BioGraphDB: a New GraphDB Collecting Heterogeneous Data for Bioinformatics Analysis

Antonino Fiannaca, Massimo La Rosa, Laura La Paglia, Antonio Messina, Alfonso Urso
ICAR-CNR, National Research Council of Italy
Palermo, Italy
Email: {fiannaca, larosa, lapaglia, messina, urso}@pa.icar.cnr.it

***Abstract*—Current bioinformatics databases provide huge amounts of different biological entities such as genes, proteins, diseases, microRNA, annotations, literature references. In many case studies, a bioinformatician often needs more than one type of resource in order to fully analyse his data. In this paper, we introduce BioGraphDB, a bioinformatics database that allows the integration of different types of data sources, so that it is possible to perform bioinformatics analysis using only a comprehensive system. Our integrated database is structured as a NoSQL graph database, based on the OrientDB platform. This way we exploit the advantages of that technology in terms of scalability and efficiency with regards to traditional SQL database. At the moment, we integrated ten different resources, storing and linking data about genes, proteins, microRNAs, molecular pathways, functional annotations, literature references and associations between microRNA and cancer diseases. Moreover, we illustrate some typical bioinformatics scenarios for which the user just needs to query the BioGraphDB to solve them.**

Keywords–Integrated database; Graph database; GraphDB; OrientDB; Bioinformatics database.

I. INTRODUCTION

In the last years, the use of computational approaches allowed researchers in bioinformatics and systems biology to produce, store and share a lot of data, such as genes, proteins, metabolic pathways, and so on. In most cases, data are collected in different databases, each of which has a proper framework and storage technology. For this reason, although the scientific community makes available to biologists and bioinformaticians a large amount of data, it is a big challenge to interconnect results from heterogeneous data sources, where each database can identify the same biological entity on one's own account. For all those reasons, it is important to provide an integrated database offering, in a modular framework, all the information contained in different available databases.

In this work, we propose BioGraphDB, an efficient bioinformatics NoSQL graph database, collecting data related to genes, microRNA (miRNA), proteins, pathways and diseases from 10 online public resources. Since we aim at integrating heterogeneous resources modelling pathways, interactions and relations among a lot of biological entities, we chose to implement a graph database; it has been highlighted by [1] that graph databases both allow for efficient queries and give advantages in scalability with respect to any relational database. The proposed database is built on the OrientDB platform [2], because previous works [3], [4] demonstrated that it outperforms the other NoSQL databases in terms of flexibility and performances.

Moreover, in this work we propose some cases of study in the field of biological and clinical research that can be resolved using the proposed database. The paper has the following structure: in Section II similar integrated DBs are presented; Section III presents the main components of the proposed DB; in Section IV it is described how the different resources have been imported and linked each other; in Section V we present four application scenarios and finally in Section VI some conclusions, as well as future developments, are drawn.

II. RELATED WORKS

Due to the overwhelming size and type of biological data, the need of biological databases that integrate many different resources has risen. The National Center for Biotechnology Information (NCBI) [5] perhaps offers the most popular platform of integrated biological databases. It includes, among the others, the Entrez database [6] consisting of 37 different databases containing data related to genes, proteins, taxonomy, gene expression and so on; the PubMed system [7] for the scientific literature, the RefSeq [8] database that hosts non-redundant sets of curated genomic, proteomic and transcriptomic sequences; and the BioSystems [9] database that integrates and cross-links information about molecular pathways. The molecular pathways are at the basis of the KEGG integrated databases project [10]. In addition to information about pathways, KEGG has also information about genes, compounds, reactions, diseases and drugs.

In recent years, with the focus on the study of non-coding RNA and especially miRNA, many integrated resources have been developed considering miRNA as their core. The miRò knowledge base [11] is a system that integrates data about miRNAs, their validated and predicted gene targets, functional annotations provided by Gene Ontology (GO) [12] and gene-disease relations taken from the Genetic Association Database (GAD). Another miRNA-centric integrated database is miRWalk 2.0 [13], [14]. Besides data about miRNAs, GO annotations, miRNA-mRNA interactions and gene-disease associations, miRWalk stores and integrates data about pathways and gene and protein classes. Moreover miRWalk web service implements several pre-defined search methods that allow the user to query the database in order to find, for example, gene-miRNA-pathway relations, gene-miRNA-GO annotations, disease-miRNA relations. Even if miRWalk integrates several type of biological data, it however only allows to query them using the above described pre-defined search tools. The proposed BioGraphDB, in turn, lets the user access all of the data in order to assemble his own set of queries, thanks to its graph structure and a specialized query language (see Section IV and Section V).

