Integrating Heterogeneous Gene Expression Data through Knowledge Graphs for Improving Diabetes Prediction

Rita T. Sousa[1,*], Heiko Paulheim[1]

[1]*Data and Web Science Group, Universität Mannheim, Germany*

Abstract

Diabetes is a worldwide health issue affecting millions of people. Machine learning methods have shown promising results in improving diabetes prediction, particularly through the analysis of diverse data types, namely gene expression data. While gene expression data can provide valuable insights, challenges arise from the fact that the sample sizes in expression datasets are usually limited, and the data from different datasets with different gene expressions cannot be easily combined.

This work proposes a novel approach to address these challenges by integrating multiple gene expression datasets and domain-specific knowledge using knowledge graphs, a unique tool for biomedical data integration. KG embedding methods are then employed to generate vector representations, serving as inputs for a classifier. Experiments demonstrated the efficacy of our approach, revealing improvements in diabetes prediction when integrating multiple gene expression datasets and domain-specific knowledge about protein functions and interactions.

Keywords

Diabetes Prediction, Expression data, Knowledge Graph, Ontology, Knowledge Graph Embedding

1. Motivation

Diabetes is a chronic health condition resulting from insufficient insulin production by the pancreas or the body's inability to utilize the insulin it generates effectively [1]. This disease has emerged as a worldwide health issue, impacting millions of people globally. According to the World Health Organization, in 2019, diabetes directly contributed to 1.5 million deaths, with 48% occurring before the age of 70. Besides that, this chronic disease is associated with the development of several comorbidities, such as blindness, kidney failure, heart attacks, strokes, and lower limb amputation.

Due to the multidisciplinary nature of diabetes, predicting and detecting this complex disease continues to pose a significant challenge. In the last decades, some approaches have demonstrated encouraging outcomes using machine learning methods to identify patterns and potential risk factors linked to diabetes, allowing not only the early detection of diabetes but also enabling tailored interventions [2, 3, 4, 5]. These machine learning approaches encompass several types of data, including electronic health records [6], imaging data [7], and demographic data [8]. Omics data, namely gene expression datasets, have also received attention since

genomics, epigenomics, and transcriptomics can help understand the critical pathways and regulatory mechanisms in diabetes [9].

While gene expression datasets are readily accessible in public databases, and gene expression analysis is a powerful tool for pinpointing genes associated with diseases, particularly in the context of diabetes prediction, a significant issue arises in handling this type of data. On the one hand, gene expression datasets often exhibit a limitation in sample size, with a relatively small number of included samples. Conversely, supervised machine learning methods are data-driven, relying on a large number of labeled data for effective training and performance. One alternative involves combining multiple expression datasets to increase the sample pool for training machine learning models. However, this brings us to the challenge of how to integrate the information about multiple expression datasets, as each dataset may measure gene expression across distinct genes. Additionally, variations in experimental platforms and designs across different studies further complicate integration efforts. Knowledge graphs (KGs) present a unique and promising solution. KGs can represent knowledge about concepts and relationships in a fully machine-readable format [10]. Moreover, several biomedical ontologies are publicly available to enrich KGs [11], enabling the representation of domain-specific knowledge. In fact, over the past few years, biomedical ontologies and KGs have emerged as a tool for biomedical data integration and have been adopted in many machine learning applications, with KG embedding approaches [12] becoming increasingly popular [13].

This work tackles the challenge of integrating heterogeneous gene expression datasets in biomedical applications, focusing on diabetes prediction. We propose a novel approach that generates a KG to incorporate both gene expression data and domain-specific knowledge and then employs KG embedding methods to generate vector representations of patients. These patient representations serve as the input for a classifier to predict the likelihood of a patient having diabetes. We conducted an evaluation of the impact of integrating multiple gene expression datasets, which showed that incorporating other expression datasets and domain-specific knowledge improves diabetes prediction, emphasizing the efficacy of our approach. This work is developed in the context of the KI-DiabetesDetektion project, funded by the German Federal Ministry of Education and Research, that aims to integrate biomedical data from various sources and apply machine learning methods to improve the early-stage detection of Diabetes.

