A semantic approach for knowledge capture of microRNA-target gene interactions

Jingshan Huang*
School of Computing, University of South Alabama
Mobile, Alabama 36688, U.S.A.

Fernando Gutierrez
CIS Dept., University of Oregon
Eugene, Oregon 97403, U.S.A.

Dejing Dou
CIS Dept., University of Oregon
Eugene, Oregon 97403, U.S.A.

Judith A. Blake
The Jackson Laboratory
Bar Harbor, Maine 04609, U.S.A.

Karen Eilbeck
School of Medicine, University of Utah
Salt Lake City, Utah 84112, U.S.A.

Darren A. Natale
Georgetown University Medical Center
Washington D.C. 20007, U.S.A.

Barry Smith
Dept. Philosophy, University at Buffalo
Buffalo, New York 14260, U.S.A.

Yu Lin
University of Miami
Miami, Florida 33146, U.S.A.

Xiaowei Wang
Washington University School of Medicine
St. Louis, Missouri 63110, U.S.A.

Zixing Liu
MCI, University of South Alabama
Mobile, Alabama 36604, U.S.A.

Ming Tan
MCI, University of South Alabama
Mobile, Alabama 36604, U.S.A.

Alan Ruttenberg
Dental Medicine, University at Buffalo
Buffalo, New York 14214, U.S.A.

***Abstract*—Research has indicated that microRNAs (miRNAs), a special class of non-coding RNAs (ncRNAs), can perform important roles in different biological and pathological processes. miRNAs' functions are realized by regulating their respective target genes (targets). It is thus critical to identify and analyze miRNA-target interactions for a better understanding and delineation of miRNAs' functions. However, conventional knowledge discovery and acquisition methods have many limitations. Fortunately, semantic technologies that are based on domain ontologies can render great assistance in this regard. In our previous investigations, we developed a miRNA domain-specific application ontology, Ontology for MIcroRNA Target (OMIT), to provide the community with common data elements and data exchange standards in the miRNA research. This paper describes (1) our continuing efforts in the OMIT ontology development and (2) the application of the OMIT to enable a semantic approach for knowledge capture of miRNA-target interactions.**

Keywords*—microRNA, non-coding RNA, target gene, biomedical ontology, ontology development, data annotation, data integration, semantic search, SPARQL query.

I. Introduction

In biological, biomedical, and clinical investigation, microRNAs (miRNAs) are considered as important non-coding RNAs (ncRNAs). Prior research [1] [2] has indicated that miRNAs are able to perform significant roles in both biological and pathological processes, thus affecting the control and regulation of various human diseases. The mechanism by which miRNAs realize their critical functions is through some special binding to respective target genes (short for targets). Therefore, the ability to effectively identify and analyze different miRNA-target interactions has become a key step to completely understand and fully delineate miRNAs' functions.

Conventionally, data end users (that is, biologists, bioinformaticians, and clinical investigators) need to search for (1) biologically validated miRNA targets (for example, by querying the PubMed database [3]) and (2) computationally putative miRNA targets (for example, by initiating inquiries on various prediction databases or websites such as miRDB [4]). Not only manual searches are necessary among all involved data sources, but also more importantly, these data sources are semantically heterogeneous among each other — in other words, the meanings of data from different sources are usually quite different from each other and thus in many cases confusing to end users. Therefore, it has been extremely challenging for users to identify and establish possible links among original data sources. As a result, there exist significant barriers during conventional miRNA knowledge discovery and acquisition, which is time-consuming, labor-intensive, error-prone, and subject to end users' limited prior knowledge. In addition, the situation can be far worse: more often than not, it is also necessary to obtain additional information for each and every miRNA target, either validated or putative, from relevant data sources such as NCBI Gene [5] and NCBI Nucleotide [6]. Likewise, these additional data sources are also highly heterogeneous with each other.

Emerging semantic technologies are believed to be able to significantly assist with handling the aforementioned challenge in the miRNA knowledge acquisition. The core of the current semantic technologies include formal specifications such as the Resource Description Framework (RDF), RDF Schema (RDFS), and Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships, as well as to enable automatic reasoning (inference) within a given domain. One way to apply semantic technologies in miRNA knowledge acquisition is to transform data obtained from miRNA-related databases into RDF by

*Corresponding author (Email: huang@southalabama.edu)

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

annotating original data with formally defined ontologies. After such data annotation we can then use SPARQL Protocol and RDF Query Language (SPARQL) [7] to issue a search query based on the RDF model.

In our previous research [8–13] we investigated the construction of an application ontology for the miRNA field, named Ontology for MIcroRNA Target (OMIT), the first of its kind that formally encodes miRNA domain knowledge. By providing a standardized metadata model to help establish miRNA data connections among heterogeneous sources, OMIT was meant to fill the gap of lacking common data elements and data exchange standards for the miRNA research, especially with regard to miRNA-target interactions.

There are two major scientific contributions in this paper: (1) our continuing efforts and significant improvements on the OMIT ontology development and (2) the application of the OMIT to enable a semantic approach for knowledge capture of miRNA-target interactions, leading to more effective miRNA data integration and knowledge discovery.

The rest of this paper is organized as follows. Section II summarizes state-of-the-art research in biomedical ontologies and semantic mapping & search, respectively; Section III reports our efforts on reconstructing the OMIT ontology; Section IV describes a set of software packages to realize miRNA semantic annotation, data integration, and semantic search; Section V reports our experimental results along with discussion; finally, Section VI concludes with some future research work.

II. Related Work

A. Related work in biomedical ontologies

Ontologies have been widely utilized in biological, biomedical, and clinical research. We briefly describe some representative bio-ontologies included in both the Open Biological and Biomedical Ontologies (OBO) Library [14] and the National Center for Biomedical Ontology (NCBO) BioPortal [15].

Gene Ontology (GO) [16]: GO is by far the most successful and widely used bio-ontology, consisting of three independent sub-ontologies: biological processes, molecular functions, and cellular components. The GO has been utilized to annotate gene products of model organisms including Homo sapiens.