Since in many cases it is needed only a limited set of bioinformatics resources, it would be useful to build a cus-

 ISBN: 978-1-61208-488-6

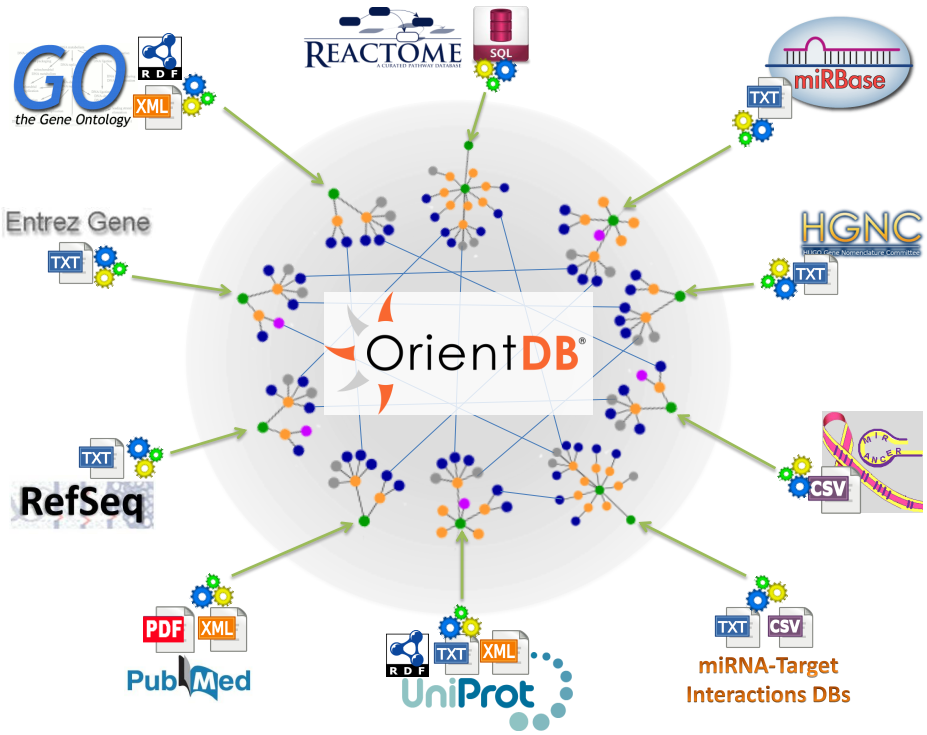

Figure 1. A graphical representation of the proposed integrated database based on OrientDB framework.

tom integrated database. The Java BioWareHouse (JBioWH) platform [15] offers a Java library that allows the importation and integration of different data sources into a SQL-based framework which defines a set of data types related to bio-entities, such as genes, proteins, pathways and drugs.

All the above described integrated databases, as well as the JBioWH library, are built upon a standard SQL architecture. With advances in the developing of NoSQL databases, which provide a more flexible and performing environment, new ways for integrating different resources have been studied. For example authors in [16] presented ncRNA-DB, a NoSQL database based on the OrientDB platform that put together many biological resources that deal with several classes of non-coding RNA (ncRNA) such as miRNA, long-non-coding RNA (lncRNA), circular RNA (circRNA) and their interactions with genes and diseases. More recently a graph-based database, called Bio4j, has been developed by [17]. Bio4j is based on a Java library that allows to build an integrated cloud-based data platform upon a graph structure. Bio4j is protein-centric, in fact it only includes data about proteins, GO and enzymes.

Since Bio4j has fewer resources rather than our proposed BioGraphDB (see Section III-B), it is difficult do directly compare them, especially because the number and type of resolvable scenarios are quite different (see [17] and Section V).