2. Related Work

Several works have been using gene expression data to predict diabetes, employing diverse methodologies and datasets. In Li *et al.* [14], a support vector machine classifier is used for the diagnosis of diabetes. While multiple datasets were extracted from the Gene Expression Omnibus database, the machine learning model was trained on only one dataset, with three additional datasets used for validation. Feature selection involved the identification of ten common genes across all datasets. Mansoori *et al.* [15] and Kazerouni *et al.* [16] focus on long non-coding RNAs potentially associated with diabetes type 2. Both studies incorporated data collected from 100 diabetic and 100 non-diabetic to train the classifiers. Mansoori *et al.* [15] employed logistic regression, whereas Kazerouni *et al.* [16] compare four classifiers (K-nearest neighbor, support vector machine, logistic regression, and artificial neural networks) to predict

diabetes type 2 using the expression values for specific long non-coding RNAs as input. Both studies suggest that increasing the dataset with a larger number of samples would likely improve the performance of the classifiers. Furthermore, some other approaches explore expression data for diabetes prediction without employing machine learning methods [17, 18, 9].

In the biomedical domain, the exploration of KGs has become increasingly prominent, with KG embedding methods emerging as particularly promising for capturing KG-based information [19]. These methods map entities and relationships in a KG into a lower-dimensional vector space while preserving graph structure and, in some cases, semantic information. Various types of KG embedding methods have been proposed to date. Translational models, exemplified by TransE [20], employ distance-based scoring functions to capture relationships between entities. On the other hand, semantic matching approaches, such as distMult [21], use similarity-based scoring functions to capture the latent semantics of entities and relations in their vector space representations. Walk-based methods, such as RDF2Vec [22], employ random walks to generate entity sequences as input to a neural language model that learns latent entity representations. Different walk-based approaches differ in their strategies for random walks and consideration of edge direction and type. In the context of biomedical KGs, characterized by rich hierarchical relations, walk-based approaches emerge as particularly well-suited, considering that these hierarchical relations can be more easily captured in walks.

3. Methodology

As discussed above, gene expression datasets typically only have few instances, and different datasets record different gene expressions. Thus, when training prediction models, one can either (1) use only one dataset, thereby having only little training data, or (2) try to combine multiple datasets. In the latter case, those are typically "incompatible" in the sense that they have different feature sets, i.e., a naive combination would lead to a larger dataset with lots of NULL values.

To overcome these challenges, we propose a methodology to integrate multiple expression datasets into a biomedical KG and then use it for diabetes prediction. Figure 1 shows an overview of this methodology. The first step corresponds to building the KG that integrates not only expression data from different datasets but also domain knowledge on protein function and protein interactions. Then, we generate a vector representation for each patient described in the biomedical KG. The last step involves giving the vectors as input for a classifier. The source code for our methodology is available on GitHub (https://github.com/ritatsousa/expressionKG).

3.1. Expression Data

Several studies have recently explored gene expression for diabetic and non-diabetic individuals, and the findings from these studies can be accessed in publicly available databases. The Gene Expression Omnibus (GEO) [23] is a public database maintained by the National Center for Biotechnology Information that archives high-throughput gene expression and other genomics datasets. Each GEO dataset represents a curated collection of biologically comparable GEO samples whose measurements are assumed to be calculated equivalently. The file associated with each dataset contains the raw gene expression data generated by microarrays. In addition to raw
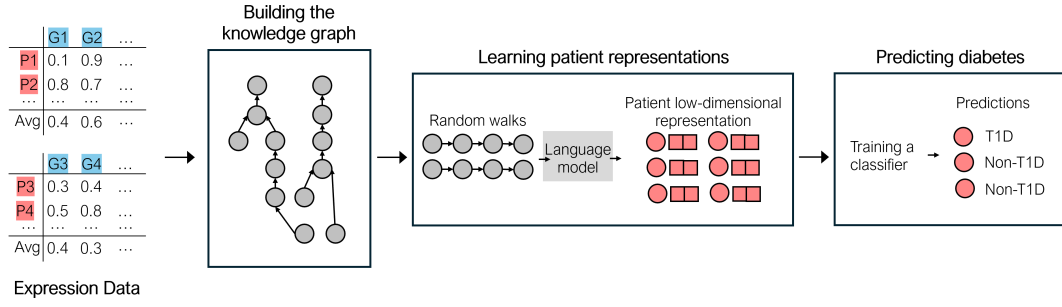
Figure 1: Overview of the proposed methodology with the main steps: building the KG, learning patient representations and predicting diabetes.