Sequence Ontology (SO) [17]: SO is an ontology to capture genomic features and the relationships that obtain between them. This ontology contains the features necessary to annotate a genome with structural features such as gene models and also the terms necessary for the annotation of genomic variants.

PRotein Ontology (PRO) [18]: Proteins are functional entities in many processes eventually impacted by the regulatory effect of ncRNAs (e.g., miRNA bindings). The PRO, with a particular focus on human proteins and disease-related variants thereof, provides an ontological representation of proteins.

RNA Ontology (RNAO) [19]: RNAO is an OBO foundry reference ontology to catalogue the molecular entities composing primary, secondary, and tertiary components of RNA. The goal of the RNAO project is to enable integration and analysis of diverse RNA datasets.

B. Related work in semantic mapping and semantic search

Our investigation in this paper is related to the research efforts to map different semantic models, such as ontologies, RDF/RDFS, and relational databases. In the early work, Premerlani [20] proposed a seven-step reverse engineering process and gave the guidelines to get mappings between semantic models and original schemas. Specially, one similar approach [21] to our database-to-RDF approach was to map a relational model to frame logic that can be represented in RDF. Another approach in the DOGMA ontology framework [22] also discussed how to translate a query written in some ontology language into a Structured Query Language (SQL) [23] query. A more recent research described in [24] provided a description logic-based ontology language to capture features from entity-relationship (ER) and Unified Modeling Language (UML) class diagrams. Their approach was proven to preserve the semantics of the constraints in relational databases.

As to be demonstrated and discussed later in this paper, our research focus is not just on representing relational models and relevant data in RDF/RDFS; but also more importantly, we aim to show how semantic search queries can be implemented as RDF SPARQL queries.

Semantic search is a research field that intends to improve the access to contents by considering the semantics behind the search process [25]. In other words, semantic search goes beyond keyword-based search by considering contextual meaning of words, the intend of the user and the search space. Ontologies can improve the search by *query expansion*. The original set of query keywords can be expanded by considering their synonyms or their relationship to other words that are not part of the query. In the work by Chauhan et al. [26], the original query was first expanded by considering synonyms, then terms with high semantic similarity were chosen from the ontology to be integrated to the search query, and the semantic similarity used for the query expansion was computed by the distance among concepts in the ontology, the position in the hierarchy, and the number of upper classes. On the other hand, ontology can also be used to translate keyword-based search into formal queries. For example, Tran et al. [25] used a set of models (mental, user, system, and query) to capture information, such as *thought* entities, language primitives, knowledge representation (KR) primitives, and query elements. These models were combined with a set of assumptions to redefine original queries, filling the gap between terms with structural information from an ontology (e.g., one term of the query is a property of another term). Similarly, *SemSearch* transformed original search queries based on concepts and instances from an ontology. Depending on the number of keywords involved in the search, queries were mapped into a set of structured templates.

III. OMIT reconstruction

We have significantly reconstructed the OMIT ontolgy, and we followed the same development procedure as we did in our previous investigations, that is, iteratively combining both top-down and bottom-up processes.

TABLE I. A SUBSET OF IMPORTED TERMS AND RELATIONS

Imported Term or Relation	Source Ontology and Original ID
BFO:has_part	BFO_0000051
RO:participates_in	RO_0000056
RO:has_participant	RO_0000057
NCRO:human_miRNA	NCRO_0000810
NCRO:hsa-miR-125b-1-3p	NCRO_0001813
NCRO:hsa-miR-125b-2-3p	NCRO_0001815
NCRO:hsa-miR-125b-5p	NCRO_0001816
NCRO:miRNA_target_gene	NCRO_0000025
NCRO:miRNA_and_target_gene_binding	NCRO_0000003
NCRO:protein_miRNA_promoter_binding	NCRO_0000011
CHEBI:chemical entity	CHEBI_24431
IAO:information content entity	IAO_0000030
IAO:measurement datum	IAO_0000109

A. Top-down process

A collection of terms and relations have been imported from both established upper ontologies and extant ontologies in the OBO Library, namely: Basic Formal Ontology (BFO) [27], Relation Ontology (RO) [28], Non-coding RNA Ontology (NCRO) [29] [30], Chemical Entities of Biological Interest Ontology (CHEBI) [31], and Information Artifact Ontology (IAO) [32].

Table I lists a subset of important terms and relations imported into the OMIT.

- All terms are shown in a normal font, whereas all relations are shown in an *italicized* font.
- The format for the left column (Imported Term or Relation) is "PREFIX:human-readable label" (e.g., "*BFO:has_part*").
- The format for the right column (Source Ontology and Original ID) is "PREFIX_unique identifier" (e.g., "BFO_0000051").

B. Bottom-up process

We continued to create new terms based on an in-depth analysis of different data sources, such as: miRBase [33], miRDB [4], TargetScan [34], miRGator [35], and miRanda [36]. In particular, we designed a software module (greater details can be found in Section IV) to generate new terms from the PubMed database [3].

C. Core design of the OMIT

The core design of the OMIT ontology is shown in Fig. 1. Compared with previously released versions, the current version contains many important new terms and relations, and some of which are listed in Tables II and III, respectively.

- Both terms and relations are represented in the format of "PREFIX:label" in Fig. 1.
- For the purpose of better readability, labels rather than identifiers are used in Tables II and III.
- Relations in Table III were either defined in or imported into the OMIT, which can be easily distinguished from each other by different prefixes used in the first column.

IV. SOFTWARE MODULES

A. Term-from-PubMed module

The purpose of the Term-from-PubMed software module is to create new terms out of PubMed abstracts. The software architecture is demonstrated in Fig. 2.

