Towards the Construction of an RNA-centered Knowledge Graph

Emanuele Cavalleri[1], Sara Bonfitto[1], Alberto Cabri[1], Jessica Gliozzo[1], Paolo Perlasca[1], Mauricio Soto-Gomez[1], Gabriella Trucco[1], Elena Casiraghi[1], Giorgio Valentini[1] and Marco Mesiti[1]

[1]*Dep. of Computer Science, Università di Milano, Via Celoria 18, 20133 Milano*

Abstract

The use of RNA molecules for developing new drugs and new vaccines is attracting more and more scientific centers all over the world that produce biological banks with different kinds of relationships existing among the different coding and non-coding molecules. Collecting and identifying relationships among the data included in these collections is of paramount importance for knowledge discovery and analysis. In this paper, we describe the initial steps in the construction of *RNA-KG*, an RNA-centered Knowledge Graph that will contain the different types of entities that can be extracted from different public databases and the relationships that can be inferred. A meta-graph reporting the main kinds of relationships that can be included by the integration of the identified data sources is finally presented. These activities are conducted in the context of the "National Center for Gene Therapy and Drugs based on RNA Technology" funded by the Italian PNRR and the NextGenerationEU program.

Keywords

RNA-based technologies, Knowledge Graphs, RNA-drug discovery

1. Introduction

RNA-based drugs represent one of the most promising advances in therapeutics, as evidenced by the recent success of mRNA-based vaccines for the COVID-19 pandemic [1]. More generally, coding and non-coding RNA molecules can potentially lead to new treatments of cancer, genetic and neurodegenerative disorders, cardiovascular and infectious diseases [2].

Conventional drugs show relevant limitations in their druggable targets because they usually consist of small molecules targeting proteins. Only about 10% of proteins have druggable binding sites and no more than 2% of the human genome is protein-coding. On the contrary, RNA drugs can target both proteins and mRNA, as well as other non-coding RNA (ncRNA). Moreover, they can encode missing or defective proteins, regulate the transcriptome, and mediate DNA or RNA

SEBD 2023: 31st Symposium on Advanced Database Systems, July 02–05, 2023, Galzignano Terme, Padua, Italy

emanuele.cavalleri@unimi.it (E. Cavalleri); sara.bonfitto@unimi.it (S. Bonfitto); alberto.cabri@unimi.it (A. Cabri); jessica.gliozzo@unimi.it (J. Gliozzo); paolo.perlasca@unimi.it (P. Perlasca); mauricio.soto@unimi.it (M. Soto-Gomez); gabriella.trucco@unimi.it (G. Trucco); elena.casiraghi@unimi.it (E. Casiraghi); giorgio.valentini@unimi.it (G. Valentini); marco.mesiti@unimi.it (M. Mesiti)

0000-0003-1973-5712 (E. Cavalleri); 0000-0002-9883-5561 (S. Bonfitto); 0000-0003-1373-8402 (A. Cabri); 0000-0001-7629-8112 (J. Gliozzo); 0000-0001-6674-2822 (P. Perlasca); 0000-0001-5977-9467 (M. Soto-Gomez); 0000-0002-5904-0969 (G. Trucco); 0000-0003-2024-7572 (E. Casiraghi); 0000-0002-5694-3919 (G. Valentini); 0000-0001-5701-0080 (M. Mesiti)

editing. Thus, RNA technology significantly broadens the set of druggable targets and is also less expensive than other technologies (e.g. drug synthesis based on recombinant proteins), due to the relatively simple structure of RNA molecules that facilitates their biochemical synthesis and chemical modifications [3].

In the framework of the NextGenerationEU funded "National Center for Gene Therapy and Drugs based on RNA Technology", we aim to support the discovery of novel RNA-based drugs by developing an RNA-centered Knowledge Graph (RNA-KG).[1] RNA-KG will collect and organize data and knowledge about RNA molecules, retrieved from public databases and/or generated from the results of the research groups involved in the National Center. It will also provide a comprehensive description of the relationships among the various kind of RNAs, diseases, drugs, phenotypes, and other bio-medical entities. RNA-KG will be the basis for the development of novel cutting-edge AI methods specifically tailored for the analysis of biological processes involving RNA. These methods could also open the way to RNA-drug prioritization, RNA drug-target prediction, and other prediction tasks for discovering new RNA drugs.