III. BioGraphDB components

All the components used in this work are discussed in the following. In details, in the next subsection we introduce the OrientDB framework, that represents the platform used to build the proposed work. In the subsection III-B we define all the databases we used in this work. Figure 1 shows a graphical representation of public databases integration.

A. OrientDB

OrientDB is an open source NoSQL database management system (DMBS) developed in Java by Orient Technologies LTD. It collects features of document databases and graph databases, including object orientation. In graph mode, referenced relationships are like edges, accessible as first-class objects with a start vertex, end vertex, and properties. This interesting feature let us represent a relational model as a document-graph model, maintaining the relationships.

OrientDB supports an extended version of SQL, to allow all sort of Create, Read, Update and Delete (CRUD) and query operations, and Atomicity, Consistency, Isolation, Durability (ACID) transactions, helpful to recover pending document at the time of crash. It is easily embeddable and customizable and it handles HTTP Requests, RESTful protocols and JSON without any 3rd party libraries or components. Finally, it is fully compliant with TinkerPop Blueprints [18], the standard of graph databases. It is distributed under the open source Apache 2 license [19], therefore it is totally free for any kind of use and its enterprise features are not limited.

B. Data Source

In order to build a database containing the most updated resources related to genes, proteins, miRNAs, metabolic pathways and their references in literature, it is useful to integrate the last versions of different publicly available data sources. For this aim, we take into account those on-line databases that represent the state-of-art in bioinformatics. In the following the list of databases we have considered for populating the proposed graph database, as showed in Figure 1.

1) miRBase [20]: The microRNA database (miRBase) is a searchable database of published miRNA sequences and annotation. It contains both hairpin and mature sequences of

 ISBN: 978-1-61208-488-6

223 species, and for each of them, it provides name, keywords, genomic location, references and annotations.

2) UniProtKB [21]: The UniProt Knowledgebase (UniProtKB) is the largest public collection of annotated functional information on proteins and it is updated every four weeks. It stores both computationally analysed and manually annotated records, including classifications, cross-references and quality indications available to scientific researchers.

3) Gene Ontology [12]: The Gene Ontology (GO) is the most complete and daily updated public resource for genes and proteins annotation. It provides annotations for gene products in biological processes, cellular components and molecular functions.

4) Reactome [22]: Reactome is a database containing validated metabolic pathways in human biology and computationally inferred pathways for 20 non-human species. Each pathway is annotated as a set of biological events, dealing with genes and proteins.

5) Entrez Gene [6]: The NCBI Entrez Gene database contains a wide set of details related to all the genes that have been studied in literature. For each gene, there is a record containing a lot of information, such as the genomic context, a list of ortholog/homolog genes, annotated pathways, interactions with other genes and so on.

6) Refseq [8]: The Reference Sequence (RefSeq) database is a collection of computationally and manually curated annotations for identification and characterization of genomes, transcripts and proteins.

7) Pubmed: Pubmed is a structured information resource on scientific publications in the field of biomedical literature. It allows to perform clinical queries for specific studies, categories and scopes. Due to copyright restrictions, only an open-access subset of this database is available for download.

8) mirCancer [23]: The microRNA Cancer association database (mirCancer) provides associations between miRNAs and related human cancers Pubmed entries. These associations are first extracted from Pubmed database by means of text mining algorithms and then manually revised. In addition, mirCancer gives, for each association, the miRNA expression profile.

9) HGNC [24]: The HUGO Gene Nomenclature Committee (HGNC) is the authority responsible for the gene nomenclatures (also known as gene symbols) for the human species. The HGNC database contains, for each gene symbol, a list of synonyms and a list of corresponding entries in the most popular gene databases (e.g. Refseq, Entrez gene). HGNC is the main source for synonyms disambiguation for genes and proteins.

10) miRNA-Target Interactions: This resource is a collection of publicly available miRNA-target interactions databases. It contains both validated and predicted interactions. The published experimentally validated interactions, including their experimental conditions, are provided by mirTarBase database [25]. A list of putative interactions are obtained by combining results of five different databases: miRNATIP [26], TargetScan [27], Diana micro-T [28], Pita [29] and miRanda [30].

