BioGrakn: A Knowledge Graph-Based Semantic Database for Biomedical Sciences

Antonio Messina[1(✉)], Haikal Pribadi[2], Jo Stichbury[2], Michelangelo Bucci[2], Szymon Klarman[2], and Alfonso Urso[1]

[1] ICAR-CNR, via Ugo La Malfa 153, 90146 Palermo, Italy
{antonio.messina,alfonso.urso}@icar.cnr.it
[2] Grakn Labs Ltd, Unit 22, 8 Hornsey Street, London, UK
{haikal,jo,michelangelo,szymon}@grakn.ai

Abstract. The proliferation of biological research data generated and shared openly online is of huge benefit to the scientific community, but there are often significant challenges to overcome before it can be integrated from different sources and re-used to gain new knowledge. This paper introduces BioGrakn, which is a graph-based deductive database, combining the power of knowledge graphs and machine reasoning. BioGrakn illustrates how data can be aggregated and integrated, modelled in all its complexity and contextual specificity, and extended as needed. Built upon GRAKN.AI, it provides an integrated, intelligent database for researchers handling complex data.

Keywords: Knowledge representation and reasoning · Semantic web · Semantic data integration · Biomedical · Databases · Knowledge graphs

1 Introduction

Nowadays, the amount of biological data available online has proliferated, but this has been accompanied by enormous challenges arising from the need to integrate and connect related information from different sources [1].

Common problems include locating resources, differing data formats, ambiguity and duplication, relationships between data and the sheer volume and granularity of the information. As yet, there is no standard memorization and query format for this kind of data, so each resource usually requires a different approach to be properly handled.

Several classes of bio-molecular data, such as transcriptional regulatory networks and protein-protein interaction networks, interact as complex networks. They can usually be modeled as graphs, where nodes (and their attributes) model biological entities and edges contain relationships between these entities. Since query languages play a key role in the success of databases, in order to allow for efficient queries, these graphs can be stored either in relational or graph databases [2], where the latter by their nature seem to be a *natural* choice.

Examples of the adoption of graph databases in bioinformatics are given by ncRNA-DB [3], Bio4J [4], and BioGraphDB [5].

L. Barolli and O. Terzo (eds.), *Complex, Intelligent, and Software Intensive Systems*, Advances in Intelligent Systems and Computing 611, DOI 10.1007/978-3-319-61566-0_28

ncRNA-DB is a NoSQL database based on OrientDB [6] that combines many biological resources to deal with several classes of ncRNA such as miRNA, long-noncoding RNA (lncRNA), circular RNA (circRNA) and their interactions with genes and diseases.

Bio4j is based on a Java library and is an integrated cloud-based data platform, built upon a graph structure on top of Neo4J [7]. For now, it includes data about proteins, GO and enzymes.

Lastly, BioGraphDB integrates several types of data sources to perform bioinformatics analysis using a comprehensive system built on top of OrientDB. It includes data about genes, proteins, microRNAs, molecular pathways, functional annotations, and associations between microRNAs and cancer diseases.

No matter the chosen underlying architecture (relational or NoSQL graphs), every solution should also address the major issue of *semantic integrity*, that is, interpreting the real meaning of data derived from multiple sources or manipulated by various tools [8].

In the biological sciences, Semantic Web database technologies have seen significant adoption over the past decade, with some of the most fundamental and broadly known resources are being the EBI RDF platform [9], BioPortal [10], and Pathway Commons [11]. The uptake of these types of system has been summarized by Pasquier [12], who goes on to analyze the improvements needed before the Semantic Web is taken up by the majority of life science researchers.

Similarly, Livingston et al. describe the problems that persist in data integration, providing a case study of a knowledge base built on 18 large biomedical data sources [13]. KaBOB (the Knowledge Base of Biomedicine) is an integrated knowledge base of biomedical data and allows the underlying data to be queried in terms of biomedical concepts (e.g., genes and gene products, interactions and processes). KaBOB illustrates the concepts of shared identity and shared meaning across heterogeneous biomedical data sources.

Here, we introduce BioGrakn, based on GRAKN.AI [14], which is a deductive database in the form of a knowledge graph, allowing complex data modelling, verification, scaling, querying and analysis.

