CoreKB: A Web-based Platform for Searching Reliable Facts over a Medical Knowledge Base*

Fabio Giachelle[1,*], Stefano Marchesin[1], Gianmaria Silvello[1] and Omar Alonso[2,†]

[1]Department of Information Engineering, University of Padua, Padua, Italy
[2]Amazon, Santa Clara, California, USA

Abstract

CoreKB is a web-based platform enabling users to search for reliable scientific facts concerning gene expression-cancer associations over a medical Knowledge Base (KB). CoreKB provides a streamlined interface for searching either using natural language queries or by exploiting structured facets providing autocomplete facilities. It is designed to simplify information access and search of scientific facts targeting healthcare stakeholders (i.e., clinicians, physicians, and researchers). CoreKB aims at presenting the user a comprehensive overview of the scientific evidence supporting a medical fact, fully connected with ontology-based entities and well-defined literature resources. In addition, CoreKB provides the user a quantitative comparison of the possible gene-cancer associations related to a specific fact, thus enabling users to assess the degree of agreement among the evidence support.

Keywords

Knowledge Discovery, Fact Search, Gene Cancer Associations

1. Introduction

The World Health Organization has identified cancer prevention as a critical public health challenge of the 21st century, with an estimated 32% increase in cancer cases by 2040[1]. To address this challenge, cancer research has increasingly relied on microarray and next-generation sequencing technologies, generating vast amounts of experimental data on gene expression-cancer interactions [2, 3], which is crucial for cancer diagnostics, prognosis, and therapies [4, 5]. However, the huge amounts of data and their heterogeneous nature poses hindrances to effective analyses and secondary data re-use. In this regard, scientific literature is a critical source to complement and validate this data. Nevertheless, extracting, combining, and interpreting multiple scientific claims to produce an explanatory overview concerning a fact is a challenging and time-consuming task [6]. Yet, it is crucial to shed a light on the subject and provide useful insights – thus simplifying its comprehension, providing deeper understanding, and making it more accessible.

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy
*This is an extended abstract for [1].
*Corresponding author.
†Work done prior to joining Amazon
fabio.giachelle@unipd.it (F. Giachelle)
0000-0001-5015-5498 (F. Giachelle); 0000-0003-0362-5893 (S. Marchesin); 0000-0002-0877-7063 (G. Silvello)

[1]https://gco.iarc.fr/tomorrow/en/dataviz/bubbles?sexes=0&mode=population

In this context, efficient and reliable methods to store and organize knowledge from various sources have become increasingly important, and KBs have emerged as essential resources for cancer researchers [7, 8]. KBs are designed to arrange structured information that can be used to support data-driven research. To this aim, a graph-based representation is usually adopted; it consists of nodes representing concepts/entities whereas the edges represent relationships among them. Specifically, the Resource Description Framework (RDF) format can be used to provide an accessible, interoperable, and machine-readable representation of the underlying data that can be queried with SPARQL [9]. However, searching within KBs can be a challenging task for multiple reasons. First, KBs may lack a fixed schema, making it difficult to comprehend the organization of data and formulate SPARQL queries to extract the required information. Furthermore, the absence of a fixed schema may result in uncertainty and confusion while interpreting search results. Secondly, even for expert users, looking for specific information in a KB via SPARQL queries can be cumbersome and time-consuming [10].

Moreover, the massive amount of information contained in a KB can pose hindrances in locating and extracting specific data of interest. This is compounded by the technical nature of the queries needed to retrieve the desired information, resulting in an additional layer of complexity, especially for non-technologically savvy users [10, 11, 12].

Cancer research also relies on specialized KBs adopting domain-specific jargon, ontologies, and terminologies resources with a large presence of complex terms, acronyms, and morphosyntactic variants, thus making query formulation and effective search even more challenging [13].

As a consequence, researchers may encounter difficulties when attempting to retrieve information of interest from KBs, thus limiting effective data discovery, re-use, and new insights. Therefore, there is an urgent need for intuitive and user-friendly search applications [10] that can facilitate and speed up the retrieval of relevant data over KBs.

