Educational-scientific Council
Faculty of Mathematics
University of Belgrade

By the decision of the Educational-scientific Council of the Faculty of Mathematics in Belgrade, adopted at the 405th session held on June 23 2023, we were appointed as members of the commission for the evaluation of the doctoral dissertation "***Semantic Unification And Searching Of Bioinformatics Databases Using Data Mining Methods***" by candidate Aleksandar Veljković, Master of Informatics. After reviewing the submitted manuscript, we submit the following

Report

Candidate's biography

Candidate Aleksandar Veljković was born on September 23, 1992 in Požarevac. He graduated from elementary school in Golubac, as a holder of diploma Vuk Karadžić, and from secondary school in Veliko Gradište, as a valedictorian. He started his bachelor‘s academic studies in 2011 at the University of Belgrade, Faculty of Mathematics, program Informatics, and graduated in 2014 with a GPA of 9.73.

He started his master’s academic studies in 2014, at the Faculty of Mathematics, program Informatics, with major in computer science, and graduated in 2016 with GPA of 9.69, with master’s thesis "New method for genome assembly based on PFG electrophoresis", under the supervision of prof. Dr. Nenad Mitić.

He started his Ph.D. studies at the Faculty of Mathematics, program Informatics, in 2016 and passed all exams with an average grade of 10.00.

From 2015 to 2023, he was a teaching assistant at the Faculty of Mathematics, University of Belgrade, at the Department of Computer Science and Informatics. He participated in exercising in several subjects at undergraduate and master studies (Programming 1, Programming 2, Algorithms and data structures, Web programming, Artificial intelligence, Data research 2, Bioinformatics, Data research in bioinformatics, Cryptography, Project Management in Industry and Science and Fundamentals of Programming for Biochemists).

Since 2023, he is working as a senior researcher in the field of cryptography in the research department of the MVP Workshop company.

He participated in the development of several technical solutions in the field of decentralized systems, cryptography and flight control and management of unmanned aerial vehicles. His main areas of interest are cryptography, data research and bioinformatics.

Published scientific papers and announcements from scientific meetings

Aleksandar Veljković has published seven papers in journals from the SCI list (M21, M22, M21a, M23, M23, M23, M23), one paper in a national journal, and gave ten presentations at international conferences (seven printed in excerpts) and one presentation at a domestic conference.

Papers in international journals from the SCI list:

1. Veljković, A. et al.: *BioGraph: Data Model for Linking and Querying Diverse Biological Metadata*, International Journal of Molecular Sciences, (2023) 24(8):6954. https://doi.org/10.3390/ijms24086954 (M21, IF=5.6)

2. Milutinović, B. et al.: *VLSI for SuperComputing: Creativity in R+D from applications and algorithms to masks and chips*, Advances in Computers, (2022). https://doi.org/10.1016/bs.adcom.2022.01.001 (M22, IF=3.067)

3. Čokić, V.P. et al.: *A comprehensive mutation study in wide deep-rooted R1b Serbian pedigree: mutation rates and male relative differentiation capacity of 36 Y-STR markers,* Forensic Science International: Genetics, Volume 41 (2019), 137-144 https://doi.org/10.1016/j.fsigen.2019.04.007 (M21a, IF=4.884)

4. Alempijević, T.M. et al.: *Change in the incidence and anatomic distribution of colorectal adenoma and cancer over a period of 20 years – A single center experience,* Vojnosanitetski pregled, Volume 75, Issue 3 (2018), 260-266 https://doi.org/10.2298/VSP160409207A (M23, IF=0.418)

5. Alempijevic T. et al.: *Doppler ultrasonography combined with transient elastography improves the non-invasive assessment of fibrosis in patients with chronic liver diseases,* Medical Ultrasonography, Volume 19 (2017), 7-15 http://dx.doi.org/10.11152/mu-921 (M23, IF=1.651)

6. Arsenijević V. e al.: *Erythropoietin in the Evaluation of Treatment Outcomes in Patients with Polytrauma,* Acta clinica Croatica, Vol. 56. No. 4 (2017), 581-587 https://doi.org/10.20471/acc.2017.56.04.01 (M23, IF=0.452)

7. Alempijevic T. et al.: *Erythropoietin in predicting prognosis in patients with acute-on-chronic liver failure,* Journal of Gastrointestinal and Liver Diseases, Voluma 25, Number 4, (2016), 473-479 https://doi.org/10.15403/jgld.2014.1121.254.jev (M23, IF=4.884)

Papers in national journals:

8. Veljković A., *Algorithm for Document Authorship Identification and Plagiarism Evaluation Based on Generalized Suffix Tree*, Pregled NCD 37 (2020), 46–51 http://www.ncd.matf.bg.ac.rs/issues/37/5.pdf

Announcements at international conferences printed in full or in excerpts:

9. Veljković A, Stojanović B, Malkov S, Beljanski M, Pavlović-Lažetić G, Mitić M. *Codon Usage-based SARS-CoV-2 Protein Classification.* Book of Abstracts Belgrade BioInformatics Conference 2021. Biologia serbica. 2021;43(1) ISSN: 2334-6590, 21-25. Jun, 2021, Belgrade. http://belbi.bg.ac.rs/wp-content/uploads/2021/06/Book_of_Abstracts_2021-1.pdf

10. Stojanović B, Veljković A, Malkov A, Beljanski M, Pavlović-Lažetić G, Mitić N. *Codon Usage Polymorphism in SARS-CoV-2 Protein Coding Sequences.* Book of Abstracts Belgrade BioInformatics Conference 2021. Biologia serbica. 2021;43(1) ISSN: 2334-6590, 21-25. Jun, 2021, Belgrade. http://belbi.bg.ac.rs/wp-content/uploads/2021/06/Book_of_Abstracts_2021-1.pdf

11. Ćirić N, Veljković A. *Identification of Differentially Expressed Genes in SARS-Cov-2 Infected Cells Using Bayesian Network Models.* Book of Abstracts Belgrade BioInformatics Conference 2021. Biologia serbica. 2021;43(1) ISSN: 2334-6590, 21-25. Jun, 2021, Belgrade. http://belbi.bg.ac.rs/wp-content/uploads/2021/06/Book_of_Abstracts_2021-1.pdf

12. Veljković A, et al. *Classification of Single Cell Types using Small Sets of Expressed Genes: Comparative Analysis of Supervised Machine Learning Methods.* 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), DOI: 10.1109/bibm52615.2021.9669844

13. Veljković A, Mitić N, *Semantic unification and search of bioinformatics databases.* Book of Abstracts Belgrade BioInformatics Conference 2023. (2023) ISBN: 978-86-82679-14-1, 19-23. Jun, 2023, Belgrade.

https://belbi.bg.ac.rs/wp-content/uploads/2023/07/BelBi2023-Book-of-Abstracts.pdf

14. Malkov, S. et al. *Clustering and classification of SARS-CoV-2 isolates using RSCU.* Book of Abstracts Belgrade BioInformatics Conference 2023. (2023) ISBN: 978-86-82679-14-1, 19-23. Jun, 2023, Belgrade. https://belbi.bg.ac.rs/wp-content/uploads/2023/07/BelBi2023-Book-of-Abstracts.pdf

15. Veljković, A, et al. *Analysis of nucleotide sequence repeats in coronaviruses.* Book of Abstracts Belgrade BioInformatics Conference 2023. (2023) ISBN: 978-86-82679-14-1, 19-23. Jun, 2023, Belgrade. https://belbi.bg.ac.rs/wp-content/uploads/2023/07/BelBi2023-Book-of-Abstracts.pdf

Announcements at international conferences:

16. Veljković A. *Zero Knowledge Machine Learning.* Data Science Conference DSC Europe 24, 2023.

17. Veljković A, et al. *BioGraph: Data Model for Linking and Querying Diverse Biological Metadata.* World congress "Systems theory, algebraic biology, artificial intelligence: mathematical foundations and applications", 2023.

18. Veljković A, et al. *BioGraph: Data Framework for Linking and Querying Diverse Biological Metadata*. VII Congress of Russian Biophysicists, 2023.

Announcements at domestic conferences printed in extract:

19. Veljković A. *Digitization of church records for creating knowledge database of personal historical records.* NCD XIX Conference 2022.

The subject of the dissertation

The subject of the doctoral dissertation is the development of methods for the recognition and unification of semantic links between metadata from different biological databases and the application of data mining methods to enable and realize the search and analysis of semantic links.

Biological data are represented in a large number of open and closed data collections, which use a number of different data formats. Each biological data includes both the data itself and a set of metadata that describe the properties of a specific biological concept or object. At the same time, different databases contain different subsets of metadata that refer to the same biological objects. The method of accessing and searching this data is specific to each of the databases in which it is located, which makes it even more difficult to connect and analyze the databases.

Identifiers of the same objects often differ between databases, so in some cases semantic search is not only appropriate but also the only solution for searching. By finding and unifying semantic links between (meta)data from different sources with potentially different structure, and applying data mining methods to such established links, their aggregation into a unified metadata structure supporting semantic search could be ensured.