1) A total of 49,447 abstracts were downloaded from the PubMed database, using the search term "microRNA or miRNA or miR."
2) We utilized the OpenNLP Library [37] to process all these abstracts and obtained a total of 488,576 nouns and noun phrases.
3) All nouns and noun phrases were mapped with existing OMIT terms or relations through an ontology-alignment tool developed in one of our previous investigations [38].
4) All unmatched nouns or noun phrases were treated as candidate terms and sorted by their cumulative frequencies among all abstracts.
5) Those candidate terms with a frequency equal to or greater than 2,452 were presented to domain experts (five cellular biologists in our research group) for further checking (some example candidate terms are demonstrated in Table IV).
6) Finally, we added a total of 117 new terms into the OMIT ontology.

V. RESULTS AND DISCUSSION

A. The refactored OMIT ontology

The updated OMIT ontology contains a total of 3,081 terms and 60 relations (besides *is_a*). Screenshots of the resultant ontology in OBO-Edit [39] and Protégé [40] are demonstrated in Fig. 3 and Fig. 4, respectively. Note that both screenshots show the scenario where the term "OMIT:computationally_asserted_evidence" was selected.

B. Semantic search results

1) Experimental setup: SPARQL queries and user interface were served from the Neurocommons server [41], a vintage 2007 server with 2 2-core Xeon X5355 processors@2.66Ghz, with 32 GB of memory and 8 SATA hard disks. Other experiments were conducted on personal computers (PCs) with the following configuration: Intel(R) Core(TM) i7-3632 QM CPU @ 2.20 GHz 2.20 GHz; 8.00 GB memory; and Windows 7 64-bit Operating System.

2) Semantic search setting: To evaluate how OMIT can help *connect* different type of information that are available for the biomedical community, we have used OMIT to answer two questions of interest for a biologist.

We have added a small set of instances from platforms such as miRDB and PubMed. From miRDB [4], we have obtained miRNA-target relationships to the NCBI database and GenBank. We have selected from PubMed publications (i.e., their summaries) which are related to the selected set of genes from miRDB. Although this type of information can be obtained and integrated through D2RQ, we have used the ontology editor, Protégé, to add the instances into the ontology

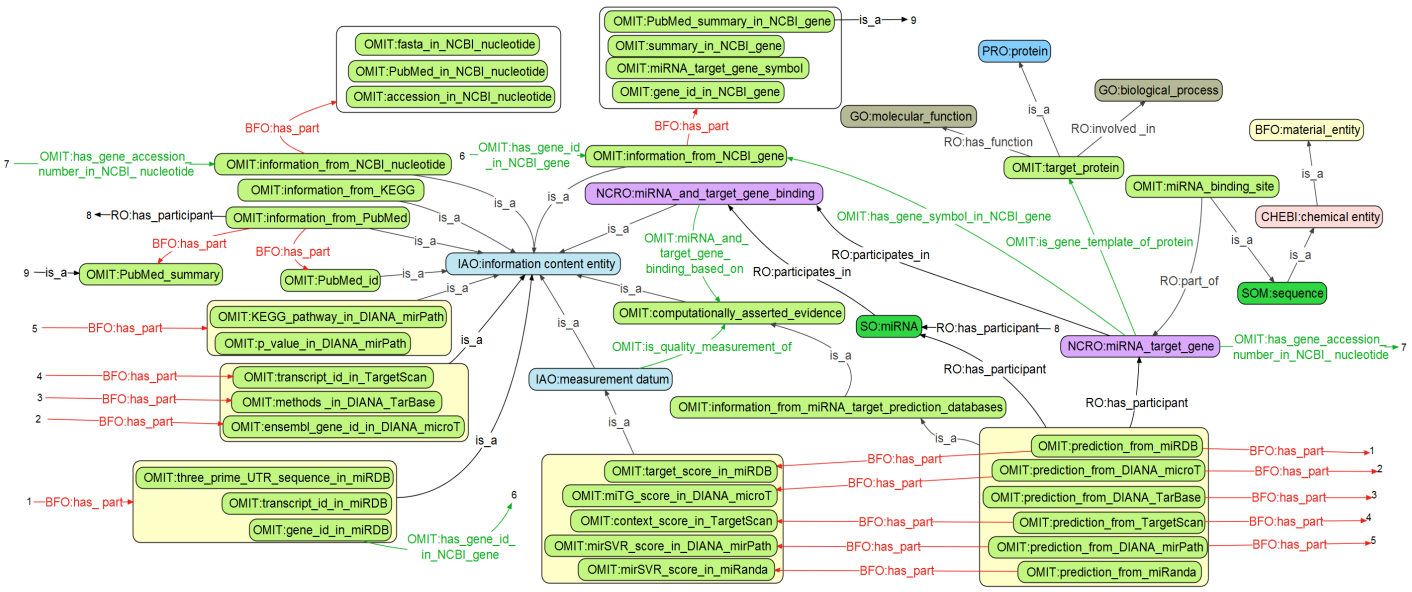

Fig. 1. The design of core terms and relations in the OMIT ontology (both terms and relations are represented in the format of "PREFIX:label").

TABLE II. A SUBSET OF NEW OMIT TERMS

OMIT New Term	Direct Parent Term	Human-Readable Explanation
computationally_asserted_evidence	IAO:information content entity	Evidence obtained from some computational methods.
information_from_miRNA_ target_prediction_database	OMIT:computationally_asserted_evidence	Records obtained from various miRNA target prediction databases.
prediction_from_miRDB	OMIT:information_from_miRNA_ target_prediction_database	Records specifically obtained from the miRDB database.
prediction_from_TargetScan	OMIT:information_from_miRNA_ target_prediction_database	Records specifically obtained from the TargetScan database.
prediction_from_miRanda	OMIT:information_from_miRNA_ target_prediction_database	Records specifically obtained from the miRanda database.
target_score_in_miRDB	IAO:measurement datum	The score of some specific miRNA-target binding prediction from the miRDB database.
context_score_in_TargetScan	IAO:measurement datum	The context score of some specific miRNA-target binding prediction from the TargetScan database.
mirSVR_score_in_miRanda	IAO:measurement datum	The mirSVR score of some specific miRNA-target binding prediction from the miRanda database.
information_from_NCBI_gene	IAO:information content entity	Records obtained from NCBI Gene according to gene IDs or gene symbols.
information_from_NCBI_nucleotide	IAO:information content entity	Records obtained from NCBI Nucleotide according to GenBank Accession numbers.
information_from_PubMed	IAO:information content entity	Records obtained from the PubMed database according to PMIDs.