data, processed files containing normalized or transformed expression values may be included. In the latter scenario, the data is structured in a tabular format, with each row corresponding to a unique sample, columns representing different genes, and the cells containing specific expression values of those genes for each respective sample.

3.2. Building the Knowledge Graph

The KG is built by integrating two types of data sources: expression data and domain-specific knowledge. Figure 2 illustrates the integration of the two data sources into a KG.

Since our approach relies on KG graph embeddings for generating patient representations and most embedding approaches are not able to handle numeric literals [24], we adopt two different strategies to include the expression data in the KG:

- The first strategy involves representing patient gene expression values in a KG using *blank nodes and binning approaches*. Following the technique proposed in [24], we create bins from the set of expression values for each gene within a given dataset. The percentage of unique values defines the number of bins. To implement this, a blank node is generated to represent the expression value attributed to a specific gene for a given patient. This establishes an association wherein a patient is connected to a blank node, which, in turn, is linked to a bin representing the expression value and the corresponding gene. Let us consider a simplified example using RDF:
  (patientID, rdf:type, :Patient)
  (:geneID, rdf:type, :Gene)
  (:patientID, :hasExpression, _:x)
  (_:x, :isExpressionOfGene, :geneID)
  (_:x :hasValue :binID)
  where *_:x* denotes a blank node.
- The second strategy employs a *linking approach between patients and genes based on expression values*. A link between a patient and a gene is created when the patient's expression value for that gene is higher than the calculated average expression value for the gene within the dataset.
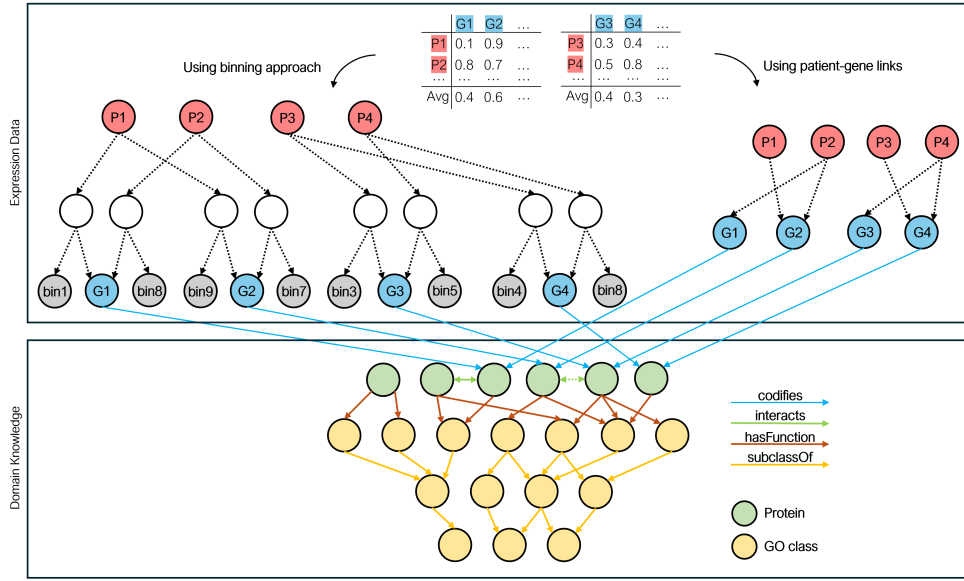
Figure 2: Schema of the two types of data sources and how they are integrated into the KG.

