Pharm Res
DOI 10.1007/s11095-011-0573-8

COMMENTARY

# OMIT: A Domain-Specific Knowledge Base for MicroRNA Target Prediction

Jingshan Huang • Christopher Townsend • Dejing Dou • Haishan Liu • Ming Tan

Received: 1 April 2011 / Accepted: 15 August 2011


**ABSTRACT** Identification and characterization of the important roles microRNAs (miRNAs) perform in human cancer is an increasingly active research area. Unfortunately, prediction of miRNA target genes remains a challenging task to cancer researchers. Current processes are time-consuming, error-prone, and subject to biologists' limited prior knowledge. Therefore, we propose a domain-specific knowledge base built upon Ontology for MicroRNA Targets (OMIT) to facilitate knowledge acquisition in miRNA target gene prediction. We describe the ontology design, semantic annotation and data integration, and user-friendly interface and conclude that the OMIT system can assist biologists in unraveling the important roles of miRNAs in human cancer. Thus, it will help clinicians make sound decisions when treating cancer patients.

**Electronic supplementary material** The online version of this article (doi:10.1007/s11095-011-0573-8) contains supplementary material, which is available to authorized users.

J. Huang (✉) • C. Townsend
School of Computer and Information Sciences
University of South Alabama
307 University Blvd. N
Mobile, Alabama, USA
e-mail: huang@usouthal.edu

D. Dou • H. Liu
Department of Computer and Information Science
University of Oregon
Eugene, Oregon, USA

M. Tan
Mitchell Cancer Institute University of South Alabama
Mobile, Alabama, USA

M. Tan
Department of Cell Biology and Neuroscience
University of South Alabama
Mobile, Alabama, USA

**KEY WORDS** human cancer · knowledge acquisition · knowledge base · microRNA (miRNA) target · ontology

## INTRODUCTION

The identification and characterization of the important roles microRNAs (miRNAs) perform in human cancer is an increasingly active research area. As a special class of small non-coding RNAs, miRNAs have been reported to perform critical roles in a variety of biological processes by regulating target genes (1,2). Moreover, miRNA expression profiling of many tumor types has identified miRNAs associated with cancer development, diagnosis, treatment, and prognosis (3,4). Unfortunately, the prediction of miRNA target genes remains a challenging task to cancer researchers. In particular, substantial time and effort have been expended in every search for available information in each small miRNA subarea. To identify miRNAs' target genes is very difficult: not only do biologists need to extract a large number of candidate target genes from existing miRNA target prediction databases, but they will also need to manually search for these genes' related information (e.g., their cellular components and biological processes) from resources other than miRNA databases for each of the hundreds of candidate target genes. The whole process is time-consuming, error-prone, and subject to biologists' limited prior knowledge. In addition, the situation is further aggravated by the great complexity and imprecise terminologies that characterize the biological and biomedical research fields. A great deal of variety has been identified in the adoption of different biological terms, along with divergent relationships among all these terms. Such variety has inhibited effective information acquisition by humans.

Published online: 31 August 2011

## OMIT FRAMEWORK

Ontologies are formal, declarative knowledge representation models, performing a key role in defining formal semantics in traditional knowledge engineering. Therefore, we explore a domain-specific knowledge base built upon the Ontology for MicroRNA Targets (OMIT) to handle challenges in miRNA target acquisition. The OMIT ontology is *the very first ontology* in the miRNA area, and the OMIT framework facilitates knowledge discovery and sharing from existing sources. As a result, the long-term objective is to assist biologists in unraveling the important roles of miRNAs in human cancer; thus, it will help clinicians make sound decisions when treating cancer patients. We aim to synthesize data from existing miRNA target databases into a comprehensive conceptual model that permits an emphasis on data semantics rather than on the forms in which the data were originally represented. Consequently, a more accurate, complete view of miRNAs' biological functions can be acquired. We designed the OMIT ontology specifically for the miRNA target domain, and then carried out the semantic annotation and data integration, based upon which a domain-specific knowledge base was created. Finally, a friendly user interface was designed to demonstrate integrated information from distributed data sources, along with newly obtained knowledge via reasoning mechanisms. The overall structure of the OMIT framework is described in this section, and more details can be found in the Supplementary Material.

### Overview of the OMIT Framework

As shown in Fig. 1, the main components of the OMIT framework are an ontology and a knowledge base. Information from distributed databases can be synthesized and presented to end users in a uniform view, integrated with additional information from the Gene Ontology. The Gene Ontology consists of three components (biological processes, cellular components, and molecular functions), and it provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data, as well as tools to access and process such data. More details are included in the Supplementary Material.

A typical knowledge acquisition process takes eight steps:

- Steps 1 and 2: User sends a search/query to the OMIT system through the user interface
- Step 3: The recognized miRNA concept in the OMIT is used to query the knowledge base
- Step 4: miRNA targets (i.e., genes) are retrieved
- Step 5: Obtained targets are utilized to acquire more gene information
- Step 6: Related gene information is returned
- Steps 7 and 8: miRNA targets and their related gene information are returned to the user

### The OMIT Ontology

The first-version OMIT ontology consists of a total of 327 concepts and 58 relationships (i.e., 28 object properties and 30 data type properties). This version has been submitted and accepted by the NCBO BioPortal. The OMIT ontology file can be freely downloaded from http://bioportal.bioontology.org/ontologies/42873

### The OMIT Knowledge Base

The first-version OMIT knowledge base contains a total of 1,889 facts (referred to as "axioms" in Protégé). These facts are specified in OWL and include 27 subclass axioms, 59 disjoint class axioms, 4 sub object property axioms, 3 inverse object property axioms, 22 object property domain axioms, 27 object property range axioms, 21 data property domain axioms, 30 data property range axioms, 166 class assertion axioms, 308 object property assertion axioms, 674 data property assertion axioms, and 248 entity annotation axioms.