TABLE III. A SUBSET OF NEW OMIT RELATIONS

New Relation in the OMIT	Domain	Range	Human-Readable Explanation
OMIT:miRNA_and_target_binding_based_on	NCRO:miRNA_and_target_gene_binding	OMIT:computationally_asserted_evidence	Specific miRNA-target binding prediction is based on some computationally asserted evidence.
OMIT:is_quality_measurement_of	IAO:measurement datum	OMIT:computationally_asserted_evidence	A piece of measurement datum (e.g., the target score in miRDB) is a quality measurement of computationally asserted evidence.
OMIT:is_gene_template_of_protein	NCRO:miRNA_target_gene	OMIT:target_protein	A miRNA target gene serves as a template of relevant protein.
RO:has_participant	OMIT:prediction_from_miRDB	SO:miRNA	Each miRNA-target binding prediction record has one miRNA as a participant.
RO:has_participant	OMIT:prediction_from_miRDB	NCRO:miRNA_target_gene	Each miRNA-target binding prediction record has one target as a participant.
BFO:has_part	OMIT:prediction_from_miRDB	OMIT:target_score_in_miRDB	Each miRNA-target binding prediction record from miRDB contains one score.
BFO:has_part	OMIT:information_from_NCBI_gene	OMIT:PubMed_summary	Each record from NCBI Gene contains one or more PubMed summaries.

TABLE IV. EXAMPLE CANDIDATE TERMS FROM PUBMED

Noun or Noun Phrase	Cumulative Frequency	Noun or Noun Phrase	Cumulative Frequency
RNA	200,243	miR	150,966
expression	82,561	cancer	38,365
ratio	29,948	tissue	15,013
RNA expression	8,698	binding	7,491
miR-12	4,075	miR-14	3,136
kinase	2,623	bone	2,452
mature miRNA	950	bronchial epithelial	100
RNA transport	39	cancer chemotherapy	36
canonical miRNA biogenesis	13	neuronal expression	6

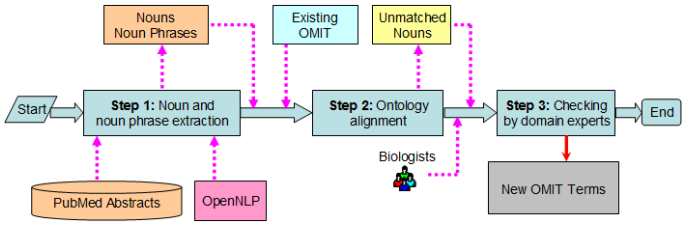

Fig. 2. The software architecture for the Term-from-PubMed module.

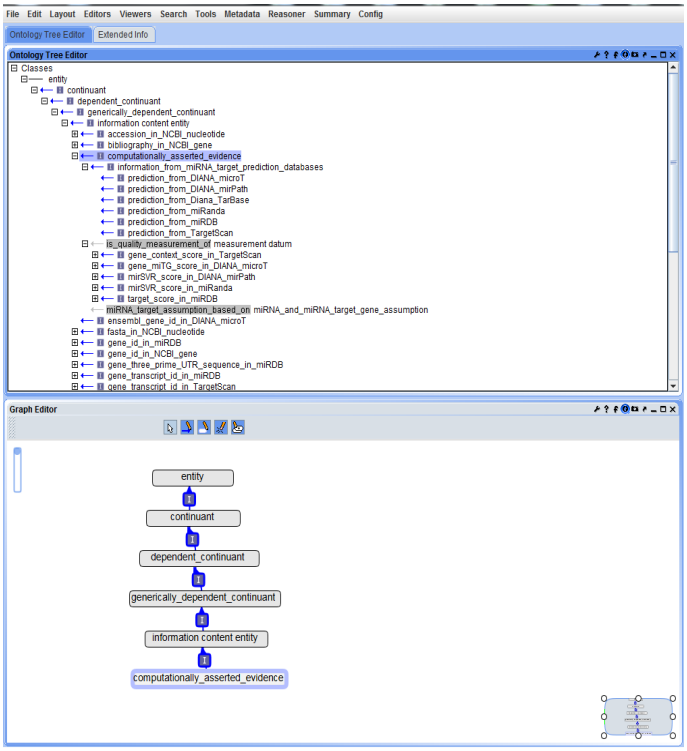

Fig. 3. A screenshot of the OMIT ontology in OBO-Edit.

because we have focused on a specific set of instances and concepts.

We have used the SPARQL engine Twinkle [42] to generate and test the queries that can provide the answers to our two evaluation questions. Twinkle provides a GUI to the SPARQL engine, which is simple enough for learning how to structure queries, yet powerful enough for more sophisticated semantic query development. Twinkle queries local and remote RDF documents, and it can perform inference over RDF Schema and OWL ontologies. The two questions queried are presented next.

