# Semi-Automated microRNA Ontology Development based on Artificial Neural Networks

Jingshan Huang
School of Computing
University of South Alabama
Mobile, Alabama 36688-0002
Email: huang@southalabama.edu

Jiangbo Dang
Corporate Technology
Siemens Corporation
Princeton, New Jersey 08540-6632
Email: jiangbo.dang@siemens.com

Xingyu Lu
School of Computing
University of South Alabama
Mobile, Alabama 36688-0002
Email: xl1001@jagmail.southalabama.edu

Min Xiong
School of Computing
University of South Alabama
Mobile, Alabama 36688-0002
Email: mx1201@jagmail.southalabama.edu

William T. Gerthoffer
College of Medicine
University of South Alabama
Mobile, Alabama 36688-0002
Email: wgerthoffer@southalabama.edu

Ming Tan
Mitchell Cancer Institute
University of South Alabama
Mobile, Alabama 36604-1405
Email: mtan@usouthal.edu

***Abstract*—microRNAs (miRNAs) are special non-coding RNAs that perform important roles through their target genes. Biologists' conventional miRNA knowledge discovery is time-consuming, labor-intensive, and error-prone. Semantic technologies, which are created upon domain ontologies, can greatly enhance miRNA knowledge discovery. Unfortunately, yet no specific miRNA domain ontologies currently exist. It thus motivates the construction of a miRNA ontology. In addition, a manual ontology development has many drawbacks. We present in this paper a semi-automated ontology development methodology. The developed ontology is the very first one of its kind that formally encodes miRNA domain knowledge. It aims to provide data exchange standards and common data elements and thus help identify novel data connections among heterogeneous sources.**

***Keywords*—*Semi-automated ontology development, microRNA ontology, artificial neural networks.***

## I. Introduction

microRNAs (miRNAs) have been indicated to perform critical roles in biological processes by regulating respective target genes ([1], [2]). To fully delineate miRNA functions, conventionally, biologists need to search both biologically validated miRNA targets (e.g., from PubMed [3] and TarBase [4]) and computationally predicted targets (from target prediction databases, e.g., miRDB [5] and miRGator [6]). Biologists also need to search additional information for all miRNA targets with regard to their associated messenger RNA molecules, related protein functions, and affiliated signaling pathways. Not only biologists are required to manually explore large amounts of data sources, but also more importantly, these involved data sources are semantically heterogeneous among each other. Therefore, significant barriers exist during conventional miRNA knowledge discovery.

Semantic technologies can greatly help in this regard. Gene Ontology (GO) [7] and UniProt [8] Consortiums are two successful examples among others. Semantic technologies are based on domain ontologies; unfortunately, yet no specific ontologies currently exist to describe miRNA entities and their relationships. It thus motivates the construction of a miRNA domain ontology. Despite the fact that it is essential to have a manual component contributed by domain experts when building ontologies, prior research ([9], [10]) has demonstrated that a completely manual ontology development has many drawbacks. We thus propose a semi-automated methodology to construct a miRNA ontology.

The rest of this paper is organized as follows. Section II briefly summarizes research in (semi)automated ontology development; Section III describes our methodology in detail (with an emphasis on a metadata/ontology alignment algorithm); Section IV reports experimental results; and Section V concludes with future research directions.

## II. Related Work

(Semi)automated ontology development has attracted a large amount of research. Existing algorithms can be divided into three categories: translation-based (e.g., [11], [12]), mining-based (e.g., [13], [14]), and external knowledge-based (e.g., [15], [16]) algorithms. Despite its importance, much more progress is still needed. In particular, while *isa* is the most common and critical ontological relationship, the importance of other relationships, especially those domain-dependent ones, has been historically underestimated in many state-of-the-art algorithms.

## III. Methodology

Our semi-automated ontology development process consists of four steps: to develop a "backbone" ontology; to extract metadata from various sources; to align these metadata with the backbone ontology; and to augment the backbone ontology.

### A. Backbone Ontology Development

The development of an initial backbone ontology was driven by domain knowledge, provided by two experimental biologists (both are co-authors of this paper). Two types of data sources were made use of. (i) Popular upper ontologies, Basic Formal Ontology (BFO) [17] for example. They provide

978-1-4799-1310-7/13/$31.00 ©2013 IEEE

general concepts that are the same across all knowledge domains. (ii) Existing bio-ontologies, GO, Foundational Model of Anatomy (FMA) [18], and Human Disease Ontology (HDO) [19] for example. By reusing and extending well-established concepts from these developed bio-ontologies, redundant efforts can be effectively reduced. In particular, special attention was placed on bio-ontologies under the Open Biological and Biomedical Ontologies (OBO) Library [20], an umbrella for ontologies shared across different biological and biomedical domains.

Seven practices proposed by OBO Foundry Initiative [21] have been observed. The ontology development procedure consists of three steps:

1) Computer scientists work together with domain experts (experimental biologists) to specify the range of concepts to be included in the ontology.
2) Definitions of these identified concepts are formalized (using Description Logic) and documented.
3) Concepts along with their properties and relationships are then implemented in a computer language.

