OMIT: Domain Ontology and Knowledge Acquisition in MicroRNA Target Prediction
(Short Paper)

Christopher Townsend[1], Jingshan Huang[1,⋆], Dejing Dou[2], Shivraj Dalvi[1], Patrick J. Hayes[3], Lei He[4], Wen-chang Lin[5], Haishan Liu[2], Robert Rudnick[1], Hardik Shah[1], Hao Sun[6], Xiaowei Wang[7], and Ming Tan[8,⋆⋆]

[1] School of Computer and Information Sciences
University of South Alabama, Mobile, AL 36688, U.S.A.
huang@usouthal.edu
http://cis.usouthal.edu/~huang/
[2] Computer and Information Science Department
University of Oregon, Eugene, OR 97403, U.S.A.
[3] Florida Institute for Human and Machine Cognition
Pensacola, FL 32502, U.S.A.
[4] College of Science and Technology
Armstrong Atlantic State University, Savannah, GA 31419, U.S.A.
[5] Institute of Biomedical Sciences
Academia Sinica, Taipei, Taiwan
[6] Department of Chemical Pathology
Chinese University of Hong Kong, Hong Kong, China
[7] Department of Radiation Oncology
Washington University School of Medicine, St. Louis, MO 63108, U.S.A.
[8] Mitchell Cancer Institute
University of South Alabama, Mobile, AL 36688, U.S.A.
mtan@usouthal.edu
http://southalabama.edu/~tan/

Abstract. The identification and characterization of important roles microRNAs (miRNAs) played in human cancer is an increasingly active area in medical informatics. In particular, the prediction of miRNA target genes remains a challenging task to cancer researchers. Current efforts have focused on manual knowledge acquisition from existing miRNA databases, which is time-consuming, error-prone, and subject to biologists' limited prior knowledge. Therefore, an effective knowledge acquisition has been inhibited. We propose a computing framework based on the Ontology for MicroRNA Target Prediction (OMIT), **the very first** ontology in miRNA domain. With such formal knowledge representation, it is thus possible to facilitate knowledge discovery and sharing from existing sources. Consequently, the framework aims to assist biologists in unraveling important roles of miRNAs in human cancer, and thus to help clinicians in making sound decisions when treating cancer patients.

⋆ Corresponding Author.
⋆⋆ Corresponding Author.

R. Meersman et al. (Eds.): OTM 2010, Part II, LNCS 6427, pp. 1162–1169, 2010.

1 Introduction

Healthcare is a typical area where advances in computing have resulted in numerous improvements. In particular, the identification and characterization of the important roles microRNAs (miRNAs) play in human cancer is an increasingly active area. MiRNAs are a class of small non-coding RNAs capable of regulating gene expression. They have been demonstrated to be involved in diverse biological functions [13,18], and miRNAs' expression profiling has identified them associated with clinical diagnosis and prognosis of several major tumor types [8,15,21]. Unfortunately, the prediction of the relationship between miRNAs and their target genes still remains a challenging task [4,6].

Ontologies are formal, declarative knowledge representation models, playing a key role in defining formal semantics in traditional knowledge engineering. We propose an innovative computing framework (Figure 1) based on the Ontology for MicroRNA Target Prediction (OMIT) to handle the aforementioned challenge. The OMIT is a domain-specific ontology upon which it is possible to facilitate knowledge discovery and sharing from existing sources. As a result, the long-term research objective of the OMIT framework is **to assist biologists in unraveling important roles of miRNAs in human cancer, and thus to help clinicians in making sound decisions when treating patients**.

