way for the development of targeted therapies. Their significance is underscored by the vast body of scientific literature dedicated to their study, emphasizing their role in advancing our understanding of cancer and its treatment.

The drug discovery process can be leveraged by Artificial Intelligence [1]. Furthermore, AI fosters the emergence of Computational Precision Medicine, allowing the design of therapies tailored to individual patients' physiology, disease features, and environmental exposures [2]. The combination of AI with big data and advanced computing has the potential to revolutionize evidence-based, personalized medicine, making treatments more efficient and tailored to individual needs [3].

As the volume of data related to cellular lines continues to grow, there is an increasing need for efficient computational representation to manage and analyze this information. Text mining, a subset of data mining, offers a promising solution. By extracting valuable insights from vast amounts of textual data, text mining can help researchers identify patterns, trends, and relationships that might otherwise go unnoticed [4]. In the context of cellular lines, text mining can facilitate the identification of novel drug targets, elucidate mechanisms of drug resistance, and streamline the drug discovery process.

Despite the wealth of information available on cellular lines, there remains a gap in the systematic computational representation of these data.

This research aims to present a comprehensive computational representation of cellular lines. Our aims are to: (1) extract relevant information from textual data sources related to cellular lines, (2) develop a computational model that captures the intricacies of cellular line data, and (3) validate the relevance of the extracted information in the context of cancer drug research.

2. Motivation

The computational representation of cellular lines has been a topic of interest in the scientific community in the last 15 years. However, there is no consensus on the most effective method for achieving this representation. The most prevalent technique employed to date is genomic representation. Though this approach offers valuable insights into the genetic makeup of cellular lines, it falls short of providing a comprehensive view. For instance, a study by Li et al. found that established cell lines are generally a poor representation of primary tumor biology, indicating that genomic representation might not capture the full spectrum of cellular characteristics [5]. Genomic representation primarily focuses on the genetic sequences and variations, often overlooking other crucial aspects such as cellular behaviors, interactions, and responses to various treatments.

In contrast, other domains within bioinformatics have witnessed significant advancements through the integration of text mining techniques. Text mining has been employed to explore publication trends in various biomedical areas, facilitating the discovery of new insights and relationships [6]. Domains such as proteomics, genomics, and systems biology have successfully employed text mining to uncover patterns, trends, and relationships, leading to novel insights and a deeper understanding of complex biological systems. Kitano discussed the integration of mathematical modeling, molecular interaction networks, and cellular structure physics, suggesting that achieving computational models predicting cellular system behaviors is increasingly feasible [7]. Yun et al. combined text mining with gene expression analysis to reveal a relationship between a specific molecule and the invasiveness of a glioblastoma cell line [6].

Thus, given the successes of text mining in these domains, it presents a compelling unaddressed approach for its application in the computational representation of cellular lines. By leveraging text mining, there is potential to extract a more holistic and nuanced understanding of cellular lines, bridging the gap left by genomic representation and offering a more comprehensive view of cellular dynamics and characteristics.

3. Materials and Methods

In our research, we employed a systematic methodology to achieve a comprehensive representation of cellular lines. Our data sources were the Cellosaurus database and PubMed, from which relevant information on cellular lines was extracted. To process these data, we utilized text mining techniques, specifically Term Frequency-Inverse Document Frequency (TF-IDF) and Support Vector Domain Description (SVDD). Following the extraction of features, we computed similarities between cell lines based on their SVDD representations. This facilitated the construction of a dendrogram, providing a hierarchical representation that elucidates the relationships between various cell lines. This structured approach ensured both the depth and breadth of our analysis, leading to meaningful insights into cellular line representations. Our methodology is represented in Figure 1.

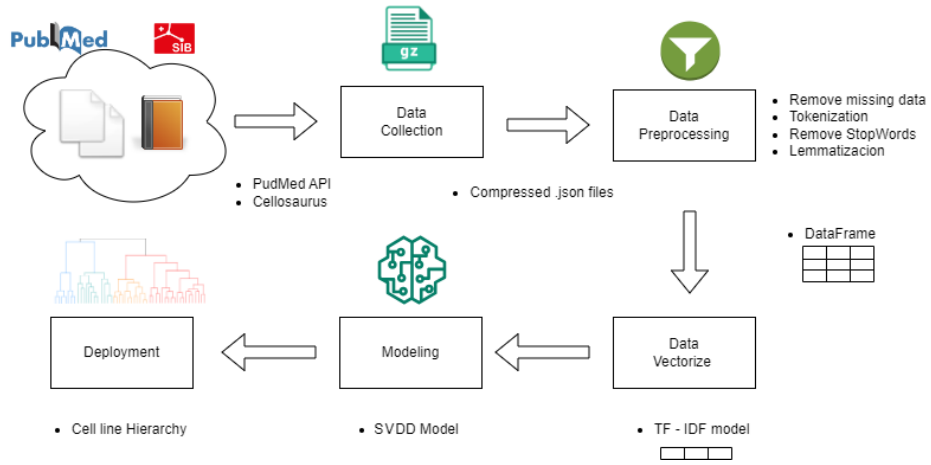


Figure 1. Methodology for data extraction and representation.