To address these issues, we propose CoreKB, an easy-to-use web-based search platform that builds upon one of the largest KBs containing fine-grained facts about gene expression-cancer associations [14]. The KB comprises over 230,000 reliable and unreliable facts extracted semi-automatically from the scientific literature by mining PubMed.

CoreKB is publicly available at http://gda.dei.unipd.it along with a demonstration video (http://gda.dei.unipd.it/static/videos/demo.mp4) presenting its features. The platform provides a streamlined interface for searching and exploring knowledge about gene expression-cancer associations. CoreKB allows users to query the system using either natural language queries or facets, enabling search for both entities and facts easily. In this regard, CoreKB provides several features (e.g., infometrics and entity cards) to support specialized users, such as medical researchers and clinicians, who are interested in searching for knowledge about gene expression-cancer associations quickly and without requiring the exact terminology or entity identifiers.

As a distinctive trait, CoreKB embraces a fact-oriented approach, which aims at providing users with a comprehensive overview of the key information concerning each fact. The overview provided consists of aggregated data including (i) the gene class distribution among the sentences extracted from the scientific literature; (ii) the number of supporting and conflicting sentences (iii) and the number of supporting publications per year, enabling users to assess the consensus supporting a fact. In summary, CoreKB offers a powerful platform for comprehensive knowledge discovery in precision medicine.

2. Related Work

Despite the importance of providing a fact-level overview for a scientific claim – i.e., showing aggregated information describing a scientific fact overall, by combining data coming from several scientific publications and resources – most of the state-of-the-art methods and tools only focuses on providing point-wise sentence-level information – thus not presenting the big-picture understanding. CoreKB aims to fill this gap by providing a comprehensive, literature-supported overview for each scientific fact. In this way, CoreKB differs from the approaches described below.

DEXTER [15], a text mining method for extracting associations from scientific literature, offers a basic search endpoint that requires users to search for gene-cancer pairs and does not support natural language queries. OncoMX [16] provides a unified and user-friendly interface for various datasets, but its literature mining capabilities are limited. OncoSearch [17] enables users to search for sentences that mention gene expression changes in cancer and offers advanced faceted search features. However, it lacks natural language search and does not provide entity cards or infometrics to support evidence. DisGeNET [18] has a well-designed faceted navigation system, but it only returns gene-disease associations at the sentence level, making it difficult to establish reliable associations for gene-disease pairs at the fact level. Finally, BioKB [19] provides access to the semantic content of biomedical literature but does not support natural language search and focuses on exploring and visualizing the structure of the underlying KB.

3. COREKB Search Platform

Figure 1 shows, on the left side, the Knowledge Base Construction (KBC) system used to build the large-scale KB on gene expression-cancer associations, whereas, on the right side, the CoreKB architecture along with its components.

3.1. Knowledge Base Construction

To construct the KB that serves as the foundation for CoreKB, we employed a system that collects text from scientific literature and processes it to obtain sentences. These sentences then undergo Named Entity Recognition and Disambiguation (NERD) annotation, which identifies gene-cancer pairs. The NERD annotations are then subjected to bootstrapping and deployment processes. During bootstrapping, fine-grained relationships between entities are manually annotated and used to (i) train Relation Extraction (RE) methods and (ii) populate the KB. In the deployment process, the trained RE methods are used to automatically extract facts from sentences and populate the KB.