IV. Data Integration

The publicly databases listed in the previous section give us a huge amount of data, that we have to integrate in an harmonious and consistent way. It is relatively easy to read and parse the various source files, but they often contain redundancies and useless data for our purpose, because, for example, at the moment we are only interested in the human species. Loading and linking the actual useful data is the goal.

Moreover, the databases are available for download in several different formats, such as tab-delimited plain-text, structured XMLs, SQL database dumps. The latest available release of OrientDB has a powerful tool to move data from and to a database by executing an Extract-Transformer-Loader (ETL) process, described by a JSON configuration file. However, its Extractor supports almost all data source types but XML. Therefore, in order to avoid mixed solutions, we decided to develop an ad hoc set of Java based ETLs.

As general rule, each biological entity and its properties have been mapped respectively into a vertex and its attributes, and each relationship between two biological entities has been mapped into an edge. If a relationship has some properties, they are also saved as edge's attributes. Vertices and edges are grouped into classes, according to the nature of the entities. For example, all the genes imported from NCBI Gene become instances of the *gene* vertex class, and all the proteins from UniProtKB become instances of the *protein* vertex class. Moreover, all the relationships between genes and proteins extracted from HGNC, in the form of "gene *G* codes for protein *P*", become instances of the *coding* edge class.

The ETLs can be grouped in the following five categories:

- *Pubmed ETL*: It is not a real ETL, because actually we do not import any Pubmed publication. It is just used to create a vertex class used to store those *Pubmed IDs* found in the other databases.
- *Tab-delimited ETLs*: They were used to import NCBI Gene, miRNA-target interactions, HGNC, and mirCancer. Because all interactions have several virtually-searchable attributes, they have been mapped to vertices and then linked to the related gene and miRNA. By using the *protein-coding gene* field from HGNC, we were able to link each gene to its encoded proteins.
- *XML ETLs*: Starting from the related XML Schema Definition (XSD) [31] file and thanks to the unmarshalling capabilities of the standard JAXB library [32], they were used to import UniprotKB and GO.
- *miRBase ETL*: miRBase is available in a EMBL format text file, hence we used the *BioJava* library [33], in order to process the data in a simple and efficient way.
- *Reactome ETL*: The Reactome database import was not so easy. It is available for download only as SQL database dumps and its schema is not documented, hence we have installed the relational DBMS MySQL [34] and followed the available installation guide [35] in order to properly load the database from the dumps. After studying the database structure and tables definitions, we have created some ad hoc SQL views to extract the useful data, afterwards exported as a set of tab-delimited text files. Finally, we were able to import pathways data and to link the proteins to their pathways.

HGNC and UniprotKB databases provide conversion tables storing the synonyms for respectively gene and protein names,

 ISBN: 978-1-61208-488-6

TABLE I. OVERALL SIZE OF BOTH IMPORTED AND PROPOSED DATABASES.

External DBs		BioGraphDB	
Public DBs size	*Overall input lines*	*Vertices*	*Edges*
$>$ 10 GB	$>$ 185 millions	$\sim$ 7.4 millions	$\sim$ 15 millions

as well as their accession IDs to the most common biobanks. In our BioGraphDB we inserted those data into two different vertexes and linked them to the corresponding gene and protein vertexes. The same strategy can be applied for managing synonyms for other kinds of data.

The imported DBs have not overlapping information, that eventually could be contradictory, because we selected one source for each considered biological entity. In any case, when a new database will be imported, its data will be labelled (as attribute) with details about the source. For instance, if we import more than a miRNA target prediction database, then each prediction will contain an attribute declaring its original source. The advantage of this representation is that a user can define specific queries implementing consensus among different predictors or apply proper filters.