The domain-specific knowledge includes the Gene Ontology (GO) [25], GO annotation data [26], and protein-protein interaction (PPI) data [27]. The GO defines a hierarchy of classes that describe protein functions that can be represented as a graph where nodes are GO classes and edges define relationships between them. The GO encompasses three distinct domains for characterizing functions: the biological processes a protein is involved in, the molecular functions a protein performs, and the cellular components where a protein is located. These three domains of GO are represented as separate root ontology classes since they do not share any common ancestor. The GO annotation data refers to assigning functions represented as GO classes to proteins represented as links in the graph. Finally, the PPI data is extracted from STRING [27], one of the largest available PPI databases that integrates physical interactions and functional associations between proteins collected from several sources.

To bridge the gap between the two types of data sources, the expression data and the domain-specific knowledge, a gene in the expression data graph is mapped to a protein in the domain-specific KG. Online ID mapping tools, namely UniProt ID Mapping tool[1], are used to convert identifiers between genes and proteins.

3.3. Learning Patient Representations

We propose to generate patient representations by leveraging the information of multiple gene expression datasets and domain knowledge. As a preliminary step, the KG is converted into a directed and labeled RDF graph, following the W3C's OWL to RDF Graph Mapping guidelines[2]. Next, our methodology employs RDF2Vec, a KG embedding method, to generate

---

[1]https://www.uniprot.org/id-mapping

[2]https://www.w3.org/TR/owl2-mapping-to-rdf/

the low-dimensional vector representations. RDF2Vec [22] is a path-based embedding method that generates random walks in a graph that take into consideration both edge direction and type, making it particularly suited to KGs. Word2vec, a language model, is then employed over random walks on the RDF graph to produce the embeddings.

Two distinct approaches are employed to represent patients: the first involves generating RDF2vec embeddings directly for the patients using the KG, while the second generates RDF2Vec embeddings for the genes present in gene expression datasets and represents patients as the weighted average of gene embeddings, determined by the respective gene expression values.

3.4. Predicting Diabetes

Diabetes prediction is formulated as a binary classification task, where the goal is to categorize a set of patients based on whether they have diabetes or not. Therefore, in the final step, the patient representations are fed into a decision tree [28] algorithm for training.

4. Evaluation

4.1. Data

Three diabetes-related GEO datasets (GSE15932, GSE30208, and GSE55098) are considered for this work (Table 1). These datasets comprise samples associated with two distinct groups: patients diagnosed with type 1 diabetes (T1D) and those serving as control subjects (non-T1D). The data from the three datasets are integrated into a KG described in Table 2.

Table 1
Number of samples, number of shared genes across different datasets, and references for each GEO dataset.

Dataset	Number of samples			Number of shared genes			Refs.
	Total	T1D	non-T1D	GSE30208	GSE15932	GSE55098	
GSE30208	63	37	26	368	0	0	[29, 30]
GSE15932	22	12	10	0	764	337	[31]
GSE55098	16	8	8	0	337	764	[32, 33]

Table 2
Number of triples, types of relations, GO classes and proteins in the KG.

	Number
Triples	2433477
Types of relations	56
GO classes	51375
Proteins	19169

4.2. Results and Discussion

To assess the efficacy of the proposed methodology, we analysed the diabetes performance on the GSE30208 dataset by enriching the training data with information from the GSE15932 and GSE55098 datasets. Since our approach involves integrating data from multiple expression datasets into a KG, we compare it against two baselines that employ the expression values of the patient directly as input for the classifier. The first baseline exclusively employs data from GSE15932 for training the classifier. The second baseline represents a more simplistic approach to adding information from other datasets. It involves merging all measured genes across datasets and setting the value to 0 when the patient does not have an expression value. We employed a stratified cross-validation strategy to ensure robust evaluation, dividing the GSE30208 dataset into five folds. The same five folds were used throughout all experiments. The reported results represent the average performance over these five folds. Figure 3 illustrates the employed cross-validation strategy.
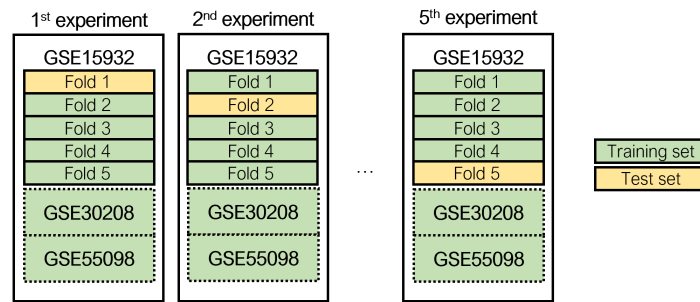

Figure 3: Experimental strategy to split the GSE30208 dataset and enrich with data from the GSE15932 and GSE55098 datasets.