In this paper, we report our initial achievements that have been recently published in [4] for the identification of a meta-graph representing the kind of relationships that can be identified among the different types of RNA molecules. This result is obtained by examining more than 50 public online repositories for non-coding RNA sequences and annotations and by studying the kinds of interactions that can exist among these molecules. The public online repositories have been selected through an extensive literature review of top journals of the sector (like NAR, BMC Bioninformatics, Science, RNA Journal, IEEE/ACM TCBB), that are periodically updated by their developers, and contain significant amounts of molecules and relationships. In the identification of the repositories, we have taken into account the presence of controlled vocabularies, thesaurus, ontologies that formally describes the repository content, and the presence of well-recognized identification schemes.

The paper is organized as follows. Section 2 introduces related work devoted to data integration and to the construction of RNA-KG starting from different heterogeneous databases. Moreover, it introduces biomedical ontologies that can be used in this context for the characterization of RNA-molecules and their interaction. Section 3 highlights the characteristics of the identified databases. Section 4 highlights the main types of relationships that can be extracted from the sources, introduces a meta-graph that shows the potential relationships that will be available among the RNA molecules, and describe the characteristics of an initial instantiation of the knowledge graph. Finally, Section 5 reports our concluding remarks.

2. Related Work

Data Integration Approaches. The data integration issue is a well-known problem in the area of data management, and many approaches have been devised to deal with relational data [5]. However, the explosion of data formats (like CSV, JSON, XML) and the variability in the representation of the same types of information [6] has pushed the need to exploit ontologies as global common models both for accessing (OBDA – Ontology-Based Data Access) and integrating (OBDI – Ontology-Based Data Integration) data sources [7, 8]. In OBDA, queries are

[1] Available at `https://github.com/AnacletoLAB/RNA-KG`

expressed in terms of an ontology, and the mappings between the ontology and the data sources' schema are described in the form of declarative mapping rules. Two approaches are usually proposed to enable access and integration to different data sources: *materialization*, where data are converted from the local schema according to the ontology concepts and relationships (i.e. data are converted into an RDF KG and locally stored in a data-warehouse of triples that can be queried by means of SPARQL); *virtualization*, where the transformation is executed on the fly during the evaluation of queries by exploiting the mapping rules and the ontology. In this case, only the data from the original sources involved in the query are accessed for generating the query result in accordance with the adopted ontology. Materialization can provide fast and accurate access to data because already organized in a centralized repository. However, data freshness can be compromised when data sources frequently change. On the other hand, virtualization allows access to fresh data but requires the application of transformations during query evaluation and can cause delay. Different approaches support the specification of mapping rules like R2RML [9] (a W3C standard for relational sources), and RML [10] which extends the standard for dealing with other formats. Moreover, SPARQL-Generate [11], YARRRML [12] and ShExML [13] were also proposed for dealing with data heterogeneity.
KG construction from bio-medical data sources. In the biological context, many efforts are nowadays devoted to the construction of KGs by integrating different public sources that exploit the materialization and virtualization approaches previously described. An approach for integrating different biological data into a biological KG was proposed in [14]. The approach designs a Connecting Ontology $\mathcal{CO}$ to integrate all the external ontologies describing the involved data sources. By exploiting algorithms for fusing and integrating annotations, an enriched KG is obtained that spans multiple data sources and is annotated by the integrated biological ontology. The effectiveness of this approach is shown by integrating `rice gene-phenotype` and `lactobacillus` data sources by gluing together the GO, Trait, Disease, and Plant Ontologies. In [15], the Precision Medicine KG (named PrimeKG) was developed to represent holistic and multimodal views of diseases. PrimeKG integrates more than 20 high-quality resources with more than 4M relations that capture information like disease-associated perturbations in the proteome, biological processes, and molecular pathways. The considered data were collected and annotated using diverse ontologies such as Disease Gene Network (DisGeNet), Mayo Clinical knowledgebase, Mondo Diseases Ontology, Bgee, and DrugBank. A virtualization approach based on an ontology-based federation of three data sources (Bgee, OMA, and UNIProtKB) was presented in [16]. Starting from a semantic model for gene expression, the authors propose using mapping rules for dealing with the different formats of the three sources and allowing the issue of joint queries across the sources by exploiting SPARQL endpoints. PheKnowLator [17] (Phenotype Knowledge Translator) is a fully automated Python 3 library for the construction of semantically rich, large-scale biomedical KGs that are Semantic Web compliant and amenable to automatic OWL reasoning, conform to contemporary property graph standards. The library offers tools to download data, transform and/or pre-processing of resources into edge lists, construct knowledge graphs, and generate a wide range of outputs. All these papers point out the difficulties that arise when trying to integrate different data sources that exploit different data models, formats, and ontologies. Specifically, data redundancies, data duplicates, and lack of common identifier mechanisms must be properly addressed.