In order to guarantee data consistency and proper relationships, ETLs were executed in a precise order. Since each imported DB has dependencies with the other ones, it is of course important that all the depending resources are already present into the graphDB when a new resource is loaded. The following importation order assures that the dependencies among the integrated resources are correctly satisfied:

1) Pubmed *(schema creation)*
2) NCBI gene *(import)*
3) miRBase *(import, links to Pubmed)*
4) mirCancer *(import, links to miRBase and to Pubmed)*
5) miRNA-target interactions *(import, links to gene and to miRBase)*
6) UniprotKB *(import, links to Pubmed)*
7) HGNC *(links from genes to proteins, gene synonyms import, links from synonyms to gene)*
8) Reactome *(import, links from pathways to proteins)*
9) GO *(import, links to genes and to pathways)*

The import process lasted several hours and most of the time was spent in the creation of the vertices and links related to the miRNA-gene interactions. The size of both imported DBs, in terms of data size and number of input lines, and BioGraphDB, in terms of number of vertices and edges, is reported in Table I. The whole graph assembled by means of the intgration of all the DBs can be traversed using proper query languages, such as Gremlin [18]. Each graph traversal represents a set of queries that are enough in order to solve several bioinformatics scenarios, and some of them will be described in Section V.

V. RESULTS

BioGraphDB can be used for the analysis in clinical research of different real life problems. Here we briefly introduce four scenarios representing typical bioinformatics problems that can be faced by means of suitable queries over the proposed DB. As an example, for the last scenario we provide a more detailed explanation and a query in the graph traversal language Gremlin [18] that resolves it.

```
g.V('name',cancer_name)
  .out('cancer2mirna')
  .out('precursorOf')
  .in('interactingMiRNA')
  .filter{it.energy<-energy_score}
  .out('interactingGene')
  .out('coding').dedup()
  .in('contains')
  .path{it.name}{it.accession}{it.accession}
       {it.transcriptId}{it.symbol}
       {it.name}{it.name}
```

Figure 2. A gremlin query for the proposed "*Target analysis of differentially expressed miRNAs in cancer*" scenario.

- *Analysis of gene functions and pathways*. Starting from the gene ID or gene sequence it is possible to investigate its role in the cellular context by exploring its functional annotations and location in pathways. Moreover it can be investigated the enrichment of that gene. This scenario requires the use of different databases: Entrez gene, RefSeq, GO, Reactome.
- *Analysis of protein motifs linked to cellular pathway*. The aim is searching the most representative protein motifs related to a specific cellular pathway. In this context, the study can be implemented by means of functional annotations related to these proteins. This scenario can be resolved using 3 databases: UniProt, Reactome, Gene Ontology.
- *Analysis of tumour-suppressor/oncogenic miRNA*. Starting from group of genes involved in a specific cellular pathway or cellular condition it is possible to identify potential miRNA targets that could have oncogenic or tumour-suppressor functions. This implies the use of 4 resources: Reactome, miRNA-target interactions, mirBase, mirCancer.
- *Target analysis of differentially expressed miRNAs in cancer*. Starting from a list of differentially expressed miRNAs linked to a specific disease, we would verify what are the major target proteins of these miRNAs belonging to particular cellular pathways. This analysis needs the use of 4 resources: mirCancer, mirBase, miRNA-target interactions, Reactome.

With regard to the last scenario, using publicly available resources, the following interactions steps are required. First of all, starting from a specific cancer type, a set of differentially expressed (DE) miRNAs can be obtained by the miRCancer database. The obtained miRNAs represent the input for the miRNA target interaction tools. Querying those tools, a list of putative miRNA targets is obtained. Filtering by energy scores, it is possible to evidence those targets that are more strongly linked to the DE miRNAs. The last step of the analysis is to verify if there are specific pathways that the selected targets belong to. This last step can be done through the use of pathways analysis tools such as Reactome. Reactome, in fact, given a list of input genes, provides a set of pathways containing those genes. A typical way in order to solve the described scenario would be to use each different DB (mirCancer, mirBase, miRNA-target predictors, Reactome) at once. In this situation, the user has to collect intermediate results and has to gain enough skill for using all the DBs. Instead of querying each biological resource singularly, all