### User Query/Search Answering

A friendly graphical user interface (GUI) to answer users' query/search has been designed with the C# language in Visual Studio 2010. As demonstrated in Fig. 2, users can specify the miRNA of interest along with expected properties of this miRNA. Both selections are made through drop-down lists so that the effort required for providing such input is minimized; corresponding values for selected properties are then retrieved and populated in a separate panel. Figure 2 exhibits part of results when "*mir-21*" and seven properties were chosen. Note that the retrieved results are regarded as integrated information in the sense that no one data source alone in our framework contains such complete knowledge. In addition to this integrated information, deep, hidden knowledge is acquired as well. Some examples include "p53 must not be a direct target of mir-885-5p" and "*mir-21* upRegulates MalignantNeoplasm." The ability to obtain previously implicit knowledge is due to the inference mechanisms applied to the knowledge base. More detailed discussion on obtaining hidden, critical domain knowledge can be found in the Supplementary Material.

## CONCLUSION

In this paper, we propose an innovative computing framework based on the miRNA-domain-specific knowl-

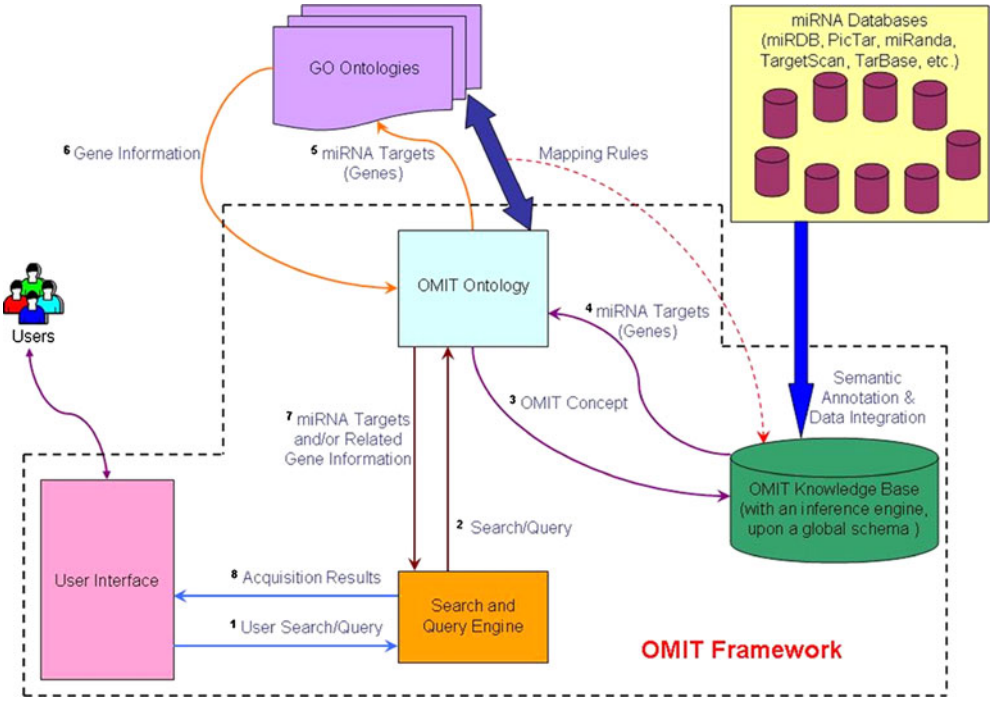


**Fig. 1** Overall structure of the OMIT framework.

edge base, OMIT, to handle the challenge of an efficient acquisition of miRNAs' candidate target genes. To the best of our knowledge, the OMIT framework is designed upon the very first ontology in the miRNA domain and includes a domain-specific knowledge base. We adopt a combination of both top-down and bottom-up approaches when designing the OMIT ontology. A deep annotation is utilized during semantic annotation and data integration, which together lead to a centralized knowledge base. The OMIT system is able to assist biologists in unraveling the important roles for miRNAs in human cancer; thus, it will help clinicians make sound decisions when treating cancer patients. This long-term research goal will be achieved via facilitating knowledge discovery and sharing from existing sources.

## ACKNOWLEDGMENTS

The authors would like to thank Hardik Shah and Robert Rudnick for helping in software implementation. The authors also appreciate the discussion with Patrick Hayes, Lei He, Wen-chang Lin, Hao Sun, and Xiaowei Wang.