Table 3 shows the accuracy, precision, recall, f-measure, weighted average f-measure and the area under the ROC curve for the baselines and the proposed methodology. The two baseline results indicate that simplistically adding information from other datasets does not enhance performance. In fact, it appears to introduce noise to the classifier. This outcome is not unexpected, as the integration of information from diverse datasets is lacking, leading to an ineffective impact on overall performance. However, by integrating the information from other datasets in a KG, it becomes evident that training a model with diverse datasets improves the performance of machine learning models in all metrics, with the exception of precision. Therefore, it confirms our hypothesis that injecting other expression datasets can improve the performance of machine learning models.

However, there are performance variations between the different alternatives of our approach. For the integration of expression data into the KG, we explore the use of blank nodes and binning approaches versus a linking method based on expression values to link patients and genes. In generating patient representations, we employed two strategies: direct learning of embeddings for patients in the KG; or learning embeddings for genes and representing patients as the weighted average of gene embeddings. This last strategy is independent of the strategy employed to represent the expression data in the KG, so Table 3 presents only three alternatives.

Table 3

Average diabetes prediction performance on the GSE30208 dataset for the baselines and the proposed methodology. Acc stands for accuracy, Pr stands for precision, Re stands for recall, F1 stands for f-measure, WAF stands for weighted average f-measure, and AUC stands for area under the ROC curve. For each metric, the best value is in bold.

	Acc	Pr	Re	F1	WAF	AUC
Baselines						
Only one dataset	0.554	**0.708**	0.561	0.578	0.529	0.560
Using all the datasets	0.442	0.650	0.314	0.396	0.422	0.474
Our methodology						
Patient rep. using weighted avg. gene emb.	**0.619**	0.677	**0.739**	**0.683**	**0.589**	**0.606**
Patient rep. using KG with binning approach	0.481	0.565	0.579	0.551	0.460	0.466
Patient rep. using KG with patient-gene links	0.583	0.638	0.604	0.595	0.567	0.578

Comparing the performance results of Table 3, the strategy involving the weighted average of gene embeddings for patient representation emerges as particularly promising because it consistently outperforms the other alternatives. Using links between patients and genes based on the expression values is the second-best strategy, and it still improves performance across several metrics compared to the baselines. Employing the binning approach achieves the worst results, performing worse for many metrics than the baseline. These results may be attributed to the inherent limitations of our path-based embedding method since genes and gene-expression values exist on separate paths.
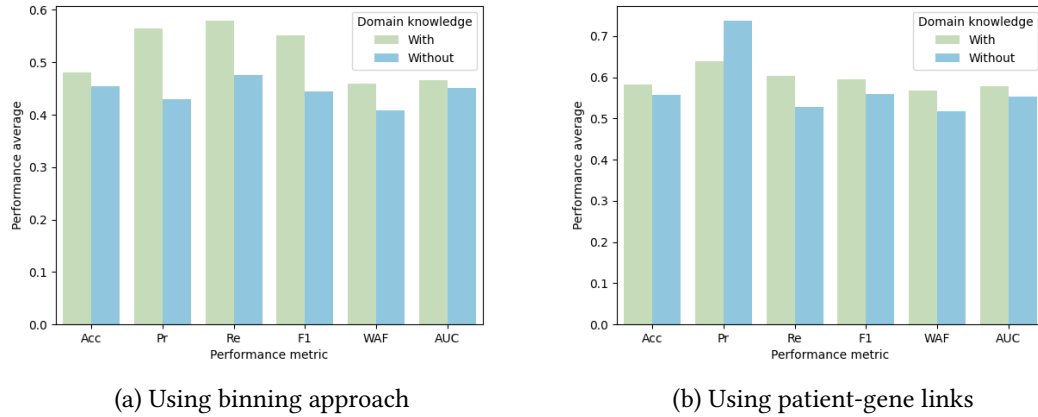

(a) Using binning approach   (b) Using patient-gene links

Figure 4: Performance comparison between using a KG with domain knowledge and without domain knowledge generated with two approaches: binning and patient-gene links. Acc stands for accuracy, Pr stands for precision, Re stands for recall, F1 stands for f-measure, WAF stands for weighted average f-measure, and AUC stands for area under the ROC curve.