 ISBN: 978-1-61208-488-6

```
gremlin> g = new OrientGraph("remote:localhost/biorient");
==>orientgraph[remote:localhost/biorient]
gremlin> graph = g.V('name','acute lymphoblastic leukemia').out('cancer2mirna').out('precursorOf').in('interactingMiRNA').filter{it.energy<-30}
                                                           .out('interactingGene').out('coding').dedup().in('contains').dedup()
                                                           .path{it.name}{it.name}{it.accession}{it.transcriptId}{it.symbol}{it.name}{it.name}
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, Regulation of pyruvate dehydrogenase (PDH) complex]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, Pyruvate metabolism]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, Pyruvate metabolism and Citric Acid (TCA) cycle]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, The citric acid (TCA) cycle and respiratory electron transport]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, Metabolism]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003csm.2, ZNF589, ZN589_HUMAN, Generic Transcription Pathway]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003csm.2, ZNF589, ZN589_HUMAN, Gene Expression]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003vrn.2, SLC35B4, S35B4_HUMAN, Transport of nucleotide sugars]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003vrn.2, SLC35B4, S35B4_HUMAN, Transport of vitamins, nucleosides, and related molecules]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003vrn.2, SLC35B4, S35B4_HUMAN, Transmembrane transport of small molecules]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003vrn.2, SLC35B4, S35B4_HUMAN, SLC-mediated transmembrane transport]
```

Figure 3. BioGraphDB response to the gremlin query depicted in Figure 2 for the proposed "*Target analysis of differentially expressed miRNAs in cancer*" scenario.

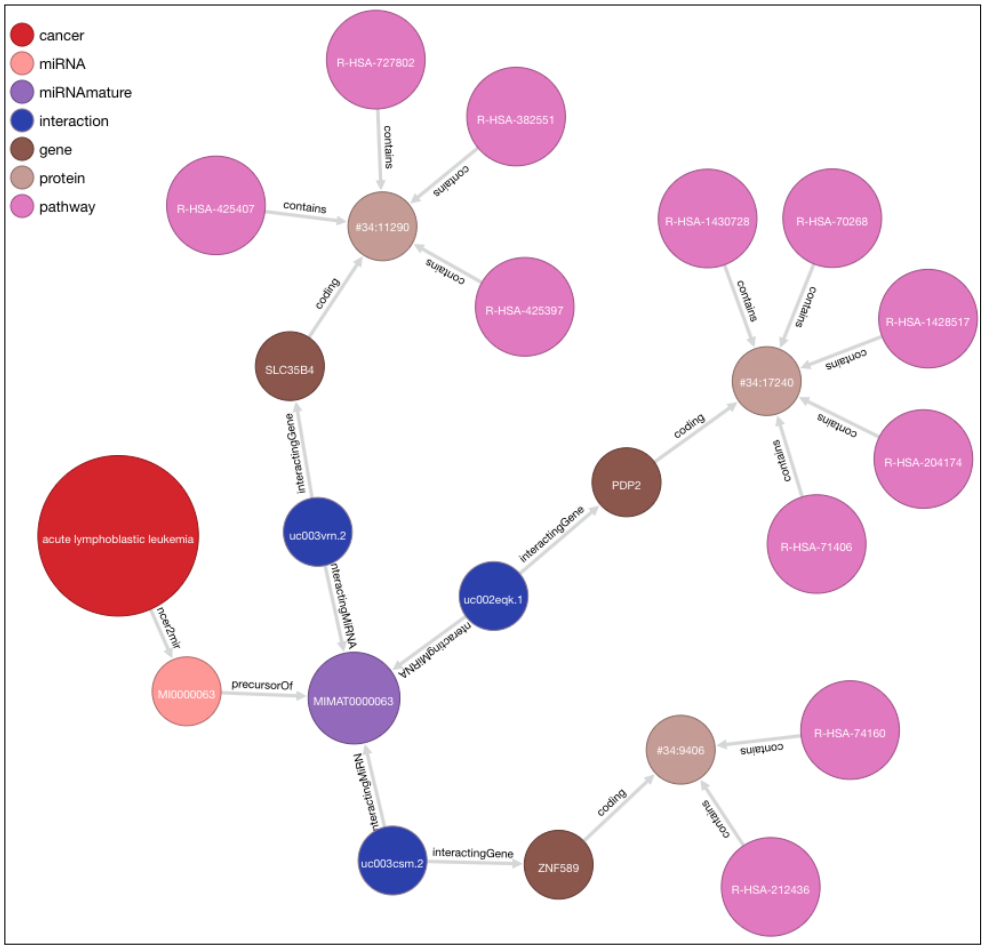

Figure 4. A graphical representation of the response produced by BioGraphDB, as seen in Figure 3. Starting from a specific disease ("*acute lymphoblastic leukaemia*"), we obtain 11 correlated biological pathways, marked with their Reactome ID.

of these steps can be easily performed using our integrated database by means of the Gremlin query shown in Figure 2.