In his work, candidate Aleksandar Veljković proposes a new model for connecting bioinformatics databases, which also includes a proposed automated protocol for discovering new semantic relationships between objects. A new automated protocol for discovering new relations of semantic similarity between objects of different databases is proposed, which is based on data mining methods. As a validation of the proposed models and protocols, the candidate implemented the *BioGraph* application based on that model, using the unified graph database which is connected to five bioinformatics databases.

Dissertation review

The manuscript has 129 (113 + XVI) pages, it was written in English and has the following structure:

1. *Introduction*
2. *Cluster analysis and association rules mining*
3. *Searching bioinformatics databases*
4. *New Data Joining Model Proposal*
5. *Model implementation and validation*
6. *Results and discussion*
7. *Conclusion*

The title page, summary and information about the mentor and committee members are written in English and Serbian. The work also includes Contents, List of Figures, List of Tables, Bibliography with 99 bibliographic units, an Appendix A with source code of database importer scripts and Candidate's Biography.

The introductory chapter presents the context of the problem and its importance. The need for automating the linking of the content of various bioinformatics databases was highlighted. Their large number and variety of formats and contents make this problem both more difficult and more significant. The concept of measuring the distance of objects, i.e. their similarity, was introduced, first by individual simple attributes, and then at the level of more complex objects. The concept of semantic similarity of objects, based on the comparison of their metadata, is presented. Known results in the field, as well as related topics, are also presented.

The second chapter introduces clustering and association rule exploration algorithms, which can be used to recognize certain patterns in data and which can be helpful in deriving and describing semantic similarity relationships between data.

Chapter 3 includes an overview of some of the most representative bioinformatics databases, including *UniProt*, *NCBI Gene*, *EMBL* and *Enssembl*. In particular, the chapter deals with the problem of searching bioinformatics data and their heterogeneous data schemas and different ways of representing and modeling certain entities. The difficulties in trying to connect them are emphasized, as well as the different data formats and protocols used to represent and describe objects in these databases, such as *CSV*, *JSON*, *RDF*, *TSV*, *PDB*, *FASTA*, *FASTQ* and others. Different types of objects that appear in those databases are presented, as well as different ways of identifying individual objects.

Chapter 4 presents a proposal of a new model for connecting bioinformatics databases. The model is designed so that it can serve as a basis for unifying bioinformatics data from multiple sources and for developing a system for deriving new relations of semantic similarity between objects of different databases.

The main goal of the new model is to enable semantic search based on data from different databases, whereby available metadata is primarily used to link them. This approach does not require to collect all data in one place and allows to create relatively smaller networks for indexing, and also provides a wealth of different information that can be found in metadata. An additional motive is their size, because biomedical data can be very large, so although for some of the data the volume of metadata can be comparable to or even larger than the data volume, in most cases the metadata is much smaller. The new model is called *BioGraph*. The *BioGraph* model is designed so that DBMSs based on different models can be used for its implementation (first of all, it is important that a relational model can be used), but in further work (in chapter 5) a graph database was used.

The basis of the model consists of three types of objects (entities, identifiers and data) and the relationships that connect these objects. Entities can have different types, which depend on the domain of specific databases, for example, they can be proteins, genes, diseases, antigens, epitopes, regions and others. In the ideal case, one object (with potentially more identifiers) is created for an entity that appears in different databases, but often this is not possible, so several different objects are created, among which appropriate relationships are later established (among which the relationship "*IS_EQUAL_TO*"). Identifier objects are linked to entities via relationship "*HAS_ID*". Additional metadata about entities is stored in data objects, to which they are linked by the "*HAS_DATA*" relationship. One entity can be associated with multiple data, just as one data can be associated with multiple entities. Data can represent different types of information about entities, from their names, through specific structural or general characteristics, to descriptions of origins and techniques for extracting entities.

Different types of relationships can also exist between entities, the establishment of which significantly improves search capabilities. Within the description of the model, it is roughly presented how the model could be implemented with a graph or relational database.

An integral part of the model is the method for deriving semantic relations between objects. Relationships between objects are established based on the analysis of the existing relationships of those objects with other objects and on the basis of similarities and differences between their metadata. The general basis of the procedure is the application of data mining algorithms on the existing relationships (matrix of selected types of relationships and related entities) and the selection, creation and addition of new relationships to the collection (knowledge graph) based on the results of the application of those algorithms. The choice of specific sets and algorithms depends on the goal. In the description of the model, it is presented how clustering, nearest neighbor graph, association rules and latent semantic analysis can be used.