3.1. Data Sources

The data for this research were sourced from two major databases. The list of cellular lines was extracted from Cellosaurus, a comprehensive cell line database that presents a large collection of cell line names and their synonyms [8]. For the descriptions of these cellular lines, we turned to PubMed, a database of scientific papers in the biomedical domain [9]. In PubMed, we searched for all the papers that refer to all known names and synonyms of cellular lines.

In our research, the Cellosaurus database was pivotal in providing both the primary names and associated synonyms of cellular lines. Upon extraction, all synonyms for a given cellular line were consolidated into a unified list. This ensured that any mention of the cellular line, regardless of the specific name or synonym used, could be identified and processed. Recognizing the variability in naming conventions, we consolidated all synonyms for each cellular line into a unified list. During data extraction from sources like PubMed, this list ensured the comprehensive capture of relevant data, regardless of naming variations. All synonyms underwent normalization to a standard format, removing discrepancies like special characters or capitalization differences. By systematically managing and incorporating these synonyms, we ensured thorough and accurate data extraction, capturing the entirety of available knowledge on the studied cellular lines. This approach was taken aiming to capture a broad spectrum of the literature, encompassing most cellular lines and their descriptions. In Figure 2, we can see an example for a query using the PubMed API. Our corpus comprises 266,790 papers from 21,844 cell lines.

```
for index, row in df_queries.iterrows():
    try:
        string_search = Entrez.read(Entrez.esearch(db="pubmed", term=row['query'] +
                                                   ' AND ("cell line" OR "cell-line" OR "cellular line")'))
        time.sleep(1)
        pubmed_ids = [int(id) for id in string_search['IdList']]
        paper_id_list.append({'cell_line': row['cell_line'], 'pubmed_ids': pubmed_ids, 'num_ids': len(pubmed_ids)})
        print("Linea Celular " + row['cell_line'])
    except Exception as e:
        print(f"Error Linea Celular {row['cell_line']}: {str(e)}")
        error_paper_id_list.append({'cell_line': row['cell_line'], 'pubmed_ids': [], 'num_ids': 0})
```

Figure 2. An example of a query using the PubMed API.

3.2. Text Mining Techniques

After extracting abstracts related to cell lines from PubMed, we utilized the Term Frequency-Inverse Document Frequency (TF-IDF) method to transform each abstract into a numerical vector, as depicted in Figure 3. The selection of the TF-IDF method for our research was driven by its ability to efficiently evaluate the importance of words within text descriptions relative to their frequency across multiple documents. TF-IDF not only emphasizes the relevance and significance of terms, but also offers scalability

and dimensionality reduction. This ensures that the most pertinent terms are highlighted, filtering out common terms and focusing on those that are truly significant, especially given the extensive data sourced from databases like PubMed.

	aa	aaa	aaaf	aaatcaatatt	aab	aabcl	aac	aacocf	aacr	aacrjourn	...	zymosan	zynaxi	zyx	zyxin	zz	zzr	zzui	zzuneui
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
266,787	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
266,788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
266,789	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 3. Textual data are transformed into numerical data using TF-IDF. Every row corresponds to a single abstract.

Following this, the Support Vector Domain Description (SVDD) technique was applied to encapsulate the abstract instances into a unified computational representation for each cell line. Unlike traditional support vector machines, SVDD creates boundaries around data points in high-dimensional spaces, representing each cellular line as a distinct sphere, visualized in Figure 4. SVDD is a technique that allows for the creation of a boundary around data points in a high-dimensional space. By using SVDD, we were able to obtain a singular, comprehensive description for each cellular line. The synergy of TF-IDF and SVDD provided a robust methodology, delivering comprehensive insights into cellular lines.