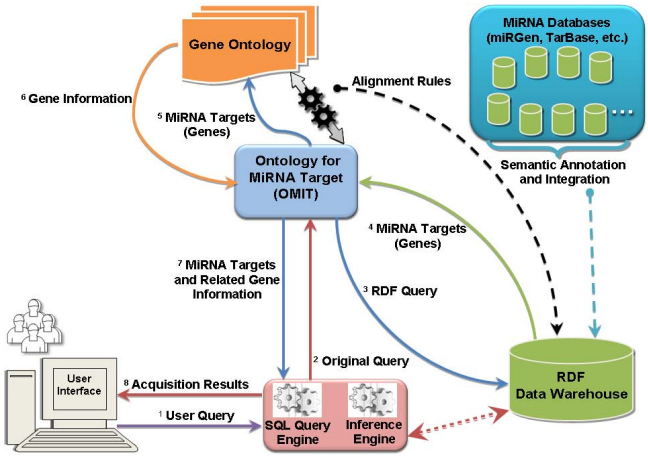

Fig. 1. OMIT System Framework

2 Background and Related Research

2.1 Background Knowledge of Ontologies

Ontology is a computational model of some portion or domain of the world [19]. The model describes the semantics of the terms used in the domain. Ontology is often captured in some form of a semantic network, i.e., a graph whose nodes are

concepts or individual objects and whose arcs represent relationships or associations among the concepts. The semantic network is augmented by properties and attributes, constraints, functions, and rules, which govern the behavior of the concepts. In brief, an ontology consists of a finite set of concepts (also known as "terms" or "classes"), along with these concepts' properties and relationships. In addition, most real-world ontologies have very few or no instances, i.e., they only have the aforementioned graphical structure (also known as "schema"). **Ontology Heterogeneity** is an inherent characteristic of ontologies developed by different parties for the same (or similar) domains. The heterogeneous semantics may occur in two ways. (1) Different ontologies could use different terminologies to describe the same conceptual model. That is, different terms could be used for the same concept, or alternatively, an identical term could be adopted for different concepts. (2) Even if two different ontologies use the same terminology, which itself is almost impossible in the real world, concepts' associated properties and the relationships among concepts are most likely to be different. **Ontology Matching** is short for "Ontology Schema Matching", also known as "Ontology Alignment," or "Ontology Mapping." It is the process of determining correspondences between concepts from heterogeneous ontologies (often designed by distributed parties).

2.2 Ontological Techniques in Biological Research

Ontological techniques have been widely applied to medical and biological research. The most successful example is the Gene Ontology (GO) project [3], which is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The GO 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. The GO's focus is to describe how gene products behave in a cellular context. Unified Medical Language System (UMLS) [22] and the National Center for Biomedical Ontology (NCBO) [10] are two other successful examples in applying ontological techniques into biological research. Besides, efforts have been carried out for ontology-based data integration in bioinformatics.

[1] discusses the issue of mapping concepts in the GO to UMLS. This study reveals the difficulties in the integration of vocabularies created in different manners, and allows for the exploitation of the UMLS semantic network to link disparate genes to clinical outcomes. The authors in [2] adopt the global gene expression profiling to identify the molecular pathways and processes affected upon toxicant exposure. Their work demonstrates that the GO mapping can identify both known and novel molecular changes in the mouse liver. [24] develops a computational approach to analyze the annotation of sets of molecules. The authors reveal trends and enrichment of proteins of particular functions within high-throughput datasets at a higher sensitivity than perusal of endpoint annotations. B. Smith et al. [20] describe a strategy, the Open Biomedical Ontologies (OBO) Foundry initiative, whose long-term goal is that the data

generated through biomedical research should form a single, consistent, cumulatively expanding and algorithmically tractable whole.

3 Methodologies

3.1 Task 1: Domain-Specific Ontology

In order to develop a conceptual model that encompasses the required elements to properly describe medical informatics (especially in human cancer), it is essential to explore and abstract the miRNA data to the semantic level. The design of the OMIT will rely on two resources: existing miRNA databases and domain knowledge from cancer biologists. Besides cancer biology experts in the project team, there are six labs from around the world, (1) Yousef Lab in Israel, (2) DIANA Lab in Greece, (3) Sun Lab in Hong Kong, China, (4) Segal Lab in Israel, (5) Lin Lab in Taiwan, and (6) Wang Lab in St. Louis, MO, that have committed to actively participate in the project by providing original data sets and undertaking an in-depth analysis of integrated data and the query that follows.