The database behind GRAKN.AI uses an ontology to facilitate the modelling of extremely complex datasets, functioning as a data schema constraint to guarantee information consistency. GRAKN.AI stores data in a way that allows machines to understand the meaning of information in the complete context of their relationships. Consequently, the semantic layer of Grakn allows computers to process complex information more intelligently, with less human intervention.

2 GRAKN.AI

GRAKN.AI is composed of two parts: Grakn (the storage), and Graql (a declarative query language).

2.1 Grakn

Grakn is built using several graph computing and distributed computing platforms, such as Apache TinkerPop and Apache Spark. Grakn is designed to be sharded and replicated over a network of distributed machines. The underlying data structure of Grakn is that of a labelled, directed hypergraph (Fig. 1).

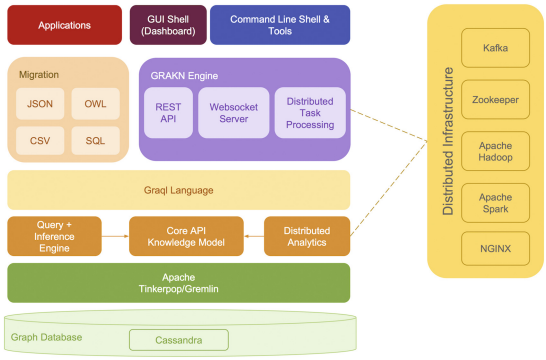

Fig. 1. The GRAKN.AI architecture

Grakn exposes a high-level knowledge model, allowing developers to represent their application domain as an ontology, specifying it in terms of entities, resources, relations, and roles. Grakn's ontology modelling constructs include, but are not limited to, data type hierarchy, relation type hierarchy, bi-directional relationships, multi-type relationships, N-ary relationships, relationships in relationships, and so on. Therefore, Grakn can model the real world and all the hierarchies and hyper-relationships contained within it.

2.2 Graql

Graql is a declarative, knowledge-oriented graph query language that uses machine reasoning to retrieve explicitly stored and implicitly derived knowledge from Grakn.

When using legacy systems, database queries have to define explicitly the data patterns they are looking for. Graql, on the other hand, will translate a query pattern into all its logical equivalents and evaluate them against the database. This includes, but is not limited to, the inference of types, relationships, context, and pattern combination. In this way, Graql can derive implicit information with concise and intuitive statements, reducing the complexity of expressing intelligent questions.

In Graql, there are different types of queries available: for matching patterns in the graph, inserting or deleting types and instances, and for computing useful information about the graph, such as statistics or shortest path between nodes.[1]

Two inference mechanisms are supported: *type inference*, based on the semantics defined in the ontology, and *rule-based inference*, that involves rules defined by expressions of the form *lhs G1 rhs G2*, where *G1* and *G2* are a pair of Graql patterns. Whenever the left-hand-side (*lhs*) pattern G1 is found in the data, the right-hand-side (*rhs*) pattern G2 can be assumed to exist and optionally materialized (inserted).

3 Data Sources

The data sources selected for database population are almost the same as those used by BioGraphDB. This way, we can build an integrated database containing resources related to genes, proteins, miRNAs, and metabolic pathways.

Getting into the details, we have considered the following:

(1) *NCBI Entrez Gene* [15]: provides a lot of genes data, such as interactions with other genes, genomic context, annotated pathways, and so on.
(2) *Gene Ontology (GO)* [16]: provides annotations for gene products in biological processes, cellular components and molecular functions.
(3) *UniProt Knowledgebase (UniprotKB)* [17]: the largest public collection of annotated functional information on proteins.
(4) *Reactome* [18]: contains validated metabolic pathways, each annotated as a set of biological events, dealing with genes and proteins.
(5) *miRBase* [19]: provides all the known miRNAs sequences and annotations, associated with names, keywords, genomic locations, and references.
(6) *mirCancer* [20]: contains associations between miRNAs and human cancers.
(7) *miRNASNP* [21]: aims to provide a resource of the miRNA-related mutations (SNPs) for human and other species.
(8) *mirTarBase* [22]: list of experimentally validated miRNA-target interactions.
(9) *miRanda* [23]: list of putative miRNA-target interactions.
(10) *HGNC* [24]: the HUGO Gene Nomenclature Committee database contains, for each gene symbol, a list of synonyms and a list of corresponding entries in the most popular genes databases.