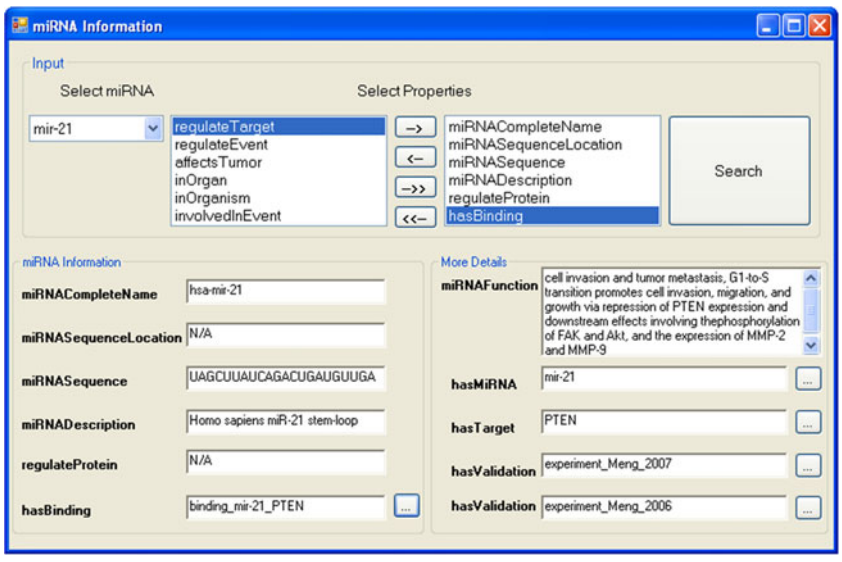


**Fig. 2** Search/query GUI in the OMIT.

## REFERENCES


1. Kobayashi T, Lu J, Cobb BS, Rodda SJ, McMahon AP, Schipani E, *et al.* Dicer-dependent pathways regulate chondrocyte proliferation and differentiation. Proc Natl Acad Sci. 2008; 105:1949–54.
2. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature. 2000;403:901–6.
3. Zhou M, Liu ZX, Zhao YH, Ding Y, Liu H, Xi Y, *et al.* MicroRNA-125b confers the resistance of breast cancer cells to paclitaxel through suppression of Bak1. J Biol Chem. 2010;285(28):21496–507.
4. Nakajima G, Hayashi K, Xi Y, Kudo K, Uchida K, Takasaki K, *et al.* Non-coding microRNAs hsa-let-7 g and hsa-miR-181b are associated with chemoresponse to S-1 in colon cancer. Cancer Genomics Proteomics. 2006;3:317–24.

**SUPPLEMENTARY MATERIAL**

## Supplementary Data in Background

*Related Work on MiRNA, Cancer, and MiRNA Target Prediction*

MiRNAs are a class of endogenous, small, non-coding, single-stranded RNAs. They regulate gene expression at the post-transcriptional and translational levels, and they constitute a novel class of gene regulators (1). Mature miRNA molecules are complementary or partially complementary to one or more messenger RNA molecules. They translationally down-regulate gene expression or induce the degradation of messenger RNAs (2). The biological functions of miRNAs include regulating proliferation, development, differentiation, migration, apoptosis, and the cell cycle (3), and miRNAs have been found to be involved in cancer development, acting as potential oncogenes or tumor suppressors (4,5). The importance of miRNA research was not fully recognized until hundreds of miRNAs in worm, fly, and mammalian genomes were identified recently (6). In addition, the miRNA gene family is one of the largest in higher eukaryotes: according to the current release of miRBase (7), more than 1,000 mature miRNAs have been identified in the human genome, and these miRNAs account for about three percent of all human genes.

Cancer is a genetic disease. The activation of oncogenes and genetic defects in tumor suppressor genes are major contributors to the development of cancer (5). Due to the ability of miRNAs to induce rapid changes in protein synthesis without the need for transcriptional activation and subsequent messenger RNA processing steps, miRNA-regulated controls provide cells with a more precise, rapid, and energy-efficient way of regulating protein expression. In contrast to messenger RNAs, miRNAs are regulatory molecules with small numbers of nucleotides (19-27 nt). The small size and relatively stable structure of miRNAs allow reliable analysis of clinically archived patient samples, and they further suggest that miRNAs may be appropriate biomarkers and potential therapeutic targets in cancer.

Two categories of approaches have been developed for identifying the targets of miRNAs: (i) experimental (direct biochemical characterization) approaches and (ii) computational approaches (8–10).

After candidate miRNA targets have been identified through computational approaches, the next step is to experimentally validate their targets. Because direct experimental methods for discovering miRNA targets are time-consuming and costly, many target prediction algorithms have been developed. In addition, computational identification of miRNA targets in mammals is considerably more difficult than in plants because most animal miRNAs only partially hybridize to their targets. Most miRNA target prediction programs adopt machine-learning techniques to construct predictors directly from validated miRNA targets. They typically depend on a combination of specific base-pairing rules and conservational analysis to score possible 3'-UTR recognition sites, then enumerate putative gene targets. Note that target predictions based solely on base pairing are subject to false positive hits. It has been estimated that the number of false positive hits can be greatly reduced by limiting hits to only those conserved in other organisms (11,12).

*Related Work on Applying Ontological Techniques into Biological Research*

Ontological techniques have been widely applied to biological research. The most successful example is the Gene Ontology (GO) project (13), which is a major bioinformatics initiative begun in 1998. The GO is a collaborative effort to build consistency of gene product descriptions, with the aim of standardizing the representation of genes across species and databases. Starting from three model organisms, many plant, animal, and microbial genomes have been assimilated into the GO. Consisting of three components, i.e., biological processes, cellular components, and molecular functions, the GO provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data in a species-independent manner, as well as tools to access and process such data. Similarly, the Unified Medical Language System (UMLS) (14) can be viewed as a comprehensive thesaurus and ontology of biomedical concepts.

In (15) M.N. Cantora et al. discuss the issue of mapping concepts in the GO to the UMLS. Such a mapping may allow for the exploitation of the UMLS semantic network to link disparate genes through their annotation in the GO to unique clinical outcomes, potentially uncovering biological relationships.

This study reveals the inherent difficulties in the integration of vocabularies created in different manners by specialists in different fields, as well as the strengths of different techniques used to accomplish such integration.

The National Center for Biomedical Ontology (NCBO) (16) is one of the seven National Centers for Biomedical Computing funded by the NIH Roadmap. Assembling the expertise of leading investigators in informatics, computer science, and biomedicine from across the country, the NCBO aims to support biomedical researchers in their knowledge-intensive work and to provide a Web portal with online tools to enable researchers to access, review, and integrate disparate ontological resources in all aspects of biomedical investigation and clinical practice. A major focus of their work involves the use of biomedical ontologies to aid in the management and analysis of data and knowledge derived from complex experiments.

**Supplementary Data in the Omit Framework**

*Ontology Development*

A particular challenge in performing miRNA target gene acquisition and prediction is to standardize the terminology and to better handle the rich semantics contained explicitly or inexplicitly in large amounts of data. Ontologies can greatly help in this regard. As formal, declarative knowledge representation models, ontologies perform a key role in defining formal semantics in traditional knowledge engineering. Therefore, ontological techniques have been widely applied to biological and biomedical research. There exist a group of well-established biological and biomedical ontologies, such as the GO in Genetics (13), the UMLS in Medicine (14), and the NCBO (16) among others. Unfortunately, there is no ontology that fits the miRNA research by providing biomedical researchers with the desired semantics in miRNA target gene acquisition and prediction. This lack of well-defined semantics necessary for the miRNA research motivates us to construct a domain-specific ontology to connect facts from distributed data sources that may provide valuable clues in identifying target genes for miRNAs of

interest. The proposed OMIT ontology, which is the very first ontology, is an integral component in our framework: it supports terminology standardization, facilitates discussions among the collaborating groups, expedites knowledge discovery, provides a framework for knowledge representation and visualization, and improves data sharing among heterogeneous sources.