3.2 Task 2: Annotation on Source Databases

Semantic annotation is the process of tagging source files with predefined metadata, which usually consists of a set of ontological concepts. We adopt a "deep" annotation that takes two steps. (1) To annotate the source database schemas, resulting in a set of mapping rules specified in the RIF-PRD format between OMIT concepts and elements from source database schemas. (2) To annotate data sets from each source, and the annotated data sets will be published in the resource description framework (RDF) [17]. Being a structure based on the directed acyclic graph model, the RDF defines statements about resources and their relationships in triples. Such generic structure allows structured and semi-structured data to be mixed, exposed, and shared across different applications, and the data interoperability is thus made easier to handle. The annotation outcomes will become the input to the next phase, i.e., data integration.

3.3 Task 3: Centralized RDF Data Warehouse

Instead of a traditional relational data warehouse, we propose to create a centralized RDF data warehouse for the data integration, which better fits the project objective. The first, and the most critical, step is to specify the correspondence between source databases and the global schema. We propose to adopt a "Globle-As-View (GAV)-like" approach. Our approach is similar to the traditional GAV approach [7] 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 also aggregated, global data sets as well. As a result, user query will be composed according to the concepts in the global schema, and the query answering will be based on the centralized data sets with an unfolding strategy.

3.4 Task 4: Query and Search in a Unified Style

When presenting a miRNA of interest, its potential targets can be retrieved from existing miRNA databases. Additional information will be further acquired from the GO, which is critical to fully understand the biological functions of the miRNA of interest. The OMIT system aims to provide users a single search/query engine that takes their needs in a nonprocedural specification format. Such search/query is *unified*, that is, although source miRNA databases are geographically distributed and usually heterogeneous among each other, the OMIT system presents users (biologists) a *uniform* view of such heterogeneous data, along with integrated information from the GO.

4 The OMIT Ontology

4.1 Design Methodology

As mentioned in Section 3, the OMIT ontology design relies on two resources: existing miRNA databases and domain knowledge from cancer biologists. Besides, unlike most existing biomedical ontologies that were developed through a top-down approach, our design methodology is a combination of both top-down and bottom-up approaches. On one hand, existing miRNA databases provide us with a general guideline (top-down) regarding which concepts are of most importance to cancer biologists, as well as these concepts' properties and their relationships among each other; on the other hand, domain experts, together with ontology engineers, can fine tune the conceptual model (bottom-up) by an in-depth analysis of typical instances in miRNA domain, e.g., *miR-21*, *miR-125a*, *miR-125b*, and *let-7*, etc.

There are currently different formats in describing an ontology based on different logics: Web Ontology Language (OWL) [14], Open Biological and Biomedical Ontologies (OBO) [11], Knowledge Interchange Format (KIF) [5], and Open Knowledge Base Connectivity (OKBC) [12]. We choose the OWL format, a standard recommended by the World Wide Web Consortium (W3C) [23]. OWL is designed for use by applications that need to process the content of information instead of just presenting it to humans. As a result, OWL facilitates greater machine interpretability of Web contents. The first version OMIT ontology has been added into NCBO BioPortal [9]. The link to access the OMIT ontology is: http://bioportal.bioontology.org/ontologies/42873.

4.2 The Alignment between the OMIT and the GO

An initial version of the OMIT ontology was designed using Protégé 4.0 [16], with 320 concepts in total, many of which are closely related to three sub-ontologies in the GO, i.e., BiologicalProcess, CellularComponent, and MolecularFunction:

- Some OMIT concepts are directly extended from GO concepts. E.g., OMIT concept *GeneExpression* is designed to describe miRNAs' regulation of gene

expression. This concept is inherited from concept *gene expression* in the BiologicalProcess ontology. This way, subclasses of *gene expression*, such as *negative regulation of gene expression*, are then accessible in the OMIT for describing the negative gene regulation of miRNAs in question.

– Some OMIT concepts are equivalent to (or similar to) GO concepts. For example, OMIT concept *PathologicalEvent* and its subclasses are designed to describe biological processes that are disturbed when a cell becomes cancerous. Although not immediately inherited from any specific GO concepts, these OMIT concepts do match up with certain concepts in the BiologicalProcess ontology. OMIT concepts *TargetGene* and *Protein* are two other examples, which correspond to individual genes and individual gene products, respectively, in the GO.