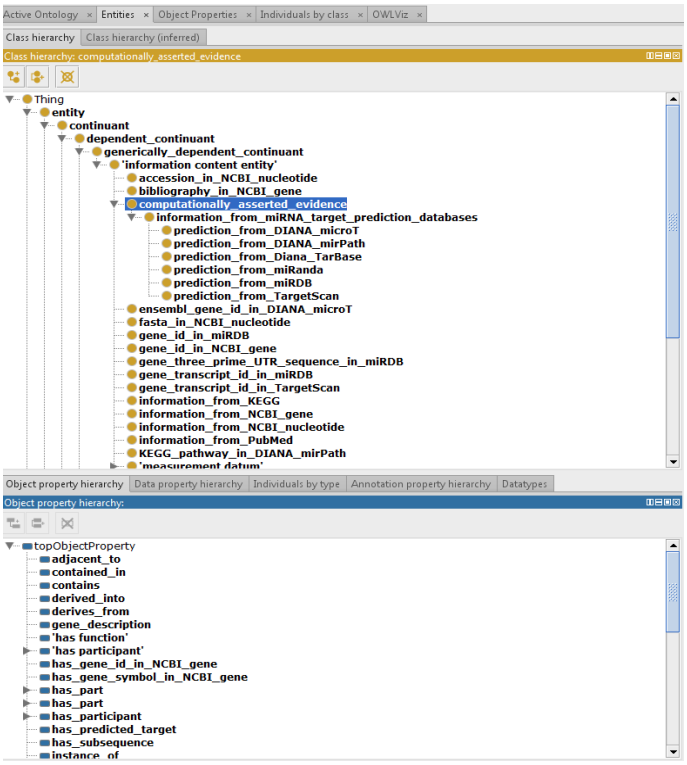

Fig. 4. A screenshot of the OMIT ontology in Protégé.

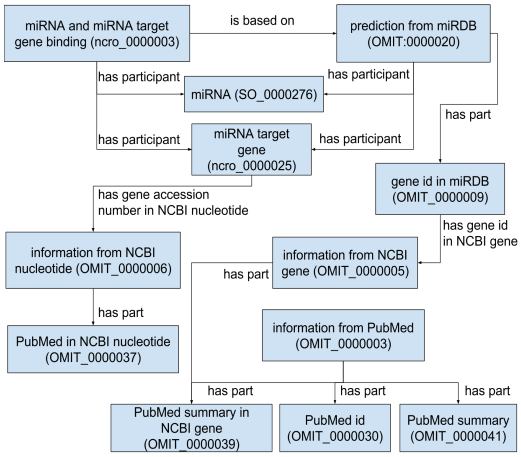

Fig. 5. Section of OMIT ontology involved in answering our evaluation questions.

3) Semantic search: Two original questions from the biologists:

- Our first evaluation question is about the role of *hsa-mir-125b* in cancer drug resistance (e.g., *IRF4*).
- The second question is regarding the role of *hsa-mir-21* in regulating *apoptosis*.

The queries for both questions use *miRNA* as the starting point. As we can see in Fig. 5, from *miRNA* we can determine predicted target genes from miRDB. The gene information gives access to PubMed publications that are related to a specific gene (i.e., *information from NCBI gene*). With the PubMed information, we can retrieve summary and identification of the relevant publications. Both question use the same approach to retrieve information. The only difference in queries between the two is that instead of filtering by *hsa-miR-125b-5p* and *IRF4*, we filter by *hsa-miR21-5p* and *PDCD4* (i.e., programmed cell death 4). The following is the SPARQL query for the first question with the results of the semantic search in Table V:

```
PREFIX rdf: <http://www.w3.org/1999/02/
                    22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/
                      rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX OMIT: <http://purl.obolibrary.org/obo/OMIT#>
PREFIX ncro: <http://purl.obolibrary.org/obo/ncro#>
PREFIX OBO_REL: <http://purl.obolibrary.org/obo/OBO_REL#>
PREFIX OMIT-obo: <http://purl.obolibrary.org/obo/
              OMIT-obo-WorkingVersion-06302015#>

SELECT ?pm_NCBI_n ?miRNA ?target_gene  ?pm_NCBI
               ?pm_id ?pm_sum
WHERE
    {?miRNA rdf:type obo:SO_0000276  .
     FILTER regex(str(?miRNA),
               'hsa*-*miR*-*125b*-*5p', 'i') .
     ?target_gene rdf:type obo:ncro_0000025 .
     FILTER regex(str(?target_gene), 'IRF4', 'i') .
     ?pm_NCBI_n rdf:type obo:OMIT_0000037 .
     ?target_gene OMIT:_has_gene_accession_number
                _in_NCBI_nucleotide ?info_NCBI_n .
     ?p_miRDB rdf:type obo:OMIT_0000020 .
     ?p_miRDB OBO_REL:_has_participant ?miRNA .
     ?p_miRDB OBO_REL:_has_participant ?target_gene .
     ?p_miRDB OMIT-obo:BFO_0000051 ?gene_id .
     ?info_pm rdf:type obo:OMIT_0000003 .
     ?gene_id OMIT:_has_gene_id_in_NCBI_gene ?info_NCBI .
     ?pm_NCBI rdf:type obo:OMIT_0000039 .
     ?info_pm OMIT-obo:BFO_0000051 ?pm_NCBI .
     ?pm_id rdf:type obo:OMIT_0000030 .
     ?pm_sum rdf:type obo:OMIT_0000041 .
     ?info_pm OMIT-obo:BFO_0000051 ?pm_id .
     ?info_pm OMIT-obo:BFO_0000051 ?pm_sum .}
```

The following is the SPARQL query for the second question with the results in Table VI:

```
...
SELECT ?pm_NCBI_n ?miRNA ?target_gene
               ?pm_NCBI ?pm_id ?pm_sum
WHERE
     {?miRNA rdf:type obo:SO_0000276  .
    FILTER regex( str(?miRNA),
        'hsa*-*miR*-*21*-*5p', 'i' ) .
     ?target_gene rdf:type obo:ncro_0000025 .
     FILTER regex(str(?target_gene), 'PDCD4', 'i') .
     ?pm_NCBI_n rdf:type obo:OMIT_0000037 .
         ?target_gene
    OMIT:_has_gene_accession_number_in
      _NCBI_nucleotide ?info_NCBI_n .
    ?p_miRDB rdf:type obo:OMIT_0000020 .
    ?p_miRDB OBO_REL:_has_participant ?miRNA .
    ?p_miRDB OBO_REL:_has_participant ?target_gene .
    ?p_miRDB OMIT-obo:BFO_0000051 ?gene_id .
    ?info_pm rdf:type obo:OMIT_0000003 .
    ?gene_id OMIT:_has_gene_id_in_NCBI_gene ?info_NCBI .
    ?pm_NCBI rdf:type obo:OMIT_0000039 .
    ?info_pm OMIT-obo:BFO_0000051 ?pm_NCBI .
    ?pm_id rdf:type obo:OMIT_0000030 .
    ?pm_sum rdf:type obo:OMIT_0000041 .
    ?info_pm OMIT-obo:BFO_0000051 ?pm_id .
    ?info_pm OMIT-obo:BFO_0000051 ?pm_sum .}
```

C. Discussion

1) The significantly restructured OMIT ontology: Important changes are summarized as follows.