	cell_line	center	radius
0	CVCL_0028	[0.00251791017518784, 0.007314898799435672, -0...	0.066173
1	CVCL_0080	[0.019450449413684838, 0.013177261231692031, -...	0.054771
2	CVCL_0081	[-0.00948665032411371, 0.027937856571742012, 0...	0.046550
3	CVCL_0082	[0.023437894045063443, 0.006965141811858809, 0...	0.038501
4	CVCL_0107	[0.005348316471685562, 0.01594712288027022, 0....	0.098690
...	...	...	...
9243	CVCL_ZZ58	[-0.007880641290901352, -0.0014381966014572116...	0.897165
9244	CVCL_ZZ59	[-0.007880641290901352, -0.0014381966014572116...	0.897165
9245	CVCL_ZZ60	[-0.007880641290901363, -0.0014381966014572116...	0.897165
9246	CVCL_ZZ61	[-0.007880641290901363, -0.0014381966014572116...	0.897165
9247	CVCL_ZZ70	[-0.0037931607355641476, -0.008175344523148489...	0.892744

Figure 4. Cell line spherical representation.

4. Results

4.1. Cell Line Hierarchy

Upon successfully acquiring the spherical representations of the cell lines, we proceeded to quantitatively assess the inter-relations between these spheres. This was achieved by calculating the distances between each spherical representation, providing a metric to gauge their relative proximities and distinctions.

The resulting distance metrics facilitated the construction of a hierarchical representation of the cell lines. Intriguingly, this derived hierarchy exhibited a striking resemblance to the established Cell Line Ontology. Such an alignment underscores the potential of our methodology to naturally capture and reflect the intrinsic classifications and relationships inherent to cellular lines within recognized biological frameworks. We employed the `scipy.cluster.hierarchy` library in Python. This library provided the necessary tools for hierarchical clustering, enabling us to create the dendrogram based on the computed distances between cellular lines, ensuring both precision and reproducibility in our representation. We obtained a dendogram for showing the relations between cell lines, shown in Figure 5. The different colors in the figure represent clusters of cellular lines.

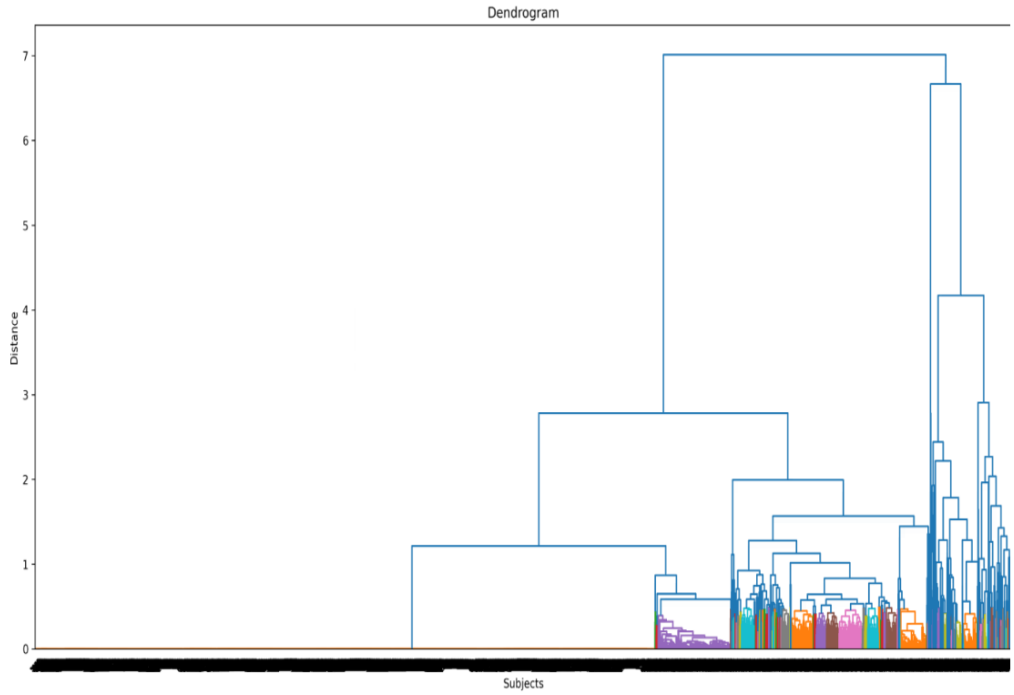


Figure 5. Cell line hierarchical representation.

4.2. Discussion

The computational representation of cellular lines, as visualized through spheres in a hyperspace, provided profound insights into the inherent relationships and distinctions among various cellular lines. By quantifying the distances between these spheres, we were able to discern patterns and clusters, indicating groups of cellular lines with similar characteristics and behaviors. This spatial representation not only facilitated a deeper understanding of the cellular lines in isolation, but also in relation to one another, highlighting potential synergies and differences that might not be evident through traditional analytical methods.