Our ontology design methodology is a unique combination of both top-down and bottom-up approaches. First, we adopt a top-down approach driven by domain knowledge and relying on three resources: (i) the GO ontologies (i.e., BiologicalProcess, CellularComponent, and MolecularFunction); (ii) existing miRNA target databases; and (iii) cancer biology experts in our project. In this iterative, knowledge-driven approach, both ontology engineers and domain experts (cancer biologists) are involved, working together to capture domain knowledge, develop a conceptualization, and implement the conceptual model. The top-down development process has taken place over many iterations, involving a series of interviews, exchanges of documents, evaluation strategies, and refinements; and revision-control procedures have been adopted to document the process for future reference. In addition, on a regular basis domain experts together with ontology engineers have fine-tuned the conceptual model (bottom-up) by an in-depth analysis of typical instances in the miRNA domain, for example, *mir-21*, *mir-125a*, *mir-125b*, *mir-19b*, *let-7*, and so on.

There are different formats for describing an ontology, all of which are popular and based on different logics: Web Ontology Language (OWL) (17), Open Biological and Biomedical Ontologies (OBO) (18), Knowledge Interchange Format (KIF) (19), and Open Knowledge Base Connectivity (OKBC) (20). We have chosen the OWL format that is recommended by the World Wide Web Consortium (W3C). OWL is designed for use by applications that need to process the content of information instead of just presenting information to humans. As a result, OWL facilitates greater machine interpretability of Web contents. As for our development tool, we have chosen Protégé (21) over other available tools such as CmapTools and OntoEdit. During development of the ontology, we have observed the seven practices proposed by the OBO Foundry Initiative (22), and we have reused and extended a subset of concepts from the Basic Formal Ontology (BFO) (23) to design top-level

concepts in the OMIT.

It is critical to present related gene information of miRNA targets to medical scientists in order for them to fully understand the biological functions of miRNAs of interest. Therefore, it is necessary to align the OMIT with the GO. Such an alignment (also known as “mapping”) is straightforward due to the fact that we have reused and extended a set of well-established concepts from the GO ontologies. We utilize RIF-PRD (W3C Rule Interchange Format–Production Rules Dialect), an XML-based language, to express such mapping rules so that they can be automatically processed by computers. Compared with SWRL (Semantic Web Rule Language), which was designed as an extension to OWL, RIF-PRD has the following advantages: (i) it supports multiple-arity predicates, whereas SWRL is limited to unary and binary predicates; (ii) it has functions, whereas SWRL is function-free; (iii) it has an extensive set of data types and built-ins, whereas the support for built-ins in SWRL is still under discussion; and (iv) it allows disjunction in rules, whereas SWRL does not.

*First-Version OMIT Ontology*

We have designed nine top-level concepts: *CommonBioConcepts*, *InfoContentEntity*, *MaterialEntity*, *ObjectBoundary*, *ProcessualEntity*, *Quality*, *RealizableEntity*, *SpatialRegion*, *TemporalRegion*, along with some core concepts: *AnatomicFeature*, *Cell*, *Disease*, *ExperimentValidation*, *GeneExpression*, *GeneSequence*, *HarmfulAgent*, *MiRNA*, *MiRNABinding*, *Organism*, *PathologicalEvent*, *Protein*, *SignsOrSympotoms*, *TargetGene*, *TargetPrediction*, *Tissue*, and *Treatment*.

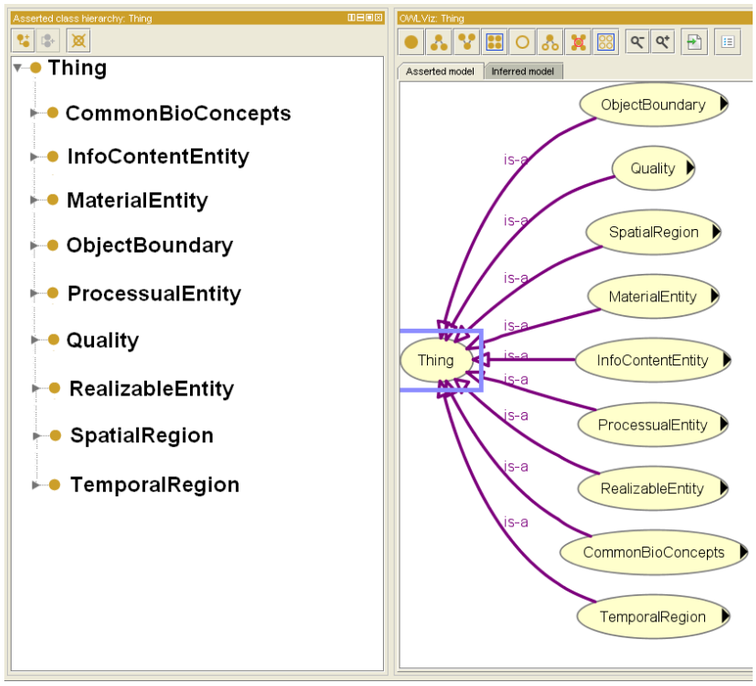


**Fig. S1** Top-level OMIT concepts


**Fig. S2** Expanded view of OMIT concepts (Portion)

Fig. S1 is a screen shot from the Protégé graphical user interface (GUI), demonstrating OMIT top-level concepts. Fig. S2 exhibits some of the core concepts and their subconcepts (also known as subclasses). Note that as mentioned earlier, the core concepts and their respective subconcepts have not been built from scratch. Instead, in order to take advantage of the knowledge contained in existing ontologies and to reduce the possibility of redundant efforts, we have reused and extended a set of well-established concepts from existing ontologies, in particular, the GO ontologies.