- The majority (around 82%) of all 3,081 terms in the updated OMIT were imported from the NCRO ontology [29] [30]. Because the NCRO is a comprehensive domain ontology in the ncRNA field, following the NCRO hierarchy will enhance the interoperability between the OMIT and future ontologies to be developed in other ncRNA sub-domains.
- All miRNAs appearing in humans were encoded, along with the information about the gene family group of each miRNA. According to miRBase [33], there are a total of 1,981 distinct human miRNAs, belonging to 320 different gene family groups. This information can be highly valuable because the fact that two or more miRNAs of interest indeed belong to the same gene family group can provide biologists, bioinformaticians, and clinical investigators with critical clues in constructing new hypothesis.
- As discussed in Section IV, a total of 117 new terms were added based on the analysis of PubMed abstracts.
- To further disseminate the ontology, and, to gather feedback from community in a more effective manner, we created a GitHub project site for the OMIT, on top of three available resources (a designated project website [43], the OBO Library [44], and the NCBO BioPortal [45]) that were established in our previous investigations.
- We also established a tracker [46] for an enhanced mechanism in handling the discussion among groups to further improve the ontology. New concepts, definitions, and their locations in the OMIT can be proposed, debated, and approved (or rejected) by an open group of individuals through this tracker.

2) Semantic search: In the process of evaluating the OMIT for semantic search, it becomes clear that the ontology is highly connected. This high connectivity between concepts in the ontology allows us to develop approaches for querying. This translates to more efficient queries and more complete answers. This high connectivity also leads to certain redundancy

TABLE V. SAMPLE OF QUERY RESULTS FOR THE FIRST EVALUATION QUESTION.

PubMed ID from NCBI Gene	Summary	PubMed ID from NCBI Nucleotide	Summary
25987254	Selective targeting of IRF4 by synthetic microRNA-125b-5p mimics induces anti-multiple myeloma activity in vitro and in vivo. Morelli E, Leone E, Cantafio ME, Di Martino MT, Amodio N, Biamonte L, Gull A, Foresta U, Pitari MR, Botta C, Rossi M, Neri A, Munshi NC, Anderson KC, Tagliaferri P, Tassone P.	8921401	Cloning of human lymphocyte-specific interferon regulatory regulatory factor (hLSIRF/hIRF4) and mapping of the gene to 6p23-p25. Grossman A, Mittrcker HW, Nicholl J, Suzuki A, Chung S, Antonio L, Suggs S, Sutherland GR, Siderovski DP, Mak TW.
24292686	No evidence for a genetic association of IRF4 with systemic lupus erythematosus in a Chinese population. Liu SS, Ye D, Lou J, Fan Z, Ye DQ.	9326949	Deregulation of MUM1/IRF4 by chromosomal translocation in multiple myeloma. Iida S, Rao PH, Butler M, Corradini P, Boccadoro M, Klein B, Chaganti RS, Dalla-Favera R.
25987254	Differentiation stage-specific expression of microRNAs in B lymphocytes and diffuse large B-cell lymphomas. Malumbres R, Sarosiek KA, Cubedo E, Ruiz JW, Jiang X, Gascoyne RD, Tibshirani R, Lossos IS.	24995979	IRF4 is a key thermogenic transcriptional partner of PGC-1α.t Kong X, Banks A, Liu T, Kazak L, Rao RR, Cohen P, Wang X, Yu S, Lo JC, Tseng YH, Cypess AM, Xue R, Kleiner S, Kang S, Spiegelman BM, Rosen ED.
24906573	Association of interferon regulatory factor 4 gene polymorphisms rs12203592 and rs872071 with skin cancer and haematological malignancies susceptibility: a meta-analysis of 19 case-control studies. Wang S, Yan Q, Chen P, Zhao P, Gu A.	24267888	A polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway. Praetorius C, Grill C, Stacey SN, Metcalf AM, Gorkin DU, Robinson KC, Van Otterloo E, Kim RS, Bergsteinsdottir K, Ogmundsdottir MH, Magnusdottir E, Mishra PJ, Davis SR, Guo T, Zaidi MR, Helgason AS, Sigurdsson MI, Meltzer PS, Merlino G, Petit V, Larue L, Loftus SK, Adams DR, Sobhiafshar U, Emre NC, Pavan WJ, Cornell R, Smith AG, McCallion AS, Fisher DE, Stefansson K, Sturm RA, Steingrimsson E.
24906573	Association of interferon regulatory factor 4 gene polymorphisms rs12203592 and rs872071 with skin cancer and haematological malignancies susceptibility: a meta-analysis of 19 case-control studies. Wang S, Yan Q, Chen P, Zhao P, Gu A.	10601358	Lineage-specific modulation of interleukin 4 signaling by interferon regulatory factor 4. Gupta S, Jiang M, Anthony A, Pernis AB.

TABLE VI. SAMPLE OF QUERY RESULTS FOR THE SECOND EVALUATION QUESTION.