Our findings have significant ramifications for drug testing. The ability to computationally represent and compare cellular lines provides a foundation for predicting their responses to various drugs. By understanding the inherent characteristics of a cellular line, researchers can potentially forecast its reaction to a specific therapeutic agent, streamlining the drug testing process. This could lead to more efficient drug trials, as compounds can be tested on cellular lines that are computationally predicted to be responsive, thereby reducing the number of ineffective trials and accelerating the discovery of potent drugs.

Though our methodology and findings are promising, they are not without limitations. The reliance on textual abstracts from the PubMed library means that our data are only as comprehensive as the abstracts themselves. Important details and nuances present in the full text of articles might be overlooked. Additionally, the dynamic nature of scientific research means that new findings are continuously emerging, and our representation might not capture the very latest advancements in the field.

Our research represents a significant step forward in the computational representation of cellular lines and offers a foundation for future studies in this domain. By bridging the gap between textual data and computational analysis, we hope to drive advancements in drug testing and personalized medicine, ultimately benefiting patients and the broader medical community.

5. Conclusions and Future Work

Our research demonstrates that text mining is not only feasible, but also effective in deriving a computational representation of cellular lines. The techniques employed

allowed for the extraction, processing, and representation of vast amounts of textual data related to cellular lines.

The similarity between the derived hierarchy from our text mining approach and the Cell Line Ontology underscores the validity of our methodology. It suggests that text-based data, when processed appropriately, can yield representations that are in alignment with established biological classifications.

As the demand for analyzing vast sources of text continues to grow, text mining applications are poised to play an increasingly pivotal role in biomedicine. Future work can improve the presented methodology, exploring additional text mining tools and techniques that can enhance the accuracy and depth of the extracted information.

Author Contributions: Conceptualization, I.C.; methodology, I.C.; software, H.G. and A.M.; writing—original draft preparation, I.C.; writing—review and editing, I.C.; project administration, I.C.; funding acquisition, I.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Escuela Politécnica Nacional grant number PII-DICC-2023-01.

Data Availability Statement: Data and codes are available at https://github.com/ivan-carrera/engproc2023 (accessed on 12 October 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
CLO	Cell Line Ontology
TF-IDF	Term Frequency-Inverse Document Frequency
SVDD	Support Vector Domain Description

References

1. Ho, D. Artificial intelligence in cancer therapy. *Science* **2020**, *367*, 982–983. [CrossRef] [PubMed]
2. Moingeon, P.; Kuenemann, M.A.; Guedj, M. Artificial Intelligence-enhanced drug design and development: Toward a computational precision medicine. *Drug Discov. Today* **2021**, *27*, 215–222. [CrossRef] [PubMed]
3. Dilsizian, S.E.; Siegel, E.L. Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment. *Curr. Cardiol. Rep.* **2013**, *16*, 441. [CrossRef] [PubMed]
4. Madhugiri, V.S.; Ambekar, S.; Strom, S.F.; Nanda, A. A technique to identify core journals for neurosurgery using citation scatter analysis and the Bradford distribution across neurosurgery journals. *J. Neurosurg.* **2013**, *119*, 1274–1287. [CrossRef] [PubMed]
5. Li, A.; Walling, J.; Kotliarov, Y.; Center, A.; Steed, M.; Ahn, S.; Rosenblum, M.; Mikkelsen, T.; Zenklusen, J.; Fine, H. Genomic Changes and Gene Expression Profiles Reveal That Established Glioma Cell Lines Are Poorly Representative of Primary Human Gliomas. *Mol. Cancer Res.* **2008**, *6*, 21–30. [CrossRef] [PubMed]
6. Yun, C.; Katchko, K.M.; Schallmo, M.S.; Jeong, S.; Yun, J.; Chen, C.H.; Weiner, J.A.; Park, C.; George, A.; Stupp, S.I.; et al. Aryl Hydrocarbon Receptor Antagonists Mitigate the Effects of Dioxin on Critical Cellular Functions in Differentiating Human Osteoblast-Like Cells. *Int. J. Mol. Sci.* **2018**, *19*, 225. [CrossRef] [PubMed]
7. Kitano, H. Computational cellular dynamics: A network–physics integral. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 163–163. [CrossRef]
8. Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *J. Biomol. Tech. JBT* **2018**, *29*, 25–38. [CrossRef] [PubMed]
9. National Center for Biotechnology Information. 2023. Available online: https://www.ncbi.nlm.nih.gov/pubmed/ (accessed on 31 August 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.