

BioGraph: Data framework for linking and querying diverse biological metadata

Veljković A.N.*, Mitić N.S.

University of Belgrade, Faculty of Mathematics, Belgrade, Serbia

* aleksandar.veljkovic@matf.bg.ac.rs

Associating biological data from different sources provides a holistic view of a domain and enables finding patterns in data that are otherwise difficult or impossible to observe by only analyzing isolated biological entities. The key issues for creating connections between data objects are the variety of biological data formats, data organization schemas, and data access methods. Connecting data from different databases can be challenging, as an entity from one database may not have the same properties or identifiers as the same entity described in another database. While some databases contain a variety of entity identifiers from various databases, the search is limited to exact property matching, and complex queries using multiple metadata attributes are not possible.

To overcome these issues, we introduce BioGraph – a novel data framework that enables linking and retrieving information from heterogeneous interconnected biological data. We tested the model and generated a knowledge graph using metadata from five distinct public datasets, DisProt [1], HGNC [2], TANTIGEN 2.0 [3], IEDB [4] and DisGeNET [5]. The resulting graph interconnects more than 17 million nodes, of which 2.5 million individual biological entity objects with over 4 million relationships. The software system allows searching and retrieving patterns and retrieving matching results from the knowledge graph using a user-friendly interface.

Numerous efforts are underway to join data from various datasets into a knowledge graph. However, the solutions currently available tend to concentrate on specific subdomains, such as proteomics and drug discovery, rather than on the comprehensive integration of biological data across different domains. We propose a novel model that facilitates simultaneous querying of biological data attributes from multiple datasets by leveraging the metadata available in the original databases. Our model aims not to replicate the data from the original datasets but to establish links between the metadata that can be utilized to efficiently execute complex queries on interconnected data. This model allows for the incorporation of properties into entities and relations and the integration of metadata from diverse data formats. To complement our model, we have developed a tool and a web interface which enables creating graphical queries in a user-friendly manner as well as exporting the query results into downloadable files. The tool and corresponding packages can be deployed locally as a standalone system, enabling offline execution of queries.

References

1. Vucetic S., Obradovic Z., Vacic V., Radivojac P., Peng K., Iakoucheva L.M., Cortese M.S., Lawson J.D., Brown C.J., Sikes J.G. et al. DisProt: a database of protein disorder. *Bioinformatics*. 2004;21:137-140.4.
2. Seal R.L., Braschi B., Gray K., Jones T.E., Tweedie S., Haim-Vilmovsky L., Bruford E.A. Genenames. org: the HGNC resources in 2023. *Nucleic Acids Research*. 2023;51:D1003-D1009.
3. Zhang G., Chitkushev L., Olsen L.R., Keskin D.B., Brusica V. TANTIGEN 2.0: a knowledge base of tumor T cell antigens and epitopes. *BMC Bioinformatics*. 2021;22:1-8.
4. IEDB. Retrieved from <http://www.iedb.org>. Accessed: 2023-02-06.
5. Piñero J., Juan Manuel R.A., Josep Saüch-Pitarch F.R., Emilio Centeno F.S., Furlong L.I. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucl. Acids Res*. 2019. Doi 10.1093/nar/gkz1021.

BioGraph: 鏈接和查詢不同生物元數據的數據框架

Veljković A.N.*, Mitić N.S.

¹ 貝爾格萊德大學數學學院，貝爾格萊德，塞爾維亞

* aleksandar.veljkovic@matf.bg.ac.rs

關聯來自不同來源的生物數據提供了一個領域的整體視圖，並能夠在數據中找到模式，而這些模式很難或不可能通過僅分析孤立的生物實體來觀察。

在數據對象之間創建連接的關鍵問題是生物數據格式、數據組織模式和數據訪問方法的多樣性。

連接來自不同數據庫的數據可能具有挑戰性，因為來自一個數據庫的實體可能與另一個數據庫中描述的同一實體不具有相同的屬性或標識符。

雖然一些數據庫包含來自各種數據庫的各種實體標識符，但搜索僅限於精確的屬性匹配，並且無法使用多個元數據屬性進行複雜查詢。

為了克服這些問題，我們引入了 BioGraph 一種新穎的數據框架，可以從異構互連的生物數據中鏈接和檢索信息。我們測試了該模型並使用來自五個不同公共數據集 DisProt [1]、HGNC [2]、Tantigen 2.0 [3]、IEDB [4] 和 DisGeNET [5] 的元數據生成了知識圖。生成的圖形將超過 1700 萬個節點互連起來，其中 250 萬個單獨的生物實體對象具有超過 400 萬個關係。該軟件系統允許搜索和檢索模式，並使用用戶友好的界面從知識圖中檢索匹配結果。

文學

1. Vucetic S., Obradovic Z., Vacic V., Radivojac P., Peng K., Iakoucheva L.M., Cortese M.S., Lawson J.D., Brown C.J., Sikes J.G. et al. DisProt: a database of protein disorder. *Bioinformatics*. 2004;21:137-140.4.
2. Seal R.L., Braschi B., Gray K., Jones T.E., Tweedie S., Haim-Vilmovsky L., Bruford E.A. Genenames. org: the HGNC resources in 2023. *Nucleic Acids Research*. 2023;51:D1003-D1009.
3. Zhang G., Chitkushev L., Olsen L.R., Keskin D.B., Brusic V. TANTIGEN 2.0: a knowledge base of tumor T cell antigens and epitopes. *BMC Bioinformatics*. 2021;22:1-8.
4. IEDB. Retrieved from <http://www.iedb.org>. Accessed: 2023-02-06.
5. Piñero J., Juan Manuel R.A., Josep Saüch-Pitarch F.R., Emilio Centeno F.S., Furlong L.I. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucl. Acids Res*. 2019. Doi 10.1093/nar/gkz1021.