PubMed ID from NCBI Gene	Summary	PubMed ID from NCBI Nucleotide	Summary
25316501	The PDCD4/miR-21 pathway in medullary thyroid carcinoma. Pennelli G, Galuppini F, Barollo S, Cavedon E, Bertazza L, Fassan M, Guzzardo V, Pelizzo MR, Rugge M, Mian C.	23541896	A genome-wide association meta-analysis of self-reported Structural mechanism of CCM3 heterodimerization with GCKIII kinases. Zhang M, Dong L, Shi Z, Jiao S, Zhang Z, Zhang W, Evans DM, St Pourcain B, Ring SM, Mountain JL, Francke U, L, Zhou Z.
19672202	Cholesteatoma growth and proliferation: posttranscriptional regulation by microRNA-21. Friedland DR, Eernisse R, Erbe C, Gupta N, Cioffi JA.	23388056	Loss of CCM3 impairs DLL4-Notch signalling: implication in endothelial angiogenesis and in inherited cerebral cavernous malformations. You C, Sandalcioglu IE, Dammann P, Felbor U, Sure U, Zhu Y.
26289851	Antisense-miR-21 enhances differentiation/apoptosis and reduces cancer stemness state on anaplastic thyroid cancer. Haghpanah V, Fallah P, Tavakoli R, Naderi M, Samimi H, Soleimani M, Larijani B.	24058906	Sporadic cerebral cavernous malformations: report of further mutations of CCM genes in 40 Italian patients. D'Angelo R, Alafaci C, Scimone C, Ruggeri A, Salpietro FM, Bramanti P, Tomasello F, Sidoti A.
26284486	Exosomal levels of miRNA-21 from cerebrospinal fluids associated with poor prognosis and tumor recurrence of glioma patients. Shi R, Wang PY, Li XY, Chen JX, Li Y, Zhang XZ, Zhang CG, Jiang T, Li WB, Ding W, Cheng SJ.	22750858	Crystallization and preliminary crystallographic studies of CCM3 in complex with the C-terminal domain of MST4. Xu X, Wang X, Ding J, Wang DC.
20508945	MiR-21 overexpression in human primary squamous cell lung carcinoma is associated with poor patient prognosis. Gao W, Shen H, Liu L, Xu J, Xu J, Shu Y.	23485406	Genomic causes of multiple cerebral cavernous malformations in a Japanese population. Tsutsumi S, Ogino I, Miyajima M, Ikeda T, Shindo N, Yasumoto Y, Ito M, Arai H.

in the data. We could modify OMIT to avoid the redundancy by eliminating or consolidating concepts and relationships. However, these changes can lead to contradictions to the domain knowledge. Another alternative is to use the redundancy as a validation mechanism for the query or the data itself. However, it is not clear how a validation mechanism could affect the performance when answering queries.

VI. CONCLUSIONS

As a special class of ncRNAs, miRNAs have been demonstrated to play important roles in various biological and pathological processes. Because miRNAs realize their functions by regulating respective targets, it is critical to identify and analyze miRNA-target interactions for a better delineation of miRNAs' functions. Emerging semantic technologies and domain ontologies have been utilized to overcome limitations identified during conventional miRNA knowledge acquisition methods. We followed the research direction identified in our previous investigations regarding the establishment of common data elements and data exchange standards in the miRNA research. Specifically, our major contributions in this paper are: (1) our continuing efforts and significant improvements on the OMIT ontology development and (2) the application of the OMIT to enable a semantic approach for knowledge capture of miRNA-target interactions, leading to more effective miRNA data integration and knowledge discovery. This paper describes our research design, methodologies, software implementation, experimental outcomes, and relevant discussion.

One obvious future research direction is to continue the development and refinement of the OMIT ontology. Another interesting piece of future work is to incorporate into our system other related data sources during the miRNA knowledge discovery and acquisition.

ACKNOWLEDGMENT

Research reported in this paper was partially supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH), under the Award Number U01CA180982. The views contained in this paper are solely the responsibility of the authors and do not represent the official views, either expressed or implied, of the NIH or the U.S. Government.

REFERENCES

[1] Y. H. Zhao, M. Zhou, H. Liu, H. T. Khong, D. H. Yu, O. Fodstad, and M. Tan, "Upregulation of lactate dehydrogenase-A by ErbB2 through heat shock factor 1 promotes breast cancer cell glycolysis and growth," *Oncogene*, vol. 28, no. 42, pp. 3689–3701, Octorber 2009.

[2] Z. Liu, H. Liu, S. Desai, D. Schmitt, M. Zhou, H. T. Khong, K. S. Klos, S. McClellan, O. Fodstad, and M. Tan, "MiR-125b functions as a key mediator for snail-induced stem cell propagation and chemoresistance," *J Biol Chem*, vol. 288, no. 6, pp. 4334–4345, February 2013.

[3] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," *Database*, January 2011.

[4] miRDB. [Online]. Available: http://mirdb.org/miRDB/

[5] NCBI Gene. [Online]. Available: http://ncbi.nlm.nih.gov/gene

[6] NCBI Nucleotide. [Online]. Available: ncbi.nlm.nih.gov/nucleotide/

[7] SPARQL Query Language for RDF. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[8] J. Huang, M. Tan, D. Dou, L. He, C. Townsend, R. Rudnick, and P. Hayes, "MiRNA Ontology for Target Prediction in Human Cancer," in *Proc. 1st ACM International Conference on Bioinformatics and Computational Biology, ACM-BCB-2010*, August 2010.

[9] C. Townsend, J. Huang, D. Dou, S. Dalvi, P. Hayes, L. He, W. Lin, H. Liu, R. Rudnick, H. Shah, H. Sun, X. Wang, and M. Tan, "OMIT: Domain Ontology and Knowledge Acquisition in MicroRNA Target Prediction," in *Proc. 9th Intl' Conference on Ontologies, DataBases, and Applications of Semantics, ODBASE-2010*, October 2010.

[10] J. Huang, C. Townsend, D. Dou, H. Liu, and M. Tan, "OMIT: A Domain-Specific Knowledge Base for MicroRNA Target Prediction," *Pharm Res*, vol. 28, no. 12, pp. 3101–3104, August 2011.