*Domain-Specific Knowledge Base*

The OMIT knowledge base is constructed upon a global schema (i.e., the OMIT ontology) and from seven distributed miRNA target databases: miRDB (24), TargetScan (25), PicTar (26), RNAhybrid (27), miRanda (28), miRBase (29), and TarBase (30). The knowledge base development consists of two steps: semantic annotation and data integration. Note that there is no clear boundary between these two steps; instead, they closely intervene with each other throughout the entire knowledge base development process.

Semantic annotation is the process of tagging source files with predefined metadata, which usually consists of a set of ontological concepts. In this paper, our annotation includes annotating both database schemas and their data sets. We refer to such annotation as “deep” annotation–this term was coined by C. Goble in the Semantic Web Workshop of WWW 02, and further investigated in (31,32). It is necessary to annotate more than just database schemas because there are situations where the opposite “shallow” annotation (annotation on schemas alone) cannot provide users with the desired knowledge. Take the schema in miRanda as an example: it combines a total of 172 fields into a single column. If users are only interested in, for example, knowledge pertaining to “AML-HL60” and “Astroblastoma-DD040800” instead of all 172 fields, then it would be extremely troublesome to retrieve the desired data for users if the conventional shallow annotation had been adopted.

Apparently, there exists some tradeoff on choosing shallow or deep annotation. As discussed and analyzed above, it is essential to annotate actual data sets in addition to schemas themselves. However, it is unavoidable for us to spend more time, resources, and human efforts during deep annotation. We believe that the extra cost associated with deep annotation is worthwhile and will pay off in the long run, if a more accurate, easy-to-understand set of retrieved data is preferable.

Our deep annotation takes two successive steps: (i) We annotate the source database schemas using OMIT concepts. During this first-level annotation, we generate a set of mapping rules between OMIT concepts and elements from source database schemas. These mapping rules are specified in the RIF-PRD format. (ii) The next step is to annotate data sets from each source. This second-level annotation is,

in fact, the data integration process.

The first–and the most critical–step in data integration is to specify the correspondence between the source databases and the global schema. This is, in fact, the first step in semantic annotation. According to the analysis in (33), there are two different categories of approaches: local-as-view (LAV) and global-as-view (GAV). In general, processing queries in the LAV approach is more difficult than that in the GAV approach. The knowledge we have regarding the data in the global schema is through the views representing the sources, which provide only partial information about the data. Because the mapping associates to each data source a view over the global schema, it is not trivial to figure out how to use the sources for the purpose of answering queries that are expressed according to the global schema. On the contrary, query processing appears to be easier in the GAV approach, because we may take advantage of the mapping that directly and explicitly specifies which source queries corresponds to the elements of the global schema. As a result, query answering can be carried out through a simple unfolding strategy. However, integrity constraints and system extensibility are two major challenges for the GAV approach.

We have adopted a “GAV-like” approach. Our approach is similar to the traditional GAV approach in that the global schema is regarded as a view over source databases and expressed in terms of source database schemas. On the other hand, our approach differs from the traditional GAV approach in that we include not only a global schema, but aggregated data sets as well. Consequently, the user search/query will be composed according to the concepts in the global schema, and the query answering process will be based on the centralized data sets with an unfolding strategy over the original query.

As illustrated in Fig. 1, an inference engine is integrated with the OMIT knowledge base. Inference engines are also known as ontology reasoners, which provide a more convenient method for querying, manipulating, and reasoning over available data sets. In particular, semantics-based queries, instead of traditional SQL queries, are thus made possible. In this paper, we have utilized the Jena2 OWL reasoner (34), a rule-based implementation of a subset of OWL Full semantics.

We exhibit some excerpts (written in the OWL format) below as an example to demonstrate how

semantics is formally encoded in the knowledge base.

```
1.  <MiRNA rdf:about="#mir-21">
2.  <miRNACompleteName rdf:datatype="&xsd;string">hsa-mir-21</miRNACompleteName>
3.  <miRNADescription rdf:datatype="&xsd;string">Homo sapiens miR-21 stem-loop</miRNADescription>
4.  <miRNASequence rdf:datatype="&xsd;string">UAGCUUAUCAGACUGAUGUUGA</miRNASequence>
5.  <upRegulateEvent rdf:resource="#hepatoCellularCarcinoma"/>
6.  <miRNAFunction rdf:datatype="&xsd;string">cell invasion and tumor metastasis, G1-to-S transition promotes
    cell invasion, migration, and growth via repression of PTEN expression and downstream effects involving the
    phosphorylation of FAK and Akt, and the expression of MMP-2 and MMP-9</miRNAFunction>
7.  <miRNASeedLocation rdf:datatype="&xsd;string">64, 1756, 3941</miRNASeedLocation>
8.  <hasTarget rdf:resource="#PDCD4"/>
9.  <hasTarget rdf:resource="#PTEN"/>
10. <hasValidation rdf:resource="#experiment_ Selbach_2008"/>
11. <hasValidation rdf:resource="#experiment_ Meng_2007"/>
12. <hasPrediction rdf:resource="#YOD1"/>
13. <targetRankInPicTar rdf:datatype="&xsd;string">1</targetRankInPicTar>
14. <targetScoreInPicTar rdf:datatype="&xsd;string">9.14</targetScoreInPicTar>
15. </MiRNA>
```

- Line 1: defines an instance, *mir-21*, for the concept MiRNA
- Line 2: the complete name of *mir-21* is hsa-mir-21
- Line 3: the description of *mir-21* is Homo sapiens miR-21 stem-loop
- Line 4: the sequence of *mir-21* is UAGCUUAUCAGACUGAUGUUGA
- Line 5: *mir-21* is able to up-regulate the hepatocellular carcinoma
- Line 6: the functions of *mir-21* are (i) cell migration, invasion, and tumor metastasis, (ii) G1-to-S cell cycle transition, (iii) promotes cell growth, and (iv) the expression of MMP-2 and MMP-9 via repression of PTEN expression and downstream effects involving the phosphorylation of FAK and Akt