Each fact in the KB is associated with a probability distribution that reflects the likelihood of a specific gene-cancer association. These probabilities are used to perform reliability tests, which determine whether facts are reliable or unreliable based on the quantity and level of consensus of collected evidence. By adopting a probabilistic approach, the system can capture the inherent uncertainty in scientific discourse and help users understand the strength of the evidence supporting a particular gene-cancer association. The KB contains 23,879 genes and

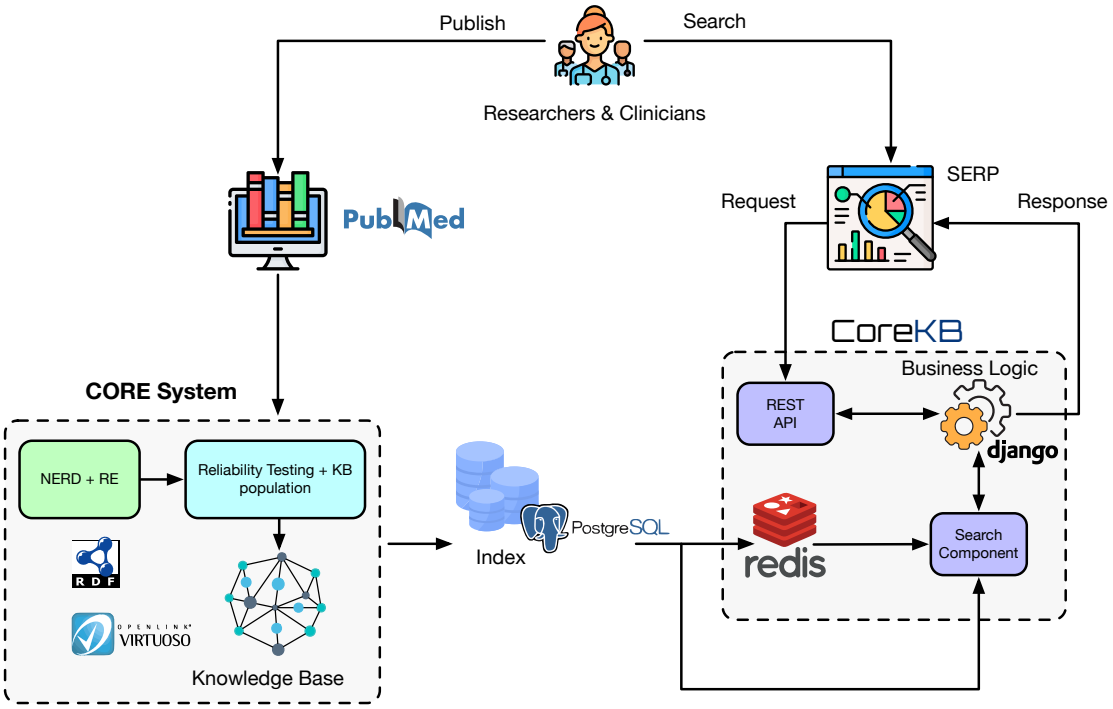

Figure 1: A schematic representation of the KBC system (left side) together with the CoreKB's architecture and its components (right side). The image was originally presented in [1].

11,530 cancer diseases, encompassing a total of 230,000 fine-grained facts. These facts are supported by 1,037,845 sentences taken from 251,038 research articles [14].

3.2. Architecture

CoreKB consists of several components including (i) a React-based web interface for the front-end; (ii) a Django-based business logic for the back-end providing support for REST APIs and services; (iii) a PostgreSQL database aligned with an instance of the Virtuoso RDF triple store for storing the KB data; (iv) a Redis server broker for efficient in-memory data store and access and (v) a Python-based search component that performs NERD on entity mentions in user queries and then performs a structured search in the database to retrieve the matching facts and the related information. To this end, the search component relies on a Redis in-memory dictionary of entities returning for each entity name and synonyms the corresponding unique identifier.

When a user-provided query is entered in the system, the search component assigns a score to each entity that is maximum in case of an exact match or proportional to the number of matching terms in case of a partial match. To avoid favoring long-named entities at the disadvantage of those with short names, the score is discounted proportionally to the entity name's length. If a single entity is recognized, all the related facts are ordered by scientific evidence support. Similarly, if multiple entities are identified, the facts regarding the most matching gene-cancer pairs are promoted in the ranking of results. It's worth noting that CoreKB is not limited to human gene-cancer pairs, instead, it contains facts regarding any living organism. Nevertheless,