Many of the above are supplied in *tab-separated values* (TSV) format, a simple text format for storing data in a tabular structure where each record in the table is one line of the text file, and each field value of a record is separated from the next by a tab character. By contrast, miRBase, GO, and UniprotKB are distributed as EMBL text file format [25] and XML format, respectively.

[1] Further information about syntax and keywords used by Graql can be found in https://grakn.ai/pages/documentation/graql/graql-overview.html.

4 BioGrakn

4.1 The Ontology

The ontology is a formal specification (in Graql) of all the relevant concepts and their meaningful associations in our domain. It must be clearly defined before loading data into the graph. This allows objects and relationships to be categorized into distinct types, enabling *automatic reasoning* over the represented knowledge, such as *inference* (extraction of implicit information from explicit data) and *validation* (discovery of inconsistencies in the data).

Grakn ontologies use four types of concepts for modeling domain knowledge. The categorization of concept types is enforced by declaring every concept type as a subtype of exactly one of the four corresponding built-in concept types: *entity*, *relation*, *role*, and *resource*.

Given the data sources considered in this work, our biological information has been associated to concepts, such as the ad-hoc defined subtypes shown in Table 1.

Table 1. Associations between Graql concepts, subtypes and biological information

Concept	Defined subtype	Biological information	Source
Entity	gene	genes	NCBI Entrez Genes
	go	functional annotations	Gene Ontology
	protein	proteins	UniProtKB
	pathway	pathways	Reactome
	mirna	miRNA precursors	miRBase
	mirnaMature	miRNA matures	miRBase
	mirnaSNP	miRNA SNPs	miRNASNP
	cancer	cancers	mirCancer
	proteinAccession	proteins accessions	UniProtKB
	geneName	genes symbols	HGNC
	interaction	miRNA-target interactions	mirTarBase, miRanda
Relation	annotation	links to annotated entities	Gene Ontology
	containing	links to entities contained in pathways	reactome
	precursorOf	precursors-matures relations	miRBase
	regulation	regulations of miRNAs in cancers	mirCancer
	snpMutation	miRNA-mutations relations	miRNASNP
	entity Reference	relations for entities synonyms	UniProtKB, HGNC
	encoding	genes-proteins coding	HGNC
	interactionMiRNA	miRNAs-interactions relations	mirTarBase, miRanda
	interaction—Gene	genes-interactions relations	mirTarBase, miRanda

4.2 Data Import

Two ways are supported for migrate data into a Grakn graph, the native migration capabilities and the *Loader Client API*. Both require the preliminary definition of an ontology for the data in Graql.

The former currently supports migration of CSV, JSON, OWL and SQL data. First, in order to map the data to the ontology, some Graql templates must be created. Then, it is possible to invoke the Grakn migration facilities through the shell or the migration API.

Even though most of the considered data sources are supplied in TSV format, a variant of CSV, their complexity and the extreme abundance of data and external references haven't allowed us to create related templates easily and quickly. Also, EMBL and XML source data files are not supported.

For this reason, we have developed an ad-hoc set of Extract-Transform-Load (ETL) tools. They have been written in Java and use the Loader Client API, in order to load large quantities of data into BioGrakn using multithreaded batch loading.

Data consistency and proper relations between entities are guaranteed by precise order of execution of the ETLs. This way, when a data source also refers to others, the presence in the database of all the depending resources is assured.

5 Results

In this section, we briefly introduce some illustrative queries and results representing typical bioinformatics problems, starting from the simplest.