- Line 7: the seed location of *mir-21* is 64, 1756, 3941
- Line 8: *mir-21* has a validated target PDCD4
- Line 9: *mir-21* has a validated target PTEN
- Line 10: validation is contained in experiment_Selbach_2008 (whose details are contained in the knowledge base but not shown here)
- Line 11: validation is contained in experiment_Meng_2007 (whose details are contained in the knowledge base but not shown here)
- Line 12: *mir-21* has a predicted target YOD1
- Lines 13 and 14: such prediction is from PicTar with a rank of 1 and a score of 9.14
- Line 15: the end of the detailed information for the instance mir-21

**Supplementary Data in Discussion**

*Enhanced Knowledge Acquisition from Existing Data Sources*

To manually search candidate target genes for a specific miRNA of interest is not only extremely time-consuming, but also error-prone most of the time. It is well known that each miRNA can have hundreds of possible target genes. Currently, there are many different target prediction databases (16 in total, to the best of our knowledge), which are geographically distributed worldwide and have adopted quite different schemas and terminologies. Moreover, in many cases, additional information for target genes is critical for biologists to fully understand these genes' biological functions. More often than not, such additional information is not available in target prediction databases. Instead, other resources such as the GO ontologies are needed for this purpose.

Taking *mir-21* as an example, miRDB, TargetScan, and PicTar report 348, 210, and 175 target genes for *mir-21*, respectively. It is very challenging, if not impossible, for biologists to manually search a total of 733 candidate target genes, let alone to further search for useful information on each gene hidden in the GO. In fact, the situation could be even worse: biologists usually make use of more than three databases in the miRNA research area. As shown in Fig. 2, the OMIT framework helps

biologists discover miRNAs' candidate target genes in a much more efficient manner: (i) knowledge from seven databases is automatically obtained, integrated, and presented to users; (ii) related information from the GO is provided for each retrieved target gene. In this manner, biologists can save a large amount of time that would have been spent if a manual search were to be carried out.

There are some other research and applications in tackling the challenge of miRNA target gene prediction through computational approaches. GOmir (36) is a typical example, and miRGator (37) is another effort. The main difference between the OMIT framework and these two systems is that, for the very first time we explore a domain-specific knowledge base approach and how it can be used to facilitate knowledge acquisition and sharing in miRNA target gene prediction. Both GOmir and miRGator, on the other hand, utilize traditional relational database techniques. The OMIT is based on a domain-specific ontology, which is a formal, declarative knowledge representation model and therefore performs a key role in knowledge engineering. The advantages of an ontology-based knowledge sharing model over a relational data model are summarized:

- Relational databases focus on syntactic representation of data, lacking the ability to explicitly encode semantics, which is critical in automated knowledge acquisition. For example, it is very hard to use relational databases to represent the subsumption relationship between different concepts, a type of data semantics that is widely adopted by domain experts.
- Powerful tools are available for capturing and managing ontological knowledge, including an abundance of reasoning tools that are readily supplied for ontological models. These ontology reasoners (also known as inference engines) make it much more convenient to query, manipulate, and reason over available data sets. In particular, semantics-based (sometimes known as logic-based) queries, instead of traditional SQL queries, are made possible.
- Advances in the miRNA domain require that changes be made on a regular basis with regard to underlying data models. In addition, more often than not, it is preferable to represent data at different levels and/or with different abstractions. There are no straightforward methods for performing such updates if relational models are adopted.

- Ontologies, along with Semantic Web technologies, better enable miRNA researchers and clinicians to append additional data sets in a more flexible way. More importantly, the formal semantics encoded in ontologies makes it possible to reuse the data in unplanned and unforeseen ways, in particular, in cases where data users are not data producers, which is now very common.

Note that out of the seven miRNA target databases currently integrated in the OMIT knowledge base, the first six (i.e., miRDB, TargetScan, PicTar, RNAhybrid, miRanda, and miRBase) are target prediction databases, whereas the last one (TarBase) is the experimentally validated database. The purpose of adding TarBase is to provide biologists with additional information to assist them in making judgements on the reliability of predicted candidate targets.

*Further Discovery of Hidden Knowledge*

Compared with traditional relational database techniques, the proposed knowledge base created upon a domain-specific ontology has the ability to help users acquire hidden knowledge that was previously implicit and unclear to biologists. Such additional information is obtained through inference engines (also known as ontology reasoners) specifically designed for OWL ontologies, as discussed earlier in this paper. In this section we discuss the following reasoning tasks that are not supported by relational databases.

*Semantic Subsumption Reasoning.* In formal logic, semantic subsumption reasoning checks whether or not it is true that a concept (i.e., class) or a relationship (i.e., property) is subsumed by another concept or relationship. Due to the well-defined concept hierarchy in the ontology, the semantic subsumption relationship can be readily retrieved and integrated into the query/search results before presenting them to users. For example, we obtain the information of "*mir-21* promotes

hepatoCellularCarcinoma" from the knowledge base; at the same time, "hepatoCellularCarcinoma" is defined as an instance of the concept *Carcinoma*, which in turn is a subclass of the concept *MalignantNeoplasm*. Therefore, a new conclusion, "*mir-21* promotes MalignantNeoplasm," is acquired by reasoning on the concept hierarchy. Similarly, another conclusion, "*mir-21* promotes Tumor," can be readily obtained as well. These extra conclusions will help cancer biologists to generalize their findings to more model systems.