Name	Abbr.	Description
Gene Ontology	GO	GO provides the terms representing gene product properties. GO covers three domains: cellular component, molecular function, and biological process.
Disease Ontology	DO	DO provides the terms representing human diseases.
Chemical Entities of Biological Interest	ChEBI	ChEBI provides the terms representing molecular entities of 'small' chemical compounds.
Non-Coding RNA Ontology	NCRO	NCRO provides the terms representing non-coding RNA molecules both of biological origin, and engineered.
Ontology for Biomedical Investigations	OBI	OBI provides the terms representing biological and clinical investigations.
Single-Nucleotide Polymorphism Ontology	SNPO	SNPO provides the terms representing formal and unambiguous representation of genomic variations.
EMBRACE Data And Methods	EDAM	EDAM provides the terms representing concepts that are prevalent within bioscientific data analysis, data management in life sciences.
Sequence Ontology	SO	SO provides the terms representing features used in biological sequence annotation.
BRENDA Tissue Ontology	BTO	BTO provides the terms representing the source of an enzyme comprising tissues, cell lines, cell types and cell cultures.
Experimental Factor Ontology	EFO	EFO provides the terms representing experimental variables. It combines parts of several biological ontologies, e.g. UBERON anatomy, ChEBI, and Cell Ontology.
Medical Subject Headings	MeSH	MeSH provides the terms used for indexing PubMed citations.

Table 1
Main Bio-Ontologies involved in RNA relationships.

Biomedical Ontologies. Several standard ontologies can be used for the characterization of RNA molecules and their interaction with other biomedical entities (see Table 1). Moreover, data formats specifically developed for biological pathways (like Panther, Reactome or Wikipathways) are used for semantically annotating the RNA molecules.

In general, a well-recognized and globally accepted ontology for the representation of any kind of ncRNA molecules is still lacking. Often for referring to ncRNA molecules, the name of the gene encoding the physically closest protein is used. Moreover, ncRNA genes with no known function are named pragmatically based on their genomic context; if there is a proximal (genomically adjacent close in physical proximity) protein coding gene (PCG) then the ncRNA genes are given a gene symbol beginning with the PCG symbol [18]. The identification scheme used for miRNA (which are the majority of data sources) is always borrowed from miRBase. This makes all the other data sources associated with miRNAs, miRBase "compliant". Furthermore, the identification scheme associated with miRNAs is partially included in NCRO, which includes miRNA transcripts from Homo Sapiens cells [19].

3. RNA-based data sources

The wide variety of RNA molecules are translated into proteins, regulate gene expression, hold enzymatic activity, and modify other RNAs. Coding RNA molecules are named messenger RNA (mRNA) molecules, translated into proteins helped by ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), and small nucleolar RNA (snoRNA) molecules. Non-coding RNA molecules having less than 200 nucleotides are named small non-coding RNA (snRNA). This category includes a wide variety of RNA molecules, such as microRNA (miRNA), short interfering RNA (siRNA), short hairpin RNA (shRNA), antisense oligonucleotides (ASO), piwi-interacting RNA (piRNA), transfer RNA fragments (tRF), guide RNA (gRNA), aptamer, riboswitch,

Type	# DBs	# molecules	# rel	Relation with
miRNA	19	58k	160M	miRNA, mRNA, lncRNA, circRNA, tRF, snoRNA, pseudogene, protein
s(h/i)RNA	1	147	147	mRNA
Aptamer	1	8k	7.8k	protein
ASO	2	2k	12k	mRNA, protein
lncRNA	9	650k	180M	mRNA, protein, miRNA, snoRNA
gRNA	1	29	3.2k	gene
Ribozyme	1	1k	17k	gene, viral RNA
Viral RNA	1	10k	17k	ribozyme
Riboswitch	1	25k	24k	protein
tRF	3	30k	215k	miRNA, tRNA
snoRNA	1	1k	2k	gene, lncRNA, miRNA, mRNA, pseudogene, rRNA, snoRNA, snRNA, tRNA
tRNA	1	10k	180k	tRF, amino acid

Table 2
For each type of RNA molecule, we report the number of available data sources, the number of molecules that can be identified, the number of relationships that can be extracted from the sources with other RNA molecules, and the molecules with which a relationship can be identified.