We have chosen Web Ontology Language (OWL) [22] format that is recommended by the World Wide Web Consortium (W3C). As for the development tool, we have chosen Protégé.

### B. Metadata Extraction

Metadata were extracted from various data sources, including existing bio-ontologies (e.g., GO, FMA, and HDO) and numerous miRNA target prediction databases (e.g., TargetScan [23], miRDB, and miRGator). These sources are either ontologies or relational databases; therefore, extracting metadata was straightforward.

### C. Metadata/Ontology Alignment

We designed an algorithm to align extracted metadata with the backbone ontology, and the alignment results are equivalent concept pairs between different metadata/ontologies. The algorithm is based on machine-learning techniques.

Given a pair of metadata/ontologies, it is reasonable to assume that contributions from different semantic aspects (i.e., concept names, concept properties, and various relationships) would hold across and therefore be independent of specific concepts. In fact, these contributions are characteristics of specific metadata/ontologies (viewed as a whole) and thus become the foundation for corresponding semantic weights. In other words, during the metadata/ontology alignment, semantic weights are determined by respective metadata/ontologies rather than individual concepts. It is thus possible to learn these weights for all concepts by training examples from a subset of concepts.

*1) Semantic similarity:* Given a pair of concepts, $C_1$ and $C_2$, a total of four semantic similarity measures were designed.

$s_1$ represents the similarity on concept names. Upon completion of some pre-processing (e.g., the removal of hyphens and underscores, transforming nouns from plural forms to single forms), if two names have an exact string matching or are synonyms of each other in WordNet[24] then $s_1$ has a value of 1; otherwise, $s_1$ is calculated according to the edit distance between two strings.

$s_2$ represents the similarity on concept properties, calculated by the percentage of matched properties between $C_1$ and $C_2$. Many domain-dependent properties specifically designed for miRNA field were considered, such as *cellLines*, *chromosomeLocation*, *miRNATargetSequence*, *miRNATargetGeneSymbol*, and *miRNATargetCompleteName*.

$s_3$ represents the similarity on *isa*, the most common and critical domain-independent relationship. First, two ancestor lists, ancestor concepts of $C_1$ and ancestor concepts of $C_2$, are calculated. Pairwise matching is then performed among concepts from these two lists using the "stable marriage" principle. Finally, an average value is calculated as $s_3$ between $C_1$ and $C_2$.

$s_4$ represents the similarity on *hasBinding*, a domain-dependent relationship specifically designed for miRNA field, calculated by the percentage of matched concepts between *hasBinding* concepts of $C_1$ and *hasBinding* concepts of $C_2$.

*2) Learning problem and ANN design:* After four similarity values are obtained, an overall similarity, $s_{overall}$, between two concepts is calculated as the weighted sum of $s_i's$, i.e., $s_{overall} = \vec{w} \cdot \vec{s} = \sum_{i=1}^{4}(w_i \cdot s_i)$. A matrix of the overall similarity between pairwise concepts can then be created. Initially, $w_1$ through $w_4$ are randomly set to some values. We utilize an artificial neural network (ANN) to learn optimal weights: the learning task is to discover equivalent concept pairs and the training experience is a set of equivalent concept pairs provided by biologists.

A two-layer, $4\times1$ ANN (Fig. 1) is designed for this learning problem, and the hypothesis space is a four-dimensional space consisting of various weights (i.e., a set of weight vectors). Gradient descent (delta rule) is adopted as the training rule to find the weight vector ($\vec{w}$) that best fits training examples, and the search strategy within the hypothesis space is to find the vector that minimizes the training error.

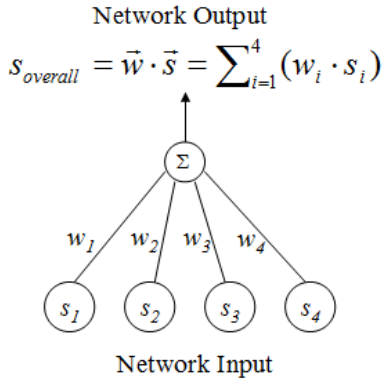


Fig. 1. ANN designed for the learning problem.

*3) Agglomerative clustering procedure:* Once the updated weight vector is obtained from the ANN, the overall similarity matrix is recalculated with learned, optimal weights. An agglomerative clustering procedure is then performed to generate equivalent concept pairs. Each concept is regarded as a *singleton* cluster, and clusters of two equivalent concepts

can be merged with each other and form a new cluster. New clusters continue to be generated until the maximum similarity between any two concepts is below a predefined threshold. Finally, newly generated clusters are output as the set of equivalent concept pairs.

### D. Backbone Ontology Augmentation

According to obtained equivalent concept pairs, it is straightforward to append additional entities (i.e., concepts along with their descendant concepts, properties, relationships, and possible instances) from one metadata/ontology into another one. The initial backbone ontology can then be augmented by entities from other metadata/ontologies.

## IV. Experimental Results

### A. Experimental setup

The backbone ontology contains a total of 53 concepts