*Semantic Contradiction Reasoning.* An ontology reasoner can help check whether or not different components in the current knowledge base (e.g., the concept hierarchy, rules, and instances) are consistent with each other. For example, prior knowledge based on domain expertise tells us that miRNAs always down-regulate instead of up-regulate their direct target genes. Such knowledge can be explicitly expressed as a constraint rule in the ontology. During the construction of our knowledge base, if a newly added instance (sometimes known as an assertion) is going to violate this rule, a warning message will be generated by the inference engine. As shown in Fig. S3, the assertion with a circle and an arrow is such a violation example. Consequently, whenever this type of contradiction is going to take place, the reasoning mechanisms integrated in the knowledge base are able to identify the scenario and prevent this contradiction from happening. Note that incorrect or outdated information contained in original data sources is just one possible cause for this situation, other events leading to such a contradiction include, but are not limited to, (i) human error; (ii) mistakes due to the annotation process; and

(iii) errors happened during the data integration.

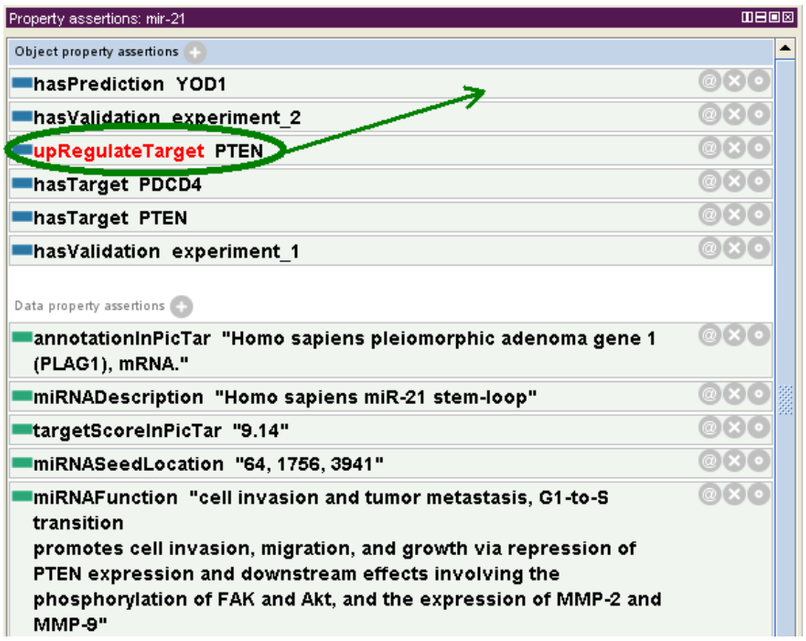


**Fig. S3** Contradiction warning by the inference engine

*Semantic disjointWith Reasoning.* The disjointWith relationship (also known as disjunction) makes the reasoning harder but also more complete because the reasoner needs to check all possible subconcepts or subproperties in disjunctions. Therefore, disjointWith reasoning may provide us with additional knowledge. For example, the following semantics is explicitly encoded in our knowledge base: (i) oncogenic miRNA is a special type of miRNA that promotes the development of some tumor; (ii) any oncogenic miRNA will be involved in at least one of three pathological events, i.e., tumor cell death, tumor cell proliferation, and tumor metastasis; and (iii) the aforementioned three pathological events are disjoint with each other. Reasoning based on such explicit semantics will make it easier for domain experts to identify newly obtained, previously hidden knowledge. Suppose that a specific

oncogenic miRNA is known not to be involved in either tumor cell death or tumor cell proliferation, and biologists are further interested in the following query: *is this miRNA involved in tumor metastasis?* When there exists at least one assertion on the involvement of this miRNA in tumor metastasis, the query answering becomes trivially "yes." Otherwise, it will be a completely different scenario. In the situation where such assertions are not found, without reasoning mechanisms, the conclusion will be "unknown," even under the open world assumption. On the other hand, if disjointWith reasoning is applied, the system is then able to draw a conclusion that the miRNA of interest must be involved in tumor metastasis. In other words, biologists obtain a positive answer to their query even before its direct or indirect support is presented.

The above analysis further explains the motivation to develop a domain-specific ontology in our research. The knowledge base created upon such an ontology not only standardizes the terminology and handles the rich data semantics, but also helps users to better acquire implicit knowledge hidden in the original data.

Ultimately, the acquired candidate target genes for miRNAs of interest will need to be experimentally validated by cancer biologists. Some of our query results have been validated by recent experimental studies. For example, the OMIT system output PTEN, a tumor suppressing gene, as one of *mir-21*'s top candidate target genes, which was experimentally confirmed by Zhang et al. (38). For those computationally predicted targets that have not yet been confirmed, we plan to perform (in our future work) western, northern, luciferase assay, and other experimental assays to validate selected high-impact candidate target genes in cancer cells.

**Future Work**

The most important future work is to continue fine-tuning the first version of our OMIT ontology and to update the knowledge base accordingly. There are ten other miRNA-related databases which we plan to integrate into our knowledge base, i.e., miRGen, DIANA-microT, RNA22, Vir-Mir, ViTa, MicroInspector, miRGator, miTarget, NBmiRTar, and PITA.

There are some inherent disadvantages in our GAV-like approach for data integration. For example, integrity constraints are generally difficult to handle and are therefore usually not taken into consideration. System extensibility is another problem. When source databases change their schemas, or, worse, when a new source database is to be included in the system, the definition of the global schema may be enormously impacted: we may need to redefine some elements and their associated views in the global schema. It is thus interesting for us to exploit the LAV approach as well. Some initial attempts have been reported to compare these two approaches. However, query answering in LAV is known to be harder, and a thorough analysis of the differences and similarities between the GAV and LAV approaches is still missing. Other questions that need further investigation include: the treatment of mutually inconsistent sources; the issue of reasoning on queries (in particular, to check the containment relationship among queries); and the incorporation of the notions of data quality and data cleaning.

