The Journal of Open Source Software

# A flexible search system for high-accuracy identification of biological entities and molecules

**Max Franz**[1], **Jeffrey V. Wong**[1], **Metin Can Siper**[2], **Christian Dallago**[3, 4, 5], **John Giorgi**[1], **Emek Demir**[2], **Chris Sander**[3, 6, 7], **and Gary D. Bader***[1, 8, 9, 10, 11]

**1** The Donnelly Centre, University of Toronto, Toronto, Ontario, M5S 3E1, Canada **2** Computational Biology Program, Oregon Health and Science University, Portland, OR 97239, USA **3** Department of Cell Biology, Harvard Medical School, Boston, MA, 02215, USA **4** Department of Systems Biology, Harvard Medical School, Boston, MA, 02215, USA **5** Department of Informatics, Technische Universität München, 85748 Garching, Germany **6** Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, 02215, USA **7** Broad Institute of MIT and Harvard, Boston, MA, 02142, USA **8** Department of Computer Science, University of Toronto, Ontario, M5S 2E4, Canada **9** Department of Molecular Genetics, University of Toronto, Ontario, M5S 1A8, Canada **10** The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, M5G 1X5, Canada **11** Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, M5G 2C1, Canada

**DOI:** 10.21105/joss.03756

**Software**

- Review
- Repository
- Archive

**Editor:** Gabriela Alessio Robles

**Reviewers:**

- @AlexanderPico
- @rabdill
- @apcamargo

**Submitted:** 10 September 2021
**Published:** 12 November 2021



## Summary

Identifying subcellular biological entities (genes, gene products, and small molecules) is essential in using and creating bioinformatics analysis tools, text mining, and accessible biological research apps. When research information is uniquely and unambiguously identified, it enables data to be accurately retrieved, cross-referenced, and integrated. In practice, biological entities are identified when they are associated with a matching record from a knowledge base that specialises in collecting and organising information of that type (e.g. genes in NCBI Gene). Our search service increases the efficiency and ease of use for identifying biological entities compared to prior approaches (Berriz & Roth, 2008; Côté et al., 2007; Juty et al., 2012; Reimand et al., 2007). A comparison of similar tools is available in the software documentation(Franz et al., 2021a). This identification service may be used to power research apps and tools, such as Biofactoid, GeneMANIA, and STRING, where colloquial entity names may be provided as input by users (Mering et al., 2003; Mostafavi et al., 2008; Wong et al., 2021).

## Statement of need

Most biologists are unaware of the concept of grounding data to database identifiers (i.e., a value, usually a piece of text, that uniquely identifies the entity) (Abeysooriya et al., 2021). When an author labels a biological entity as 'IL6,' for instance, they may not consider that this label could be ambiguous. Is this IL6 for *Homo sapiens* or for *Mus musculus*? If for *Mus musculus*, there is more than one gene that is called by that name. By mapping the user's entity to a database identifier, in this case, 3569 in NCBI Gene, the data becomes disambiguated.

*Corresponding author

Because many biologists are unaware of the utility of database identifiers, they often have the perception that the common or canonical name of an entity is sufficient for use with analysis tools. These users can be confused by traditional database grounding interfaces, where it is the user's responsibility to select a particular identifier from a long list manually. This sort of grounding interface is incongruent with users' mental model: What purpose could this list of entities have when the entity has already been identified by name?

Our search service can be used to power interfaces that allow the user to identify an entity by name, as per their mental model. This makes grounding accessible to a wider set of researchers, with high ease of use. The service returns its results quickly, i.e., in less than 100 milliseconds on a 2.4 GHz dual-core processor with 8 GB of RAM, so that the result may be shown to a user interactively and without impeding the user's actions. As output, the service returns a ranked list of possible groundings in descending order of relevance. The first entry in this list is the predicted identifier with the highest confidence given to the user's input. The remainder of the list exists only to allow the user to recover from an incorrect first identifier. The service is customisable to accommodate various use cases, e.g., interactive grounding interfaces that can dynamically build up heuristics during a user's session with the system.

## Performance evaluation

In the rare case where the service does not return the correct result as the first result, the app which uses the service may present one or more of the following entries in the list. For instance, the app may allow the user to indicate a correction manually — e.g., "I meant this IL6, not that one." To facilitate this, each identifier in the results list has included a corresponding set of descriptions and metadata.

To verify the accuracy of the service's results, a test suite was created. The tests include entity names used in PubMed Central, Pathway Commons, and Biofactoid projects. Primarily, the test cases were prioritised based on popularity to measure how well commonly-researched entities are correctly assigned by the service. In all, there are currently over 750 test cases. Of those test cases, 91% returned the expected result as the first entry in the returned list in our tests. In nearly 98% of cases, the expected result was within the top ten entries in the returned list.

## Mechanism

The database is built using a dynamic indexing approach with Elasticsearch. The system includes facilities to download the latest set of identifiers and associated metadata from NCBI Gene (including DNA, genes, and proteins in *H. sapiens*, *M. musculus*, *S. cerevisiae*, *D. melanogaster*, *E. coli*, *C. elegans*, *A. thaliana*, *R. norvegicus*, *D. rerio*, *SARS-CoV-2*, and extensible to other organisms), ChEBI (including small molecules, e.g., drugs), and UniProt (including proteins for cross-referencing NCBI Gene identifiers) to build the database. This indexing process is exposed as a top-level command, integrated with the search server itself, so that data sources can automatically be kept fresh regularly. The indices of the grounding service can be exported to an external repository (e.g. Zenodo) and directly imported in order to provide a means of referencing a particular version of an index and to reduce the need to manually build an index for researchers that reuse the grounding service in their own projects.