Since we are interested in investigating the impact of domain-specific knowledge on integrating data from different datasets, we evaluated the diabetes prediction performance using a KG that only contains gene expression data. Figure 4 illustrates the performance variations

observed when employing a KG with domain knowledge alongside expression data, compared to utilizing a KG with expression data alone. The performance decreases when the domain knowledge is removed for both strategies of building the KG. This demonstrates that knowledge about protein functions and interactions can play an important role in integrating data from datasets measuring gene expression across different genes.

5. Conclusion

Several diabetes prediction approaches rely on the analysis of expression data, which provide a detailed molecular profile reflecting gene activity and regulation and therefore can uncover relationships between specific genes and the development of diabetes. However, exploring expression data in machine learning presents its own set of challenges. Existing expression datasets related to diabetes have a very low number of samples what can be a limitation for data-driven methods such as machine learning algorithms. Therefore, the integration of multiple expression datasets can address the issue of limited samples and, at the same time, offer a comprehensive perspective on the complex factors influencing diabetes.

We have developed an approach that enables a comprehensive representation of gene expression data from different datasets within a KG. Through semantic links and domain-specific knowledge, KGs can create a unified knowledge space to connect datasets from distinct studies. In this work, we have explored different strategies to include the expression data in the KG and different strategies to represent the patients within the KG using KG embedding methods. The results of our experiments showed that integrating gene expression data in a KG is able to improve the performance of diabetes prediction.

The proposed approach is versatile and can be extended to the prediction of other diseases. In addition, since graph neural networks has gained substantial traction recently, as future work, we aim to investigate how can these architectures explicitly designed for graph structures can be used rather than the conventional process of generating embeddings and given them as input for classical machine learning methods such as decision trees.

Acknowledgments

References

[1] D. Care, Care in diabetes—2022, Diabetes care 45 (2022) S17.

[2] V. Jaiswal, A. Negi, T. Pal, A review on current advances in machine learning based diabetes prediction, Primary Care Diabetes 15 (2021) 435–443.

[3] P. Sonar, K. JayaMalini, Diabetes prediction using different machine learning approaches, in: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2019, pp. 367–371.

[4] A. Mujumdar, V. Vaidehi, Diabetes prediction using machine learning algorithms, Procedia Computer Science 165 (2019) 292–299.

[5] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, M. Hasan, Diabetes prediction using ensembling of different machine learning classifiers, IEEE Access 8 (2020) 76516–76531.

[6] D. Bertsimas, N. Kallus, A. M. Weinstein, Y. D. Zhuo, Personalized diabetes management using electronic medical records, Diabetes care 40 (2017) 210–217.

[7] Y. Tang, R. Gao, H. H. Lee, Q. S. Wells, A. Spann, J. G. Terry, J. J. Carr, Y. Huo, S. Bao, B. A. Landman, Prediction of type ii diabetes onset with computed tomography and electronic medical records, in: Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures: 10th International Workshop, ML-CDS 2020, and 9th International Workshop, CLIP 2020, Held in Conjunction with MICCAI 2020, Springer, 2020, pp. 13–23.

[8] H. Xiao, J. Gao, L. Vu, D. S. Turaga, Learning temporal state of diabetes patients via combining behavioral and demographic data, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 2081–2089.

[9] J. Liu, S. Liu, Z. Yu, X. Qiu, R. Jiang, W. Li, Uncovering the gene regulatory network of type 2 diabetes through multi-omic data integration, Journal of Translational Medicine 20 (2022) 604.

[10] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Computing Surveys (Csur) 54 (2021) 1–37.