5.1 Search for Genes Linked to a Particular Gene Ontology Annotation

Let's consider the Gene Ontology annotation "*platelet activating factor biosynthetic process*", that has *GO:0006663* as identifier. In order to find annotated genes, the *annotation* relation, with the functional annotation member equal to our starting identifier, points out all the related annotated entities, from which we extract the genes, printing their symbols and names. The following Graql query returns the desired results, shown in Fig. 2 in graph form:

```
match $go has goId "GO:0006663";
  (functionalAnnotation: $go; annotatedEntity: $gene) isa annotation;
  $gene isa gene;
```

5.2 Search for Pathways Linked to a Particular Gene

At a first sight, this seems like the previous problem. However, genes cannot be directly linked to pathways, because Reactome just provides pathway-to-proteins associations. Therefore, we have to go through two relations: *encoding*, that links genes to proteins, and *containing*, that links pathways to proteins. Thus, the Graql query is formed as follows (Fig. 3):

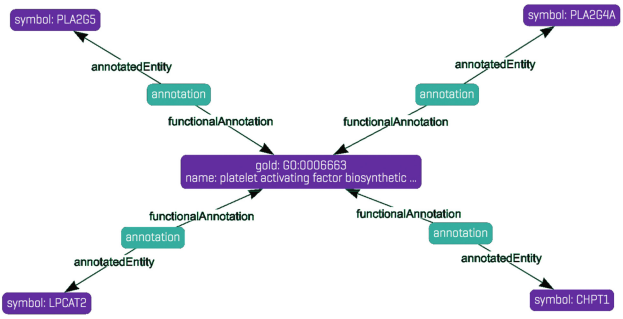

Fig. 2. Graphic results of a search for genes linked to GO annotation *GO:0006663*.

```
match $gene has symbol "LYPLA1";
    (encoder: $gene, encoded: $protein) isa encoding;
    (container: $path, contained: $protein) isa containing;
    $path isa pathway;
```

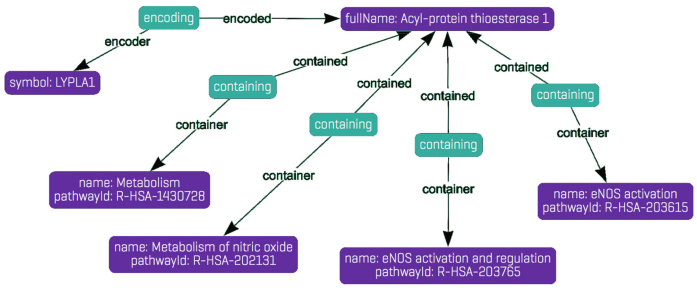

Fig. 3. Graphic results of a search for pathways linked to gene *LYPLA1*.

5.3 Differentially Expressed miRNAs Having SNPs in Cancer

Starting from a specific cancer, such as, for example, the *colorectal cancer*, we want to find all the up-regulated differentially expressed (DE) miRNAs that also have validated mutations. Because we are just interested in SNPs existence instead of their details, we can exclude them in the output, by selecting only entities of interest. Results for the following query are shown in Fig. 4:

```
match $cancer isa cancer has name "colorectal cancer";
    (upRegulator: $mirna, upRegulated: $cancer) isa upRegulation;
    (precursor: $mirna, mature: $mature) isa precursorOf;
    (mutated: $mature, snp: $snp) isa snpMutation;
    select $cancer, $mirna, $mature;
```

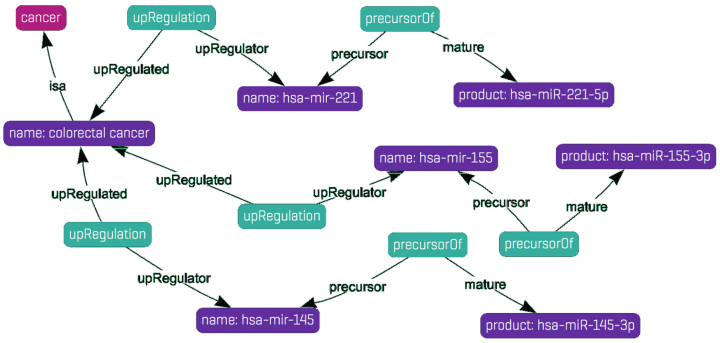

Fig. 4. Looking for DE miRNAs having SNPs for "*colorectal cancer*".

5.4 Reasoning on Our Biological Data

It is interesting to note how queries can be rearranged when using inference rules, especially the ones corresponding to typical query templates used in the domain of biological sciences.