4.3 Software Implementation

The back end of the OMIT system is implemented in the C# language, following an object-oriented approach. A class diagram is demonstrated in the left portion of Figure 2. We represent each OWL entity of interest, i.e., concepts, object properties, and data properties, as its own class. Our *OWLOntology* class constitutes the external interface to this data structure, and it stores each entity in a private hash table for quick lookup by name.

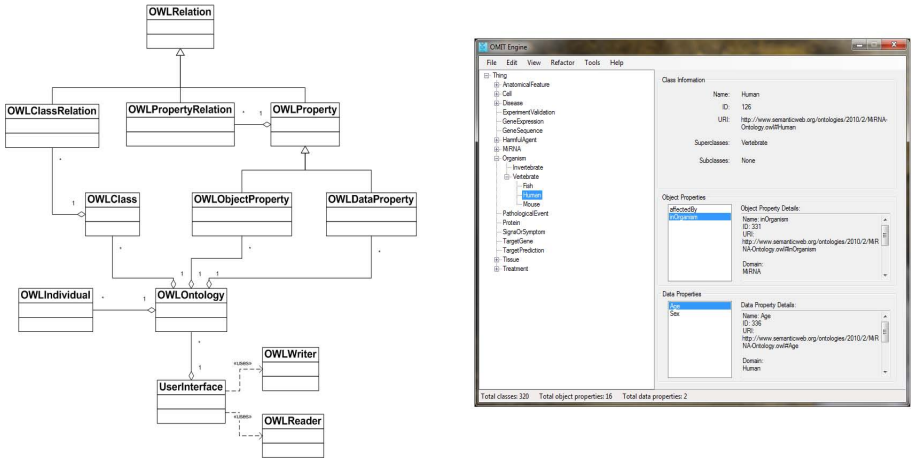

Fig. 2. OMIT Software Class Diagram and GUI Design

We have developed our own parser for OWL ontology files based on the built-in XML parsing capabilities in C#. The right portion of Figure 2 shows a friendly GUI when running our code on the OMIT ontology. In addition, we have deployed the project website (Figure 3) at http://omit.cis.usouthal.edu/, which features an interactive online discussion forum in addition to other materials, e.g., publications, software and tools, and data sets, etc.

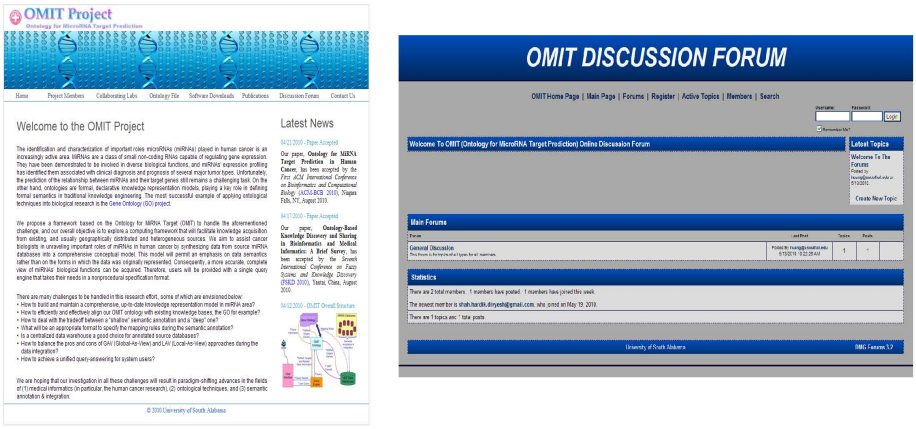

Fig. 3. Project Website Homepage and Interactive Online Discussion Forum

5 Conclusions

We propose an innovative computing framework based on the miRNA-domain-specific ontology, OMIT, to handle the challenge of predicting miRNAs' target genes. The OMIT framework is designed upon *the very first* ontology in miRNA domain. It will assist biologists in better discovering important roles of miRNAs in human cancer, and thus help clinicians in making sound decisions when treating cancer patients. Such long-term research goal will be achieved via facilitating knowledge discovery and sharing from existing sources. In this work-in-progress paper, we first discuss proposed approaches and anticipated challenges in the OMIT framework; then our efforts have focused on the development of a domain ontology. We adopt a unique combination of both top-down and bottom-up approaches when designing the OMIT ontology, whose first version has been added into NCBO BioPortal. Future investigation will be carried out according to the research tasks defined in the framework.