Relation ID	Name	Abbreviation
RO:0002429	involved in positive regulation of	`regulates+`
RO:0002430	involved in negative regulation of	`regulates-`
RO:0002434	interacts with	`interacts with`
RO:0002436	molecularly interacts with	`m. interacts with`
RO:0010002	is carrier of	`carries`
RO:0002204	gene product of	`gene product of`
RO:0002526	overlaps sequence of	`overlaps`
RO:0002528	is upstream of sequence of	`is upstream`
RO:0002529	is downstream of sequence of	`is downstream`
RO:0002202	develops from	`develops from`
RO:0002203	develops into	`develops into`

Table 3
Main relations among bio-entities involving RNA with the RO identifier.

and ribozyme molecules. Non-coding RNA molecules with more than 200 nucleotides are named long non-coding RNA (lncRNA). Circular RNA (circRNA) are lncRNA molecules produced from alternative splicing events. Further details on the role and meaning of these molecules can be found in [4].

Table 2 provides an overview of the identified databases organized according to the main molecule that they make available. Specifically, for each kind of molecule, the table reports the number of available databases, the number of molecules, the number of relationships that can be extracted, and the list of molecules with which relationships can be established. Details on the databases can be found in [4] with the bibliographic references.

Besides the sequences, these data sources also contain different kinds of relationships that can be represented according to the Relation Ontology (RO) [20]. Table 3 reports the main identified relationships. For each relation, Table 3 reports the RO identifier, the corresponding

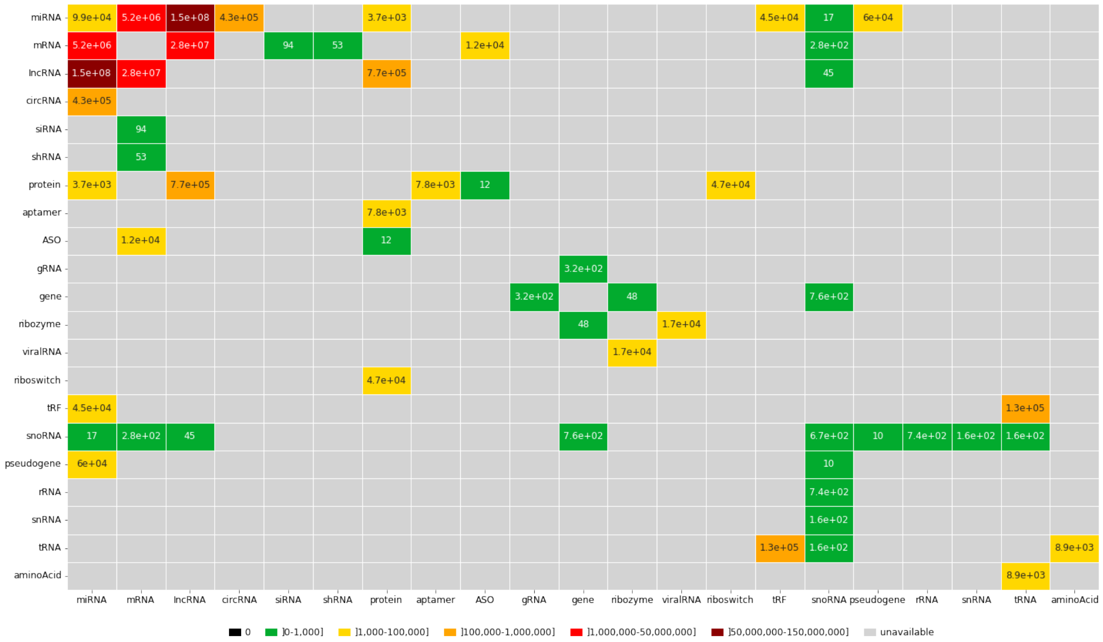

Figure 1: Available relations involving RNA molecules.

meaning, and an abbreviated form used in our paper. The general relationships “interacts with" available in RO with the meaning “A relationship that holds between two entities in which the processes executed by the two entities are causally connected" has been declined in the most specific relationships “molecularly interacts with" in our classification to represent the situation in which the two partners are molecular entities that directly physically interact with each other (e.g. via a stable binding interaction or a brief interaction during which one modifies the other). We use this relationship when we wish to represent a specific interaction process at the molecular level (e.g. complementary base pairing occurring in RNAi in miRNA-mRNA interaction or tRNA molecule charged with a specific amino acid). Figure 1 summarizes the relationships among RNA molecules that we have identified in the different data sources. More details can be found in [4].