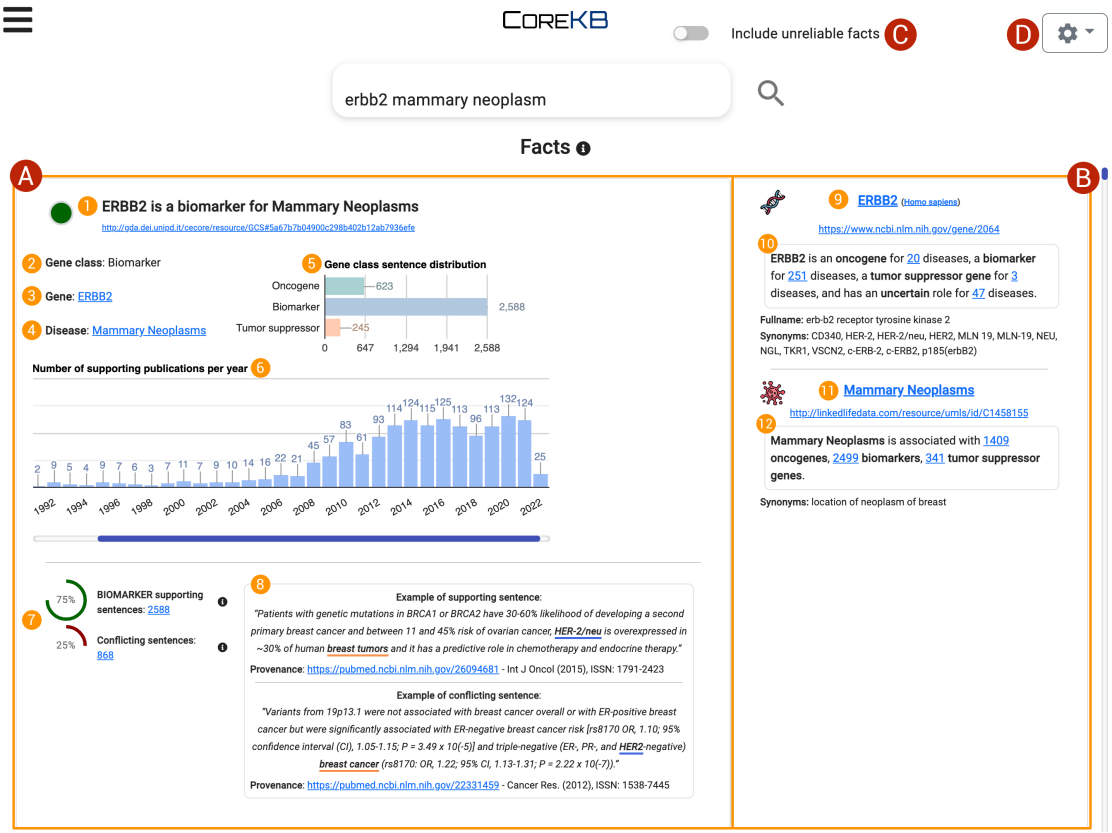

Figure 2: CoreKB Search Engine Results Page (SERP) for the query "*erbb2 mammary neoplasm*". The retrieved facts are organized as cards (A) providing information about the gene (3), cancer (4), and their relationship (2). The fact illustrated in the figure is "*ERBB2 is a biomarker for Mammary Neoplasms*" (1). Regarding this fact, there are 2588 supporting sentences and 868 conflicting sentences (7). The supporting sentences indicate *ERBB2* as a biomarker for *Mammary Neoplasms* whereas the conflicting ones propose two different gene classes: *oncogene* (623 sentences in favor) and *tumor suppressor* (245 sentences in favor). Since the majority of the scientific evidence available propose *ERBB2* as a *biomarker*, the fact is considered reliable. Note that the gene *ERBB2* is a human gene, as indicated between parentheses (9).

human-specific genes are ranked on top.