For instance, if we set the cancer_name to "*acute lymphoblastic leukaemia*" and the energy_score threshold to "−30", we obtain as result eleven pathways, as showed in Figure 3. Figure 4 reports a graph representation of this result: starting from the "*acute lymphoblastic leukaemia*" disease, we obtain the "*hsa-let-7b*" DE miRNA, that interact with three genes (*SLC35B4*, *ZNF589*, *PDP2*). Each gene codes for a protein, that, in turns, is contained in at least a biological pathway. In this scenario, the query provides eleven Reactome pathways, marked with their Reactome ID. The complete set of results is summarized in Table II, where we reported the miRNA name and the pathway descriptions lacking in Figure 4.

VI. Conclusions and Future Works

In this work, we proposed BioGraphDB, an integrated graph database for biological data. This database was designed to overcome problems related to the lack of a structural organization and interoperability of publicly available biological

 ISBN: 978-1-61208-488-6

TABLE II. RESULT OF THE GREMLIN QUERY IN FIGURE 2.

Pathology	Mature miRNA	Gene	Reactome Pathway	Pathway Description
Acute lymphoblastic leukemia	hsa-let-7b-5p	SLC35B4	R-HSA-425407	SLC-mediated transmembrane transport
			R-HSA-727802	Transport of nucleotide sugars
			R-HSA-382551	Transmembrane transport of small molecules
			R-HSA-425397	Transport of vitamins, nucleosides, and related molecules
		PDP2	R-HSA-1430728	Metabolism
			R-HSA-70268	Pyruvate metabolism
			R-HSA-1428517	The citric acid cycle and respiratory electron transport
			R-HSA-204174	Regulation of pyruvate dehydrogenase
			R-HSA-71406	Pyruvate metabolism and Citric Acid
		ZNF589	R-HSA-74160	Gene Expression
			R-HSA-212436	Generic Transcription Pathway

resources. Finally we presented some cases of study where the use of the database can give a concrete advantage to the scientific community. Because our BioGraphDB stands at a prototypal stage, we are unable to provide at the moment a full performance evaluation, that will be done in future works.

Further developments will be done in the near future. Of course, thanks to the flexibility of the proposed database, other biological resources will be integrated where necessary. At the same time, we are developing proper automated mechanism in order to update on a regular schedule our BioGraphDB with the latest releases of its integrated DBs. After the data sources integration, we will develop a collection of web services with a common user-friendly web-interface and explicit search methods implementing proper database views. This way, it will be possible to solve some of the most common bioinformatics scenarios, like the ones proposed in this paper. In addition, we are working on a web service in order to provide the users a computer aided methodology to build their own custom views and search methods.

REFERENCES

[1] C. T. Have and L. J. Jensen, "Are graph databases ready for bioinformatics?" *Bioinformatics*, vol. 29, no. 24, pp. 3107–3108, 2013.

[2] Orient Technologies LTD, "OrientDB." [Online]. Available: http://orientdb.com [accessed: 2016-02-19]

[3] M. Dayarathna and T. Suzumura, "XGDBench: A benchmarking platform for graph stores in exascale clouds," in *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*. IEEE, 2012, pp. 363–370.

[4] A. Messina, P. Storniolo, and A. Urso, "Keep it simple, fast and scalable: a Multi-Model NoSQL DBMS as an (eb)XML-over-SOAP service," in *The 30th IEEE International Conference on Advanced Information Networking and Applications (AINA-2016)*. IEEE, 2016, pp. 220–225.

[5] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 41, no. D1, pp. D8–D20, 2013.

[6] G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans, "Entrez: molecular biology database and retrieval system." *Methods in enzymology*, vol. 266, pp. 141–62, 1996.

[7] "PubMed" [Online] Available: http://www.ncbi.nlm.nih.gov/pubmed [accessed: 2016-02-19]

[8] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 35, no. Database, pp. D61–D65, 2007.

[9] L. Y. Geer, et al., "The NCBI BioSystems database." *Nucleic acids research*, vol. 38, no. Database issue, pp. D492–6, 2010.