4. A meta-graph for modelling RNA-centered relationships

Starting from the analysis of the data sources, the meta-graph in Fig. 2 has been realized. Colored edges represent uni-direction relationships (e.g. tRF regulates miRNA). The graphical representation provides a global overview of the richness of information that is currently provided. Moreover, the meta-graph points out the presence of a central hub, named “GENE/mRNA", that is bound to many kinds of ncRNA. This characteristic might have a deep impact on the discovery of new unconsidered interactions among ncRNA molecules. To simplify the visualization of the meta-graph, we omitted most of the non-RNA biomolecular and medical entities that are known

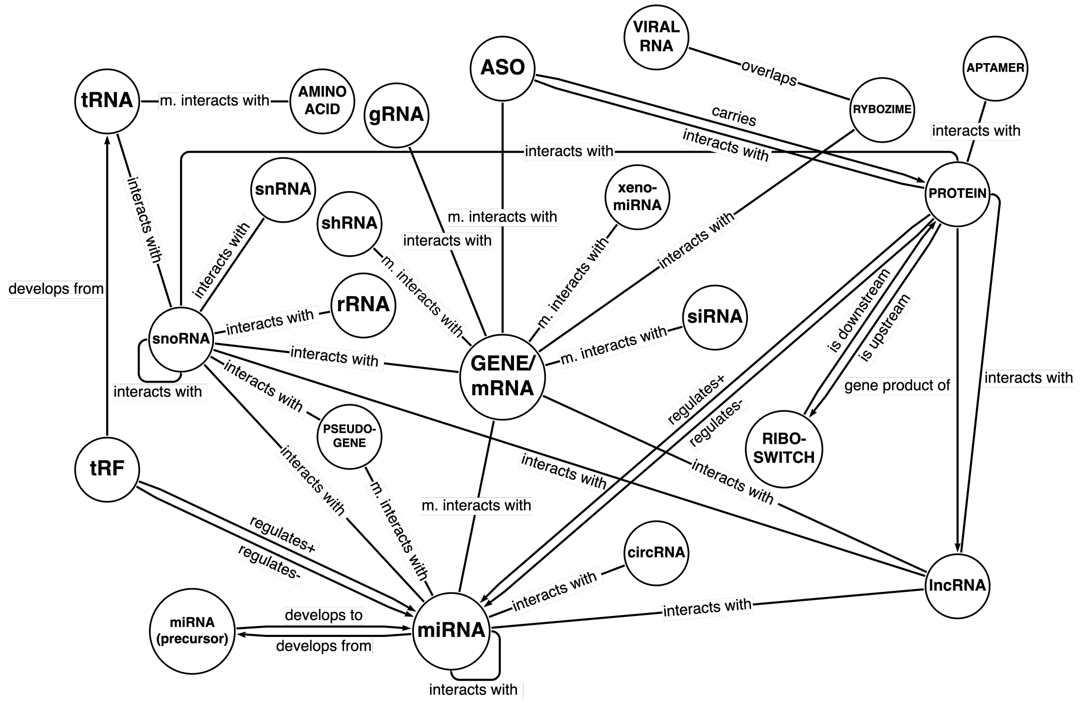

Figure 2: RNA-centered meta-graph.

to play an important role to study the biology and support the discovery of novel RNA drugs. Indeed the meta-graph in Fig. 2 can be further extended with other nodes representing other biological entities (e.g. diseases, epigenetic modifications, small molecules, tissues, biological pathways, cellular compartments) and relationships relevant to the analysis of RNA-KG.

5. Conclusion and Future Work

This paper reports the initial results of an ongoing project for the creation of a biomedical knowledge graph for the representation of non-coding RNA molecules and their relationships made available in different publicly available data sources.

The first release of RNA-KG can be accessed through a SPARQL endpointfor which we used an AllegroGraph triplestore that offers a graphical user interface for performing queries. The code used for the integration of the different sources is available on our GitHub repository. The knowledge graph has been realized by exploiting the primitives made available in PheKnowLator because they are effective and well-documented.