For instance, considering the example in Subsect. 5.2, we have the following statements, that can be seen as a *set of premises*:

```
if genes codify proteins
if proteins belong to pathways
```

Thus, it is possible to infer the following *fact*:

```
then genes can be linked to pathways
```

Therefore, we can write an inference rule that infers genes-pathways links:

```
$genesInPathways isa inference-rule
   lhs {
      $gene isa gene; $protein isa protein;
      (encoder: $gene, encoded: $protein) isa encoding;
      (container: $pathway, contained: $protein) isa containing;
   }
   rhs {
      (container: $pathway, contained: $gene) isa containing;
   }
```

This rule allows us to rewrite the query reported in Subsect. 5.2 this way:

```
match $gene has symbol "LYPLA1";
   (container: $pathway, contained: $gene) isa containing;
```

As expected, the graphic results now show direct links from gene to pathways (Fig. 5).

Similarly, we can heavily rewrite the query in Subsect. 5.4 thanks to an inference rule like this:

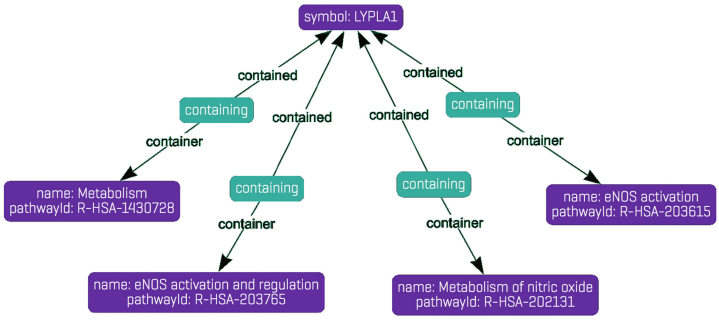

Fig. 5. Graphic results of reasoning on gene-pathways links.

```
$matureWithSNPsByCancer isa inference-rule
 lhs {
  $cancer isa cancer; $mirna isa mirna; $mature isa mirnaMature;
  (upRegulator: $mirna, upRegulated: $cancer) isa upRegulation;
  (precursor: $mirna, mature: $mature) isa precursorOf;
  (mutated: $mature, snp: $snp) isa snpMutation;
 }
 rhs {
  (byCancer: $cancer, mirnaWithSNP: $mature) isa matureWithSNPsByCancer;
 }
```

The rewritten query and its results are shown below (Fig. 6).

```
match $cancer isa cancer has name "colorectal cancer";
(byCancer: $cancer, mirnaWithSNP: $mature) isa matureWithSNPsByCancer;
```

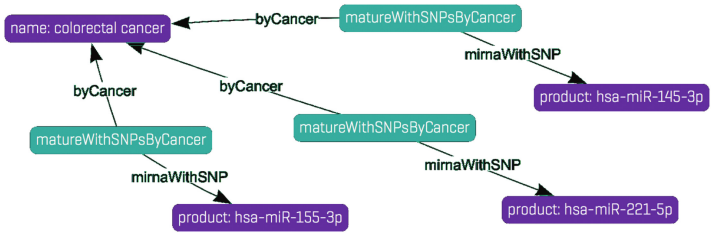

Fig. 6. Graphic results of reasoning on cancers and miRNAs with SNPs.

6 Conclusions and Future Works

In this paper, we propose BioGrakn, a graph-based semantic database that takes advantage of the power of knowledge graphs and machine reasoning, to solve problems in the domain of biomedical science. The database has been designed to overcome problems related to the lack of a structural organization and interoperability of publicly available biological resources, ensuring the semantic integrity of data by design.

BioGrakn has been built on top of GRAKN.AI, a distributed knowledge graph database which allows complex data modeling, verification, scaling, querying and analysis. A key step is the definition of an ontology, which facilitates the modeling of complex datasets and guarantees information consistency.

Inference rules allow the extraction of implicit information from explicit data, to achieve logical reasoning over the represented knowledge.

In the short term, further developments are expected, such as the integration of other publicly available biological resources, the use of the native GRAKN.AI migration tools for data migration procedures, and the deployment of an user-friendly web interface.