References

1. Castano, S., Ferrara, A., Montanelli, S.: H-MATCH: An Algorithm for Dynamically Matching Ontologies in Peer-based Systems. In: Proc. the first VLDB International Workshop on Semantic Web and Databases, SWDB 2003 (2003)
2. Currie, R., Bombail, V., Oliver, J., Moore, D., Lim, F., Gwilliam, V., Kimber, I., Chipman, K., Moggs, J., Orphanides, G.: Gene ontology mapping as an unbiased method for identifying molecular pathways and processes affected by toxicant exposure: application to acute effects caused by the rodent non-genotoxic carcinogen diethylhexylphthalate. Journal of Toxicological Sciences 86, 453–469 (2005)
3. Gene Ontology Website (August 2010), `http://www.geneontology.org/index.shtml`

4. Hsu, S., Chu, C., Tsou, A., Chen, S., Chen, H., Hsu, P., Wong, Y., Chen, Y., Chen, G., Huang, H.: miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. Nucleic Acids Research 36(D), 165–169 (2008)
5. KIF (August 2010), http://logic.stanford.edu/kif/
6. Kim, S., Nam, J., Lee, W., Zhang, B.: miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics 7(411) (2006)
7. Lenzerini, M.: Data Integration: A Theoretical Perspective. In: Proc. the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2002) (June 2002)
8. Nakajima, G., Hayashi, K., Xi, Y., Kudo, K., Uchida, K., Takasaki, K., Ju, J.: Non-coding microRNAs hsa-let-7g and hsa-miR-181b are associated with chemoresponse to S-1 in colon cancer. Cancer Genomics and Proteomics 3, 317–324 (2006)
9. NCBO BioPortal (August 2010), http://bioportal.bioontology.org/
10. NCBO Website (August 2010), http://www.bioontology.org/
11. OBO (August 2010), http://www.obofoundry.org/
12. OKBC (August 2010), http://www.ai.sri.com/~okbc/
13. Olsen, P., Ambros, V.: The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev. Biology 216, 671–680 (1999)
14. OWL (August 2010), http://www.w3.org/TR/owl-features/
15. Pradervand, S., Weber, J., Thomas, J., Bueno, M., Wirapati, P., Lefort, K., Dotto, G., Harshman, K.: Impact of normalization on miRNA microarray expression profiling. RNA 15, 493–501 (2009)
16. Protégé Website (August 2010), http://protege.stanford.edu/
17. RDF Website (August 2010), http://www.w3.org/RDF/
18. Reinhart, B., Slack, F., Basson, M., Pasquinelli, A., Bettinger, J., Rougvie, A., Ruvkun, G.: The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403, 901–906 (2000)
19. Singh, M., Huhns, M.: Service-Oriented Computing - Semantics, Processes, Agents, 1st edn. Wiley, Chichester (2005)
20. 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., Lewis, S.: The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25(11), 1251–1255 (2007)
21. Sorrentino, A., Liu, C., Addario, A., Peschle, C., Scambia, G., Ferlini, C.: Role of microRNAs in drug-resistant ovarian cancer cells. Gynecologic Oncology 111, 478–486 (2008)
22. UMLS (August 2010), http://www.nlm.nih.gov/research/umls/
23. W3C (The World Wide Web Consortium) (August 2010), http://www.w3.org/
24. Wolting, C., McGlade, C., Tritchler, D.: Cluster analysis of protein array results via similarity of Gene Ontology annotation. BMC Bioinformatics 7(338) (2006)