Another possible future research direction is to create a centralized RDF data warehouse for data integration. RDF is a standard model recommended by the World Wide Web Consortium (W3C) for data interchange on the Web (35). Due to its ability to facilitate data merging even if the underlying schemas differ from each other, the RDF specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. Being a structure based on the directed acyclic graph (DAG) model, the RDF defines statements about resources and their relationships in triples, each of which consists of a subject, a predicate, and an object. This generic structure of RDF allows structured and semi-structured data to be mixed, exposed, and shared across different applications, and data interoperability is thus made easier to handle.

**SUPPLEMENTARY REFERENCES**

1. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, and Ruvkun G. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403:901-906, 2000.
2. Zeng Y, Yi R, and Cullen BR. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. *Proc. the National Academy of Sciences*, 100:9779-84, 2003.
3. Huang Q, Gumireddy K, Schrier M, le Sage C, Nagel R, Nair S, Egan DA, Li A, Huang G, Klein-Szanto AJ, Gimotty PA, Katsaros D, Coukos G, Zhang L, Puré E, and Agami R. The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis. *Nature Cell Biology*, 10:202-210, 2008.
4. Zhao JJ, Lin J, Yang H, Kong W, He L, Ma X, Coppola D, and Cheng JQ. MicroRNA-221/222 negatively regulates estrogen receptor and is associated with tamoxifen resistance in breast cancer. *The Journal of Biological Chemistry*, 283:31079-86, 2008.
5. Fujita Y, Kojima K, Hamada N, Ohhashi R, Akao Y, Nozawa Y, Deguchi T, and Ito M. Effects of miR-34a on cell growth and chemoresistance in prostate cancer PC3 cells. *Biochemical and Biophysical Research Communications*, 377:114-9, 2008.
6. Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, and Tuschl T. New microRNAs from mouse and human. *RNA*, 9:175-9, 2003.
7. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, and Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34:D140-4, 2006.
8. Northcott PA, Fernandez-L A, Hagan JP, Ellison DW, Grajkowska W, Gillespie Y, Grundy R, Meter TV, Rutka JT, Croce CM, Kenney AM, and Taylor MD. The miR-17/92 polycistron is upregulated in sonic hedgehog-driven medulloblastomas and induced by n-myc in sonic hedgehog-treated cerebellar neural precursors. *Cancer Research*, 69:3249-55, 2009.
9. Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, Lefort K, Dotto GP, and Harshman K. Impact of normalization on miRNA microarray expression profiling. *RNA*, 15:493-501, 2009.
10. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, and Rajewsky N. Widespread

changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58-63, 2008.

11. John B, Enright AJ, Aravin A, Tuschl T, Sander C, and Marks DS. Human microRNA targets. *PLoS Biology*, 2:1862-79, 2004.
12. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, and Burge CB. Prediction of mammalian microRNA targets. *Cell*, 115:787-98, 2003.
13. Gene Ontology Website. http://www. geneontology.org/index.shtml (Accessed in March 2011).
14. Lindberg D, Humphries B, and McCray A. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291, 1993.
15. Cantora MN, Sarkara IN, Gelmana R, Hartelb F, Bodenreiderc O, and Lussiera YA. An evaluation of hybrid methods for matching biomedical terminologies: mapping the gene ontology to the UMLS. *Studies in Health Technology and Informatics*, 95:62-67, 2003.
16. The National Center for Biomedical Ontology. http://www.bioontology.org/ (Accessed in March 2011).
17. OWL. http://www.w3.org/2004/OWL/ (Accessed in March 2011).
18. OBO. http://www.obofoundry.org/ (Accessed in March 2011).
19. KIF. http://logic.stanford.edu/kif/ (Accessed in March 2011).
20. OKBC. http://www.ai.sri.com/okbc/ (Accessed in March 2011).
21. Protégé. http://protege.stanford.edu/ (Accessed in March 2011).
22. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg L, Eilbeck K, Ireland A, Mungall C, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S, Scheuermann R, Shah N, Whetzel P, and Lewis S. The OBO foundry: coordinated evolution of Ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007.
23. BFO. http://www.ifomis.org/bfo/ (Accessed in March 2011).
24. miRDB. http://mirdb.org/miRDB/ (Accessed in March 2011).
25. TargetScan. http://www.targetscan.org/ (Accessed in March 2011).
26. PicTar. http://pictar.mdc-berlin.de/ (Accessed in March 2011).

27. RNAhybrid. http://mirnamap.mbc.nctu.edu.tw/ (Accessed in March 2011).

28. miRanda. http://www.microrna.org/microrna/home.do (Accessed in March 2011).

29. miRBase. http://www.mirbase.org/search.shtml (Accessed in March 2011).

30. TarBase. http://diana.cslab.ece.ntua.gr/tarbase/ (Accessed in March 2011).

31. Handschuh S, Staab S, and Volz R. On deep annotation. *Proc. the Twelfth International World Wide Web Conference (WWW 03)*, Budapest, Hungary, May, 2003.

32. Handschuh S, Volz R, and Staab S. Annotation for the deep Web. *IEEE Intelligent Systems*, 18(5): pp.42-48, 2003.

33. Lenzerini M. Data integration: a theoretical perspective. *Proc. the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 02)*, pp.233-246, Madison, Wisconsin, June, 2002.

34. Jena2 OWL Reasoner. http://jena.sourceforge.net/inference/ (Accessed in March 2011).

35. RDF Website. http://www.w3.org/RDF/ (Accessed in March 2011).

36. Roubelakis MG, Zotos P, Papachristoudis G, Michalopoulos I, Pappa KI, Anagnou NP, and Kossida S. Human microRNA target analysis and gene ontology clustering by GOmir, a novel stand-alone application. *BMC Bioinformatics*, 10 (Suppl 6):S20, 2009.

37. Nam S, Kim B, Shin S, and Lee S. miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Research*, Epub, 1-6, 2007.

38. Zhang JG, Wang JJ, Zhao F, Liu Q, Jiang K, and Yang GH. MicroRNA-21 (miR-21) represses tumor suppressor PTEN and promotes growth and invasion in non-small cell lung cancer (NSCLC). *Clin Chim Acta*, 3:411(11-12):846-52, 2010.