The grounding service operates by making queries on the database, with the query string normalised – i.e., the string has punctuation removed, the case is normalised (e.g., "TNF-a" is considered the same as "tnf alpha"). A fuzzy query and a precise query are made to the database to ensure that both exact matches and near matches are included in the initial results. These initial results are processed with a multithreaded ranking approach. The Sørensen–Dice

coefficient is used to rank the initial results based on each entity's official name and synonyms, considering each entity's best-case score.

The Sørensen–Dice coefficient, $s_i$, for the $i$th synonym, is given as follows, where: * $b_t$ is the total number of bigrams in both the $i$th synonym and the query string, * $b_i$ is the total number of bigrams in the $i$th synonym, and * $b_q$ is the total number of bigrams in the query string:

$$s_i = (\frac{2b_t}{b_i + b_q})$$

For tie-breaking, a series of further measures are used to order the results. These measures include an organism ranking preference, molecular charge (with a preference for charged, aqueous molecules), and the Sørensen–Dice coefficient of the official name. Finally, the data source (namespace) filter may be applied, if specified, in order to include results only of a particular type (e.g., small molecules from ChEBI).

## Usage

With a complete index, the grounding service's server can be started. The server exposes a REST-like API to enable client applications to query the index (Franz et al., 2021b). The main endpoint in the API is search. A search contains a number of input parameters, the user's typed entity name chief among them. Other parameters are optional. They may be implicitly specified by the user, or they may be specified by the client application. The target namespace is an optional parameter that filters the search result by a particular data source (e.g., NCBI Gene). When unspecified, no namespace filtering is applied. A second optional parameter is an organism ranking, which can inform the service of the relative likelihood of a search's pertinence to a particular organism. An organism ranking based on general popularity, measured by PubMed mentions, is used by default (Maglott et al., 2010).

## Discussion

An existing app, Biofactoid, provides an example of how this grounding service empowers novice users to ground entities to database identifiers automatically. Biofactoid users typically take one to five minutes to summarise the molecular interactions with a paper. The database identifiers in author's Biofactoid documents have been accurate thus far, as evaluated by extensive manual spot-checking, and these novice users require no training in or understanding of database identifiers in order to create these documents.

Apps that previously required a user to explicitly specify an organism may instead use heuristics paired with the grounding service's API to provide intelligent results without user intervention. On the other hand, organism-specific apps may leverage the organism ordering to provide results only for the relevant organism: The first instance of a non-specified organism in one of the returned entries indicates that there are no further grounding entries for the specified organism.

Further, more natural search systems may emerge as a result of the approaches of this grounding service. A biological data search engine may allow for natural language queries, powered by the grounding service. A user may type a search, such as "interactions of tnf in human," in order to get relevant results. If the user types an ambiguous query, such as "tnf," the search engine may use the ranked results to provide intelligent follow-up questions: "Did you mean tnf (mouse)?" These results and follow-up questions may go so far as to be user-personalised.

A user that predominantly searches for mouse genes may have the search engine provide custom-tailored results, based on a search history used to inform the grounding service.

Modern research apps and tools motivate the need for robust, fast, reusable grounding tools that allow for easily identifying biological entities from their common names. Our grounding service can be used in apps to provide users with an easy-to-use experience in line with their mental model of biological entities.

## Acknowledgements

This project was funded by the US National Institutes of Health (NIH) [U41 HG006623, U41 HG003751, R01 HG009979 and P41 GM103504].

## References

Abeysooriya, M., Soria, M., Kasu, M. S., & Ziemann, M. (2021). Gene name errors: Lessons not learned. *bioRxiv*. https://doi.org/10.1101/2021.03.30.437702

Berriz, G. F., & Roth, F. P. (2008). The synergizer service for translating gene, protein and other biological identifiers. *Bioinformatics*, *24*(19), 2272–2273. https://doi.org/10.1093/bioinformatics/btn424

Côté, R. G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., & Hermjakob, H. (2007). The protein identifier cross-referencing (PICR) service: Reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, *8*(1), 1–14. https://doi.org/10.1186/1471-2105-8-401

Franz, M., Wong, J. V., & Siper, M. C. (2021a). GitHub: Pathway commons grounding search service. In *GitHub repository*. GitHub. https://github.com/PathwayCommons/grounding-search

Franz, M., Wong, J. V., & Siper, M. C. (2021b). Pathway commons grounding search service. In *GitHub documentation*. GitHub. https://grounding.baderlab.org

Juty, N., Le Novere, N., & Laibe, C. (2012). Identifiers. Org and MIRIAM registry: Community resources to provide persistent identification. *Nucleic Acids Research*, *40*(D1), D580–D586. https://doi.org/10.1093/nar/gkr1097

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2010). Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research*, *39*(suppl_1), D52–D57. https://doi.org/10.1093/nar/gki031

Mering, C. von, Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. (2003). STRING: A database of predicted functional associations between proteins. *Nucleic Acids Research*, *31*(1), 258–261. https://doi.org/10.1093/nar/gkg034

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., & Morris, Q. (2008). GeneMANIA: A real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, *9*(1), 1–15. https://doi.org/10.1186/gb-2008-9-s1-s4

Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). G: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, *35*(suppl_2), W193–W200. https://doi.org/10.1093/nar/gkm226

Wong, J. V., Franz, M., Siper, M. C., Fong, D., Durupinar, F., Dallago, C., Luna, A., Giorgi, J. M., Rodchenkov, I., Babur, Ö., & others. (2021). Capturing scientific knowledge in computable form. *bioRxiv*. https://doi.org/10.1101/2021.03.10.382333