[10] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 44, no. D1, pp. D457–D462, 2016.

[11] A. Lagana, et al., "miRo: a miRNA knowledge base," *Database*, vol. 2009, 2009.

[12] The Gene Ontology Consortium, "Gene Ontology Consortium: going forward," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1049–D1056, 2015.

[13] H. Dweep, N. Gretz, and C. Sticht, "miRWalk Database for miRNA-Target Interactions." *Methods in Molecular Biology*, vol. 1182, pp. 289–305, 2014.

[14] H. Dweep and N. Gretz, "miRWalk2.0: a comprehensive atlas of microRNA-target interactions," *Nature Methods*, vol. 12, no. 8, pp. 697–697, 2015.

[15] R. Vera, Y. Perez-Riverol, S. Perez, B. Ligeti, A. Kertesz-Farkas, and S. Pongor, "JBioWH: an open-source Java framework for bioinformatics data integration," *Database*, vol. 2013, pp. bat051–bat051, 2013.

[16] V. Bonnici, F. Russo, N. Bombieri, A. Pulvirenti, and R. Giugno, "Comprehensive Reconstruction and Visualization of Non-Coding Regulatory Networks in Human," *Frontiers in Bioengineering and Biotechnology*, vol. 2, 2014.

[17] P. Pareja-Tobes, R. Tobes, M. Manrique, E. Pareja, and E. Pareja-Tobes, "Bio4j: a high-performance cloud-enabled graph-based data platform," Era7 bioinformatics, Tech. Rep., 2015.

[18] Apache Software Foundation, "Apache TinkerPop." [Online]. Available: http://tinkerpop.incubator.apache.org [accessed: 2016-02-19]

[19] Apache Software Foundation, "Apache License Version 2.0." [Online]. Available: http://www.apache.org/licenses/LICENSE-2.0 [accessed: 2016-02-19]

[20] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data." *Nucleic acids research*, vol. 39, no. Database issue, pp. D152–7, 2011.

[21] The UniProt Consortium, "UniProt: a hub for protein information," *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, 2015.

[22] D. Croft,et al., "The Reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 42, no. D1, pp. D472–7, 2014.

[23] B. Xie, Q. Ding, H. Han, and D. Wu, "miRCancer: a microRNA-cancer association database constructed by text mining on literature," *Bioinformatics*, vol. 29, no. 5, pp. 638–644, 2013.

[24] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford, "Genenames.org: the HGNC resources in 2015," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1079–D1085, 2015.

[25] S.-D. Hsu, et al., "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions," *Nucleic Acids Research*, vol. 42, no. D1, pp. D78–D85, 2014.

[26] A. Fiannaca, M. La Rosa, L. La Paglia, R. Rizzo, and A. Urso, "MiRNATIP: a SOM-based miRNA-target interactions predictor," *BMC Bioinformatics*, in press.

[27] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[28] M. D. Paraskevopoulou,et al., "DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows," *Nucleic Acids Research*, vol. 41, no. W1, pp. W169–W173, 2013.

[29] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microRNA target recognition." *Nature genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[30] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human microRNA targets," *PLoS Biology*, vol. 2, no. 11, 2004.

[31] World Wide Web Consortium (W3C), "W3C XML

Schema Definition Language (XSD) 1.1." [Online]. Available: https://www.w3.org/TR/xmlschema11-1/ [accessed: 2016-02-21]

[32] Java Community Process, "JSR 222: Java Architecture for XML Binding (JAXB) 2.0." [Online]. Available: https://jcp.org/en/jsr/detail?id=222 [accessed: 2016-02-21]

[33] A. Prlic, et al., "BioJava: an open-source framework for bioinformatics in 2012," *Bioinformatics*, vol. 28, no. 20, pp. 2693–2695, 2012.

[34] Oracle Corporation, "MySQL." [Online]. Available: http://www.mysql.com [accessed: 2016-02-22]

[35] Website 3 Installing SOP, "Reactome." [Online]. Available: http://wiki.reactome.org/index.php/Website_Installing_SOP [accessed: 2016-02-22]

 ISBN: 978-1-61208-488-6