[11] J. Huang, J. Dang, X. Lu, D. Dou, J. Blake, W. Gerthoffer, and M. Tan, "An Ontology-Based MicroRNA Knowledge Sharing and Acquisition Framework," in *Proc. BHI Workshop at 2012 IEEE International Conference on Bioinformatics and Biomedicine, BIBM-2012*, October 2012.

[12] J. Huang, J. Dang, X. Lu, M. Xiong, W. Gerthoffer, and M. Tan, "Semi-Automated microRNA Ontology Development based on Artificial Neural Networks," in *Proc. 2013 IEEE International Conference on Bioinformatics and Biomedicine, BIBM-2013)*, December 2013.

[13] J. Huang, J. Dang, G. Borchert, H. Zhang, M. Xiong, W. Gerthoffer, K. Eilbeck, J. Blake, and M. Tan, "OMIT: A Dynamic microRNA Domain Ontology for Microgenomics Knowledge Discovery, Unification, and Bio-Curation," *PLoS One*, vol. 9, no. 7, July 2014.

[14] OBO Library. [Online]. Available: http://obofoundry.org

[15] NCBO BioPortal. [Online]. Available: https://bioportal.bioontology.org/

[16] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, no. 1, pp. 25–29, May 2000.

[17] K. Eilbeck, S. Lewis, C. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner, "The Sequence Ontology: a tool for the unification of genome annotations," *Genome Biol*, vol. 6, no. 5, April 2005.

[18] D. Natale, C. Arighi, W. Barker, J. Blake, C. Bult, M. Caudy, H. Drabkin, P. D'Eustachio, A. Evsikov, H. Huang, J. Nchoutmboube, N. Roberts, B. Smith, J. Zhang, and C. Wu, "The Protein Ontology: a structured representation of protein forms and complexes," *Nucleic Acids Res*, vol. 39, pp. D539–D545, January 2011.

[19] R. Hoehndorf, C. Batchelor, T. Bittner, M. Dumontier, K. Eilbeck, R. Knight, C. Mungall, J. Richardson, J. Stombaugh, E. Westhof, C. Zirbel, and N. Leontis, "The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structure data," *Applied Ontology*, vol. 6, no. 1, pp. 53–89, January 2011.

[20] W. J. Premerlani and M. R. Blaha, "An Approach for Reverse Engineering of Relational Databases," *Commun. ACM Journal*, vol. 37, no. 5, pp. 42–49, 134, 1994.

[21] L. Stojanovic, N. Stojanovic, and R. Volz, "Migrating data-intensive web sites into the Semantic Web," in *Proceedings of ACM symposium on Applied computing*. ACM Press, 2002, pp. 1100–1107.

[22] P. Verheyden, J. D. Bo, and R. Meersman, "Semantically Unlocking Database Content Through Ontology-Based Mediation," in *SWDB*, 2004, pp. 109–126.

[23] E. Codd, "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM*, vol. 13, no. 6, pp. 377–387, June 1970.

[24] L. Lubyte and S. Tessaris, "Extracting Ontologies from Relational Databases," in *Proceedings of Description Logics*, 2007, pp. 122–126.

[25] T. Tran, P. Cimiano, S. Rudolph, and R. Studer, "Ontology-based interpretation of keywords for semantic search," in *The Semantic Web*, K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudr-Mauroux, Eds. Springer Berlin Heidelberg, 2007, vol. 4825, pp. 523–536.

[26] R. Chauhan, R. Goudar, R. Sharma, and A. Chauhan, "Domain ontology based semantic search for efficient information retrieval through automatic query expansion," in *Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on*, March 2013, pp. 397–402.

[27] BFO. [Online]. Available: http://www.ifomis.org/bfo/

[28] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, and C. Rosse, "Relations in biomedical ontologies," *Genome Biology*, vol. 6, no. 5, April 2005.

[29] NCRO. [Online]. Available: https://github.com/OmniSearch/ncRO-ontology-files

[30] NCRO Project Site. [Online]. Available: http://omnisearch.soc.southalabama.edu/OntologyFile.aspx

[31] CHEBI. [Online]. Available: http://www.obofoundry.org/cgi-bin/detail.cgi?id=chebi

[32] IAO. [Online]. Available: http://www.obofoundry.org/cgi-bin/detail.cgi?id=information_artifact

[33] miRBase. [Online]. Available: http://www.mirbase.org/

[34] TargetScan. [Online]. Available: http://www.targetscan.org

[35] miRGator. [Online]. Available: http://genome.ewha.ac.kr/miRGator

[36] miRanda. [Online]. Available: http://www.microrna.org

[37] OpenNLP. [Online]. Available: https://opennlp.apache.org/

[38] J. Huang, J. Dang, M. Huhns, and W. Zheng, "Use Artificial Neural Network to Align Biological Ontologies," *BMC Genomics*, vol. 9, no. Suppl 2, p. S16, 2008.

[39] OBO-Edit. [Online]. Available: http://oboedit.org/

[40] Protégé. [Online]. Available: http://protege.stanford.edu/

[41] A. Ruttenberg, J. Rees, M. Samwald, and M. Marshall, "Life sciences on the Semantic Web: the Neurocommons and beyond," *Brief Bioinform*, vol. 10, no. 2, pp. 193–204, 2009.

[42] [Online]. Available: http://www.ldodds.com/projects/twinkle/

[43] OMIT Project Site. [Online]. Available: http://omnisearch.soc.southalabama.edu

[44] OMIT in OBO Library. [Online]. Available: http://www.obofoundry.org/cgi-bin/detail.cgi?id=omit

[45] OMIT in NCBO BioPortal. [Online]. Available: http://bioportal.bioontology.org/ontologies/OMIT

[46] OMIT Tracker. [Online]. Available: https://github.com/OmniSearch/OMIT-ontology-files/issues