3.3. User Interaction and Interface features

The CoreKB platform enables users to search for factual information related to gene-cancer associations supported by scientific literature, as well as entities such as genes and cancer diseases. Users can search using free-text or structured search facilities with autocomplete features. The interface allows users to switch between the two search modes via the settings menu and provides clickable sample queries to demonstrate the system's capabilities. The search results are presented in cards that display gene class, cancer label, statistics on supporting and conflicting sentences, key examples of supporting and conflicting sentences, and bibliometrics

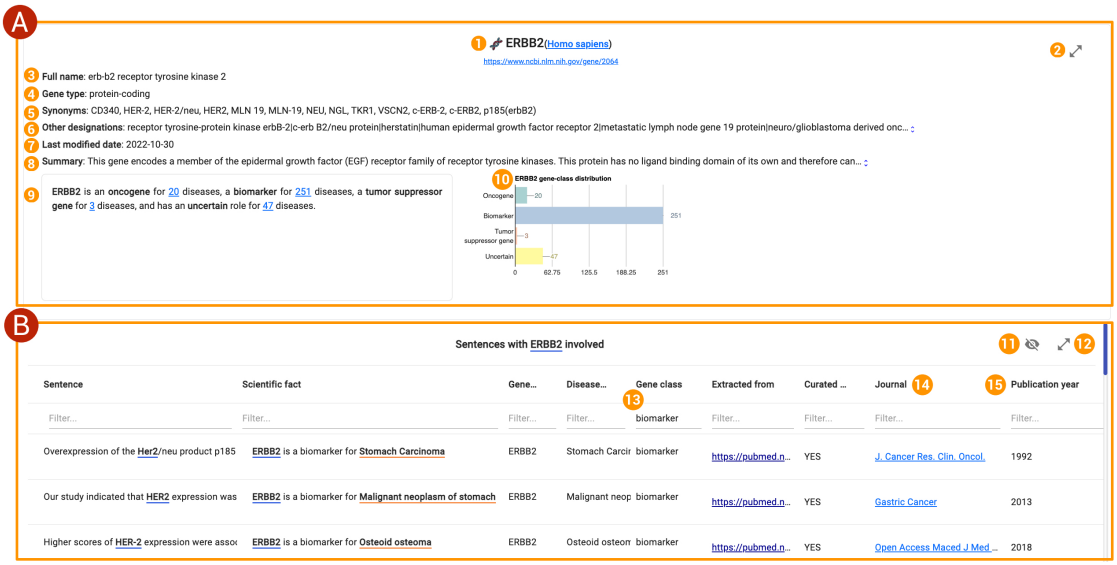

Figure 3: Landing page for the gene *ERBB2*. The interface consists of two cards; the first (A) reports gene-specific information, while the second (B) shows the sentences where *ERBB2* is involved, according to the user-specified column filters and sorting. In the figure, only the sentences where *BRAF* is involved as a biomarker are shown due to the filtering on the *Gene class* column (13). Gene/cancer mentions are highlighted in the sentence text respectively in blue and orange. For each sentence, it is shown information about the publication from which it is extracted, such as the PubMed URL, the journal name (14), and the publication year (15).

on the number of supporting publications per year. The reliability of the fact is indicated by a circle that is colored differently according to the informativeness and reliability of the fact - i.e., green for reliable facts, gray for facts missing proper support, and red for unreliable facts. On the right side of each fact card, specific information about the gene and cancer involved is presented, including links to the corresponding entries in the National Center for Biotechnology Information (NCBI)[2] and Linked Life Data[3] platforms. Users can access additional quantitative and aggregated information on dedicated landing pages, including related facts and supporting and conflicting sentences. The landing page for a particular gene displays all the available information about the gene, including its symbol, full name, type, synonyms, designations, summary, and gene class distribution for different cancer diseases. The sentences involving the gene are presented in a table that users can filter and sort arbitrarily. Action buttons enable users to expand or collapse each card and choose the columns to be shown in the sentence table.