References

1. Cheung, K.H., Smith, A.K., Yip, K.Y., Baker, C.J., Gerstein, M.B.: Semantic web approach to database integration in the life sciences. In: Semantic Web. Springer, US (2007)
2. Have, C.T., Jensen, L.J.: Are graph databases ready for bioinformatics? Bioinformatics **29**(24), 3107–3108 (2013)
3. Bonnici, V., Russo, F., Bombieri, N., Pulvirenti, A., Giugno, R.: Comprehensive reconstruction and visualization of non-coding regulatory networks in human. Front. Bioeng. Biotechnol. **2**, 69 (2014). doi:10.3389/fbioe.2014.00069
4. Pareja-Tobes, P., Tobes, R., Manrique, M., Pareja, E., Pareja-Tobes, E.: Bio4j: a high-performance cloud-enabled graph-based data platform. Era7 bioinformatics, Technical Report (2015)
5. Fiannaca, A., La Rosa, M., La Paglia, L., Messina, A., Urso, A.: BioGraphDB: a new GraphDB collecting heterogeneous data for bioinformatics analysis. In: BIOTECHNO 2016: The Eighth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies, Lisbon, Portugal, pp. 28–34 (2016)
6. Orient Technologies Ltd: OrientDB Community Edition. http://orientdb.com
7. Neo Technology Inc: Neo4J https://neo4j.com
8. Pettifer, S., Thorne, D., McDermott, P., Marsh, J., Villéger, A., Kell, D.B., Attwood, T.K.: Visualising biological data: a semantic approach to tool and database integration. BMC Bioinform. **10**(6), S19 (2009)
9. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., Jenkinson, A.M.: The EBI RDF platform: linked open data for the life sciences. Bioinformatics **30**, 1338–1339 (2014). (Oxford, England)
10. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A.: BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Res. **39**(Web Server issue), W541–W545 (2011)
11. Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., Sander, C.: Pathway commons, a web resource for biological pathway data. Nucleic Acids Res. **39**(Database issue), D685–D690 (2011)
12. Pasquier, C.: Biological data integration using semantic web technologies. Biochimie **90**(4), 584–594 (2008)
13. Livingston, K.M., Bada, M., Baumgartner, W.A., Hunter, L.E.: KaBOB: ontology-based semantic integration of biomedical databases. BMC Bioinform. **16**(1), 126 (2015)

14. Grakn Labs Ltd: GRAKN.AI. https://grakn.ai
15. Schuler, G.D., Epstein, J.A., Ohkawa, H., Kans, J.A.: Entrez: molecular biology database and retrieval system. Methods Enzymol. **266**, 141–162 (1996)
16. The gene ontology consortium: gene ontology consortium: going forward. Nucleic Acids Res. **43**(D1), 1049–1056 (2015)
17. The UniProt: a hub for protein information. Nucleic Acids Res. **43**(D1), 204–212 (2015)
18. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., D'Eustachio, P.: The reactome pathway knowledgebase. Nucleic Acids Res. **42**(D1), 472–477 (2014)
19. Kozomara, A., Griffiths-Jones, S.: miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. **39**, 152–157 (2011). Database issue
20. Xie, B., Ding, Q., Han, H., Wu, D.: miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics **29**(5), 638–644 (2013)
21. Gong, J., Tong, Y., Zhang, H.M., Wang, K., Hu, T., Shan, G., Sun, J., Guo, A.Y.: Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Hum. Mutat. **33**(1), 254–263 (2012)
22. Hsu, S.-D., Tseng, Y.-T., Shrestha, S., Lin, Y.-L., Khaleel, A., Chou, C.-H., Chu, C.-F., Huang, H.-Y., Lin, C.-M., Ho, S.-Y., Jian, T.-Y., Lin, F.-M., Chang, T.-H., Weng, S.-L., Liao, K.-W., Liao, I.-E., Liu, C.-C., Huang, H.-D.: miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. **42**(D1), 78–85 (2014)
23. John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., Marks, D.S.: Human microRNA targets. PLoS Biol. **2**(11), e363 (2004)
24. Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., Bruford, E.A.: Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. **43**(D1), 1079–1085 (2015)
25. Baker, W., Van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G., Tuli, M.A.: The EMBL nucleotide sequence database. Nucleic Acids Res. **28**(1), 19–23 (2000)