We are currently working on further integrating specific databases on RNA. Moreover, PheknowLator provides 12 Open Biological and Biomedical Foundry Ontologies and 31 publicly available resources that can be integrated with our ongoing RNA-KG. The resulting RNA-KG will be analyzed with cutting-edge AI graph representation learning algorithms [21], developed in the context of the National Center for Gene Therapy and Drugs based on RNA Technology, to support the discovery of novel RNA drugs. Finally, we would like to develop graphical facilities for supporting the user in the data acquisition process and thus reducing the manual effort required for mapping the data available in the different data sources into RNA-KG [22].

References

[1] A. Barbier, et al. The clinical progress of mRNA vaccines and immunotherapies, Nature Biotechnology 40 (2022) 840–865.

[2] T. R. Damase, et al. The limitless future of RNA therapeutics, Frontiers in Bioengineering and Biotechnology 9 (2021). doi:10.3389/fbioe.2021.628137.

[3] K. Paunovska, et al. Drug delivery systems for RNA therapeutics., Nat Rev Genet 23 (2022) 265–280.

[4] E. Cavalleri, et al. A meta-graph for the construction of RNA-KG, in: 10th Int'l Work-Conference on Bioinformatics and Biomedical Engineering, 2023. To appear.

[5] A. Halevy, Information Integration, Springer, 2009. doi:10.1007/978-0-387-39940-9_1069.

[6] M. Mesiti, et al. XML-based approaches for the integration of heterogeneous bio-molecular data, BMC Bioinformatics 10 (2009). doi:10.1186/1471-2105-10-S12-S7.

[7] A. Poggi, et al. Linking data to ontologies, in: J. on Data Semantics X, Springer, 2008.

[8] D. Calvanese, et al. Accessing scientific data through knowledge graphs with Ontop, in: Patterns, CellPress, 2021. doi:10.1016/j.patter.2021.100346.

[9] S. Das, et al. R2RML: RDB to RDF mapping language, www.w3.org/TR/r2rml/, 2012.

[10] A. Dimou, et al. RML: a generic language for integrated RDF mappings of heterogeneous data, in: Proc. of the 7th Workshop on Linked Data on the Web, volume 1184 of *CEUR Workshop Proc.*, 2014.

[11] M. Lefrançois, et al. A SPARQL extension for generating RDF from heterogeneous formats, in: The Semantic Web, Springer, 2017, pp. 35–50.

[12] P. Heyvaert, et al. Declarative rules for linked data generation at your fingertips!, in: The Semantic Web: ESWC 2018 Satellite Events, Springer, 2018, pp. 213–217.

[13] H. García-González, et al. ShExML: improving the usability of heterogeneous data mapping languages for first-time users, PeerJ Computer Science 6 (2020) 27. doi:10.7717/peerj-cs.318.

[14] S. Zhang, Y. Tang, et al. A graph-based approach for integrating biological heterogeneous data based on connecting ontology, in: IEEE Int'l Conf. on Bioinformatics and Biomedicine, 2021, pp. 600–607. doi:10.1109/BIBM52615.2021.9669700.

[15] P. Chandak, et al. Building a knowledge graph to enable precision medicine, Sci Data 10 (2023) 67. doi:10.1038/s41597-023-01960-3.

[16] A. C. Sima, et al. Enabling semantic queries across federated bioinformatics databases, Database 2019 (2019). doi:10.1093/database/baz106.

[17] T. J. Callahan, et al. A framework for automated construction of heterogeneous large-scale biomedical knowledge graphs, bioRxiv (2020). doi:10.1101/2020.04.30.071407.

[18] M. W. Wright, A short guide to long non-coding rna gene nomenclature, Human Genomics 8 (2014) 7. doi:10.1186/1479-7364-8-7.

[19] J. Huang, et al. The non-coding RNA ontology (NCRO): a comprehensive resource for the unification of non-coding RNA biology, J. of Biomedical Semantics 7 (2016) 24. doi:10.1186/s13326-016-0066-0.

[20] E. Ong, et al. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration, Nucleic Acids Res. 45 (2016) D347–D352. doi:10.1093/nar/gkw918.

[21] Xia, F. et al.: Graph Learning: A Survey. IEEE Transactions on Artificial Intelligence 2(2) (2021) 109–127.

[22] S. Bonfitto, et al. Easy-to-use interfaces for supporting the semantic annotation of web tables, in: Int'l Workshop on Data Platforms Design, Management, and Optimization, 2023.