Figure 3 shows the landing page for a given entity, that is, in this case the human gene *ERBB2*. The interface comprises two major cards (A, B). The first card (A) displays comprehensive information about the gene, including its symbol (1), full name (3), type (4), synonyms (5), designations (6), last modified date (7), summary (8), and the gene class distribution – i.e., the role of the gene in a specific gene-cancer association which can assume one of alternatives

[2]https://www.ncbi.nlm.nih.gov/gene/
[3]http://linkedlifedata.com/

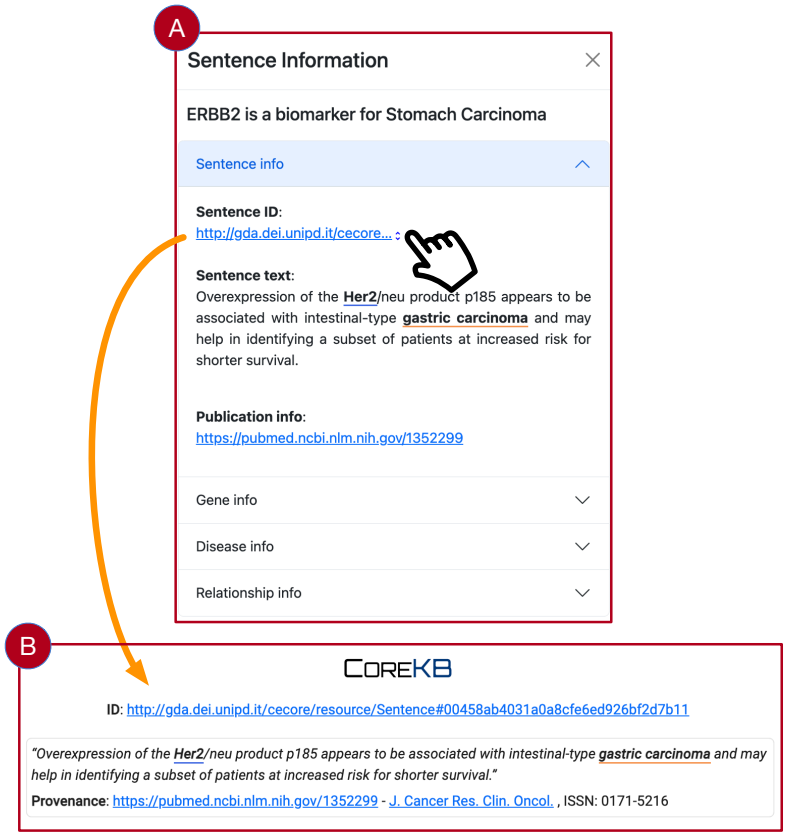

Figure 4: (A) Pop-up reporting the information concerning the sentence clicked in the table in Figure 3.B by the user. (B) Landing page for the sentence corresponding to the permanent link identifier clicked by the user.

oncogene, *biomarker*, *tumor suppressor gene* – for several cancer diseases presented both as a textual statement (9) and depicted also in a horizontal bar chart (10). It is worth noting that all the numbers reported in the textual statement are clickable and allows the users to visualize on a new tab the specific gene-cancer associations of interest (e.g., for which cancer diseases the *ERBB2* gene is indicated to be an oncogene). To save space on the interface, long textual information such as the gene summary is truncated by default, but it can be expanded or collapsed by clicking on the dedicated button after the ellipsis.

The second card (B) provides a list of sentences related to the gene. The sentences are presented in a dynamic table that allows users to filter and sort them according to their preferences. For instance, in Figure 3, only the sentences where *ERBB2* is reported as a biomarker are displayed because of the user-provided value *biomarker* for the *Gene class* column (13). In case of long sentences, users can either (i) resize columns; (ii) going over with the mouse on a sentence to view the full sentence text in a small tool-tip or (iii) click on the sentence to visualize it in a separate pop-up. Moreover, there are two buttons reported on the top-right corner of each card so that users can expand/collapse each card alternatively to fit the full-screen size (2, 12) and choose the columns to be shown in the sentence table (11).

