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
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 ontology, 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.
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

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

TABLE III. A SUBSET OF NEW OMIT RELATIONS

<table>
<thead>
<tr>
<th>New Relation in the OMIT</th>
<th>Domain</th>
<th>Range</th>
<th>Human-Readable Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMIT:miRNA_and_target_binding_based_on</td>
<td>NCRO:miRNA_and_target_gene_binding</td>
<td>OMIT:computationally asserted evidence</td>
<td>Specific miRNA-target binding prediction is based on some computationally asserted evidence.</td>
</tr>
<tr>
<td>OMIT:is_quality_measurement_of</td>
<td>IAO:measurement datum</td>
<td>OMIT:computationally asserted evidence</td>
<td>A piece of measurement datum (e.g., the target score in miRDB) is a quality measurement of computationally asserted evidence.</td>
</tr>
<tr>
<td>RO:has_participant</td>
<td>OMIT:prediction_from_miRDB</td>
<td>SO:miRNA</td>
<td>Each miRNA-target binding prediction record has one miRNA as a participant.</td>
</tr>
<tr>
<td>RO:has_participant</td>
<td>OMIT:prediction_from_miRDB</td>
<td>NCRO:miRNA_target_gene</td>
<td>Each miRNA-target binding prediction record has one target as a participant.</td>
</tr>
<tr>
<td>BFO:has_part</td>
<td>OMIT:prediction_from_miRDB</td>
<td>OMIT:target_score_in_miRDB</td>
<td>Each miRNA-target binding prediction record from miRDB contains one score.</td>
</tr>
<tr>
<td>BFO:has_part</td>
<td>OMIT:information_from_NCBI_gene</td>
<td>OMIT:PubMed_summary</td>
<td>Each record from NCBI Gene contains one or more PubMed summaries.</td>
</tr>
</tbody>
</table>
TABLE IV. EXAMPLE CANDIDATE TERMS FROM PUBMED

<table>
<thead>
<tr>
<th>Noun or Noun Phrase</th>
<th>Cumulative Frequency</th>
<th>Noun or Noun Phrase</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA</td>
<td>200,243</td>
<td>miR</td>
<td>150,966</td>
</tr>
<tr>
<td>expression</td>
<td>82,361</td>
<td>cancer</td>
<td>38,365</td>
</tr>
<tr>
<td>ratio</td>
<td>29,484</td>
<td>tissue</td>
<td>15,013</td>
</tr>
<tr>
<td>RNA expression</td>
<td>8,698</td>
<td>binding</td>
<td>7,491</td>
</tr>
<tr>
<td>miR-12</td>
<td>4,075</td>
<td>miR-14</td>
<td>3,136</td>
</tr>
<tr>
<td>kinase</td>
<td>2,623</td>
<td>bone</td>
<td>2,452</td>
</tr>
<tr>
<td>mature miRNA</td>
<td>950</td>
<td>bronchial epithelial</td>
<td>100</td>
</tr>
<tr>
<td>RNA transport</td>
<td>39</td>
<td>cancer chemotherapy</td>
<td>36</td>
</tr>
<tr>
<td>canonical miRNA biogenesis</td>
<td>13</td>
<td>neuronal expression</td>
<td>6</td>
</tr>
</tbody>
</table>

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

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

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-mirR21-5p 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 xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ncro: <http://purl.obolibrary.org/obo/ncro#>
PREFIX OMIT: <http://purl.obolibrary.org/obo/OMIT#>

WHERE {
  FILTER regex(str(?miRNA), 'hsa\-*miR\-*125b\-*5p', 'i') .
  FILTER regex(str(?target_gene), 'IRF4', 'i') .
  ?pm_NCBI_n rdf:type obo:OMIT_00000276 .
  FILTER regex(str(?pm_NCBI_n), 'IRF4', 'i') .
  ?pm_NCBI_n rdf:type obo:OMIT_00000276 .
  ?pm_NCBI rdf:type obo:OMIT_0000037 .
  ?pm_NCBI_n rdf:type obo:OMIT_0000037 .
  ?pm_NCBI rdf:type obo:OMIT_0000037 .
  ?info_pm rdf:type obo:OMIT_0000030 .
  ?pm_id rdf:type obo:OMIT_0000041 .
  ?pm_NCBI rdf:type obo:OMIT_0000030 .
  ?pm_sum rdf:type obo:OMIT_0000041 .
  ?pm_NCBI ?pm_id ?pm_sum .
}
```

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

```
WHERE {
  FILTER regex(str(?miRNA), 'hsa\-*miR\-*125b\-*5p', 'i') .
  FILTER regex(str(?target_gene), 'PDCD4', 'i') .
  ?pm_NCBI_n rdf:type obo:OMIT_0000037 .
  ?pm_NCBI ?pm_id ?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
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