A basic set of relationship types (*IS_EQUAL*, *IS_INSTANCE*, *IS_VARIANT*, *FROM*, *CONTAINS*, *HAS_ROLE*, *RELATED_WITH*, *SIMILAR_TO*) is presented within the model, but that set can be expanded as needed and according to the domains of the specific databases being searched.

Chapter 5 describes the implementation details of the proposed model and system for deriving new semantic similarity relations based on available data. The implementation includes five databases (*DisProt*, *DisGeNET*, *Tantigen*, *IEDB* and *HGNC*). The system's core implementation provides a *REST-API* for searching. Queries are expressed in an internal query language specifically developed for these needs. The user interface is implemented in the form of a web interface to the core of the system. Everything was developed in the *JavaScript* programming language and using the *NodeJS* environment (server).

The system architecture consists of basic services, data importers, indexers, database adapters and a *REST* interface:

- Basic services represent the central component of the *BioGraph* system. They allow other components to add, link, or search data. They use transactional mechanisms to maintain data, thus maintaining the integrity of the system.

- Data importers represent the entry point of the system. At least one data importer is implemented for each database that is connected. The importer downloads the data from the database (by downloading prepared packages or searching the database website or in some other way), transforms them into the appropriate form and populates the *BioGraph* database with the transformed data using basic services.

- Indexers are services for storing and indexing all data in the system. In particular, data on data import, entities, identifiers and descriptions (metadata) are indexed.

- Database adapters allow the system to be implemented using different DBMSs. They represent the basis of system portability and flexibility. The *BioGraph* implementation uses the *Neo4J* graph base and the appropriate adapter. An adapter for *MySQL* was also developed for experiments and the validation of the concepts.

- The *BioGraph* system *API* was developed using *REST*.

An internal query language, modeled on *Cypher*, was also designed to be used as a query language for *BioGraph*. Queries are written in *JSON* format.

Chapter 6 discusses the characteristics of the proposed model and specific *BioGraph* implementation. The model is compared with other similar existing models and the differences and similarities between them are highlighted. In terms of the most important criteria, the proposed model has significant advantages, which is especially evident when considering the flexibility and extensibility of the model. Although the model is basically independent of the data model used (relational or some other), the implementation for a different model can be significantly different.

Scientific contribution of the dissertation

In his dissertation, candidate Aleksandar Veljković presents a proposal for a completely new model for linking different bioinformatics databases, as well as a description of an implementation of that model on a very specific and significant example.

The key scientific contributions of the dissertation are (1) the proposed new model for connecting bioinformatics databases, which also includes a proposed automated protocol for discovering new semantic relationships between objects, (2) the proposed automated protocol for discovering new semantic similarity relationships between objects of different databases, based on on data mining methods, and (3) publicly available implementation of the proposed model (https://github.com/aleksandar-veljkovic/biograph) on the example of connecting five bioinformatics databases with the use of a graph database, which, in addition to practical applicability, is also a confirmation of the usability of the model and its good foundation.

Conclusion

The results of the candidate’s research in the field of semantic unification of bioinformatics databases, which are presented in the manuscript "***Semantic Unification and Searching of Bioinformatics Databases Using Data Mining Methods***") represent a valuable scientific contribution in the fields of bioinformatics, databases and applications of data mining in bioinformatics.

The topic and subject of the research are from the narrower scientific field of Bioinformatics, and the research methods are from the narrower scientific field of Data Mining. Some of the research results may have a wider application in the field of Databases. Candidate Aleksandar Veljković demonstrated excellent knowledge of these fields and ability to perform independent scientific work. The obtained results have significant potential for further application and continued research.

Considering all of the above, we propose to the Educational-scientific Council of the Faculty of Mathematics to accept the manuscript "***Semantic Unification and Searching of Bioinformatics Databases Using Data Mining Methods***" of candidate **Aleksandar Veljković** as a doctoral dissertation and assign a committee for its defense.

In Belgrade, 25.12.2023.

Members of the evaluation committee

____

(dr Saša Malkov, associate professor)

____

(dr Jovana Kovačević, assistant professor)

____

(dr Aleksandar Kartelj, associate professor)

____

(dr Natalija Polović, full professor
University of Belgrade, Faculty of Chemistry)

____

(dr Yuriy Orlov,
PhD, DrSci, Professor of the Russian Academy of Sciences,
The Digital Health Institute, I.M. Sechenov First Moscow State Medical University
of the Ministry of Health of the Russian Federation, Moscow, Russia)