Most of the information reported in the sentence table of Figure 3.B are clickable, including (i) the sentence; (ii) the scientific fact; (iii) the gene; (iv) the disease; (v) the PubMed link to the publication from which the sentence has been extracted and (vi) the publication journal. When the user clicks on a sentence, a pop-up appears to show all the relevant information concerning the sentence such as the sentence identifier, its textual content, the gene, the cancer disease, and their relationship. Figure 4 shows the informative pop-up for the sentence's fact "*ERBB2 is a biomarker for Stomach Carcinoma*", which is reported on top of the window card. Then the pop-up reports other information concerning the sentence, the gene, the cancer disease, and their association in separated collapsible menus. Each sentence and fact in CoreKB is univocally identified with a permanent URL enabling point-wise access to the information specific for an individual resource. Indeed, when the user clicks on the sentence identifier in Figure 4.A the corresponding landing page for the sentence is shown, as depicted in Figure 4.B. Note that in the landing page, users can access provenance information about the sentence, that is, the PubMed URL to the original publication and a link to the corresponding journal entry in the SCImago Portal[4]. Similarly, when users click on a fact the corresponding landing page for the scientific claim is shown providing a holistic perspective enabling to assess overall the literature consensus among the supporting/conflicting evidences.

4. Conclusions

CoreKB is a web-based search platform designed to provide users with reliable scientific facts on gene expression-cancer associations. The platform offers natural language query support as well as structured facet search features integrated with autocomplete facilities. As a distinguishing feature, CoreKB focuses on presenting fact-oriented information rather than adopting a sentence-level approach, where only single, independent results are presented. To this end, CoreKB combines multiple aggregated information from several literature resources either supporting or conflicting with the scientific claim of interest. This approach offers a comprehensive overview of the scientific evidence associated with a given fact. Moreover, CoreKB provides users with a quantitative comparison of the possible gene-cancer associations, making it easier to determine if there is a consensus on a specific gene's role. Although CoreKB focus is on gene expression-cancer associations, its model can be adapted and reused to accommodate other types of relationships with the appropriate modifications.

The platform aims to support clinicians and researchers by providing fast search, access, and consultation of reliable scientific findings along with validating literature evidence. Hence, simplifying knowledge discovery and promoting a serendipity-oriented perspective. As future work, we plan to improve CoreKB according to the requisites and feedback of clinicians, which we plan to collect by conducting a user study.

Acknowledgments

The work was supported by the ExaMode project as part of the EU H2020 program under Grant Agreement no. 825292.

[4] https://www.scimagojr.com

References

[1] F. Giachelle, S. Marchesin, G. Silvello, O. Alonso, Searching for reliable facts over a medical knowledge base, in: Proc. of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023. (in print).

[2] C. Manzoni, D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis, R. Ferrari, Genome, Transcriptome and Proteome: the Rise of Omics Data and Their Integration in Biomedical Sciences, Briefings in Bioinformatics 19 (2016) 286–302.

[3] P. Borry, H. B. Bentzen, I. Budin-Ljøsne, M. C. Cornel, H. C. Howard, O. Feeney, L. Jackson, D. Mascalzoni, Á. Mendes, B. Peterlin, B. Riso, M. Shabani, H. Skirton, S. Sterckx, D. Vears, M. Wjst, H. Felzmann, The Challenges of the Expanded Availability of Genomic Information: an Agenda-Setting Paper, J. Community Genet. 9 (2018) 103–116.

[4] B. Neary, J. Zhou, P. Qiu, Identifying Gene Expression Patterns Associated with Drug-Specific Survival in Cancer Patients, Scientific Reports 11 (2021) 1–12.

[5] S. Dugger, A. Platt, D. Goldstein, Drug development in the era of precision medicine, Nat. Rev. Drug. Discov. 17 (2018) 183–196.

[6] S. Marchesin, F. Giachelle, N. Marini, M. Atzori, S. Boytcheva, G. Buttafuoco, F. Ciompi, G. M. Di Nunzio, F. Fraggetta, O. Irrera, H. Müller, T. Primov, S. Vatrano, G. Silvello, Empowering Digital Pathology Applications through Explainable Knowledge Extraction Tools, Journal of Pathology Informatics 13 (2022) 100139. doi:https://doi.org/10.1016/j.jpi.2022.100139.