[11] D. L. Rubin, N. H. Shah, N. F. Noy, Biomedical ontologies: a functional perspective, Briefings in bioinformatics 9 (2008) 75–90.

[12] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Transactions on Knowledge and Data Engineering 29 (2017) 2724–2743.

[13] M. Kulmanov, F. Z. Smaili, X. Gao, R. Hoehndorf, Semantic similarity and machine learning with ontologies, Briefings in Bioinformatics 22 (2021) bbaa199.

[14] J. Li, J. Ding, D. Zhi, K. Gu, H. Wang, et al., Identification of type 2 diabetes based on a ten-gene biomarker prediction model constructed using a support vector machine algorithm, BioMed Research International 2022 (2022).

[15] Z. Mansoori, H. Ghaedi, M. Sadatamini, R. Vahabpour, A. Rahimipour, M. Shanaki, L. Saeidi, F. Kazerouni, Downregulation of long non-coding rnas linc00523 and linc00994 in type 2 diabetes in an iranian cohort, Molecular biology reports 45 (2018) 1227–1233.

[16] F. Kazerouni, A. Bayani, F. Asadi, L. Saeidi, N. Parvizi, Z. Mansoori, Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding rnas expression: a comparison of four data mining approaches, BMC bioinformatics 21 (2020) 1–13.

[17] L. Saeidi, H. Ghaedi, M. Sadatamini, R. Vahabpour, A. Rahimipour, M. Shanaki, Z. Mansoori, F. Kazerouni, Long non-coding rna ly86-as1 and hcg27_201 expression in type 2 diabetes mellitus, Molecular biology reports 45 (2018) 2601–2608.

[18] H. Zhu, X. Zhu, Y. Liu, F. Jiang, M. Chen, L. Cheng, X. Cheng, et al., Gene expression profiling of type 2 diabetes mellitus by bioinformatics analysis, Computational and Mathematical Methods in Medicine 2020 (2020).

[19] D. Chang, I. Balažević, C. Allen, D. Chawla, C. Brandt, R. A. Taylor, Benchmark and best

practices for biomedical knowledge graph embeddings, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, volume 2020, NIH Public Access, 2020, p. 167.

[20] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Proceedings of NIPS 2013, Curran Associates Inc., Red Hook, NY, USA, 2013, p. 2787–2795.

[21] B. Yang, W. tau Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, 2015.

[22] P. Ristoski, H. Paulheim, RDF2Vec: RDF graph embeddings for data mining, in: Proceedings of the 15th International Semantic Web Conference, Springer International Publishing, Cham, Switzerland, 2016, pp. 498–514.

[23] E. Clough, T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, et al., Ncbi geo: archive for gene expression and epigenomics data sets: 23-year update, Nucleic Acids Research 52 (2024) D138–D144.

[24] P. Preisner, H. Paulheim, Universal preprocessing operators for embedding knowledge graphs with literals (2022).

[25] G. Consortium, The Gene Ontology resource: enriching a GOld mine, Nucleic Acids Research 49 (2021) D325–D334.

[26] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, C. O'Donovan, The GOA database: gene ontology annotation updates for 2015, Nucleic Acids Research 43 (2015) D1057–D1063.

[27] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, et al., The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets, Nucleic Acids Research 49 (2021) D605–D612.

[28] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1986) 81–106.

[29] Series GSE30208, 2014. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30208.

[30] H. Kallionpää, L. L. Elo, E. Laajala, J. Mykkänen, I. Ricano-Ponce, M. Vaarma, T. D. Laajala, H. Hyöty, J. Ilonen, R. Veijola, et al., Innate immune activity is detected prior to seroconversion in children with hla-conferred type 1 diabetes susceptibility, Diabetes 63 (2014) 2402–2414.

[31] Series GSE15932, 2012. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15932.

[32] Series GSE55098, 2014. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55098.

[33] M. Yang, L. Ye, B. Wang, J. Gao, R. Liu, J. Hong, W. Wang, W. Gu, G. Ning, Decreased mi r-146 expression in peripheral blood mononuclear cells is correlated with ongoing islet autoimmunity in type 1 diabetes patients 1, Journal of diabetes 7 (2015) 158–165.