[7] X. Li, J. L. Warner, A Review of Precision Oncology Knowledgebases for Determining the Clinical Actionability of Genetic Variants, Front. Cell Dev. Biol. 8 (2020). URL: https://www.frontiersin.org/articles/10.3389/fcell.2020.00048. doi:10.3389/fcell.2020.00048.

[8] S. Marchesin, G. Silvello, TBGA: a large-scale gene-disease association dataset for biomedical relation extraction, BMC Bioinform. 23 (2022) 111.

[9] G. Weikum, X. L. Dong, S. Razniewski, F. M. Suchanek, Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases, Found. Trends Databases 10 (2021) 108–490.

[10] W. Wu, Proactive natural language search engine: tapping into structured data on the web, in: Proc. of the Joint 2013 EDBT/ICDT Conferences, EDBT '13, Genoa, Italy, March 18-22, 2013, ACM, 2013, pp. 143–148. URL: https://doi.org/10.1145/2452376.2452394. doi:10.1145/2452376.2452394.

[11] H. Bast, B. Buchhold, E. Haussmann, Semantic Search on Text and Knowledge Bases, Found. Trends Inf. Retr. 10 (2016) 119–271. URL: https://doi.org/10.1561/1500000032. doi:10.1561/1500000032.

[12] A. Badan, L. Benvegnù, M. Biasetton, G. Bonato, A. Brighente, A. Cenzato, P. Ceron, G. Cogato, S. Marchesin, A. Minetto, L. Pellegrina, A. Purpura, R. Simionato, N. Soleti, M. Tessarotto, A. Tonon, F. Vendramin, N. Ferro, Towards Open-Source Shared Implementations of Keyword-Based Access Systems to Relational Data, in: Proc. of the Workshops of the EDBT/ICDT 2017 Joint Conference (EDBT/ICDT 2017), Venice, Italy, March 21-24, 2017, volume 1810 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-1810/KARS_paper_01.pdf.

[13] M. Agosti, S. Marchesin, G. Silvello, Learning Unsupervised Knowledge-Enhanced Representations to Reduce the Semantic Gap in Information Retrieval, ACM Trans. Inf. Syst. 38 (2020) 38:1–38:48. URL: https://doi.org/10.1145/3417996. doi:10.1145/3417996.

[14] S. Marchesin, L. Menotti, G. Silvello, O. Alonso, CORE: Gene Expression-Cancer Knowledge Base, 2023. URL: https://doi.org/10.5281/zenodo.7577127. doi:10.5281/zenodo.7577127.

[15] S. Gupta, H. Dingerdissen, K. E. Ross, Y. Hu, C. H. Wu, R. Mazumder, K. Vijay-Shanker, DEXTER: disease-expression relation extraction from text, Database J. Biol. Databases Curation 2018 (2018) bay045.

[16] H. M. Dingerdissen, F. Bastian, K. Vijay-Shanker, M. Robinson-Rechavi, A. Bell, N. Gogate, S. Gupta, E. Holmes, R. Kahsay, J. Keeney, H. Kincaid, C. H. King, D. Liu, D. J. Crichton, R. Mazumder, OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data, JCO Clin. Cancer Inform. (2020) 210–220.

[17] H. J. Lee, T. C. Dang, H. Lee, J. C. Park, OncoSearch: cancer gene search engine with literature evidence, Nucleic Acids Res. 42 (2014) 416–421.

[18] J. P. González, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L. I. Furlong, The DisGeNET knowledge platform for disease genomics: 2019 update, Nucleic Acids Res. 48 (2020) D845–D855.

[19] M. Biryukov, V. Grouès, V. P. Satagopam, BioKB - Text mining and semantic technologies for the biomedical content discovery, in: Proc. of the 10th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4LS 2017), Rome, Italy, December 4-7, 2017, volume 2042 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: http://ceur-ws.org/Vol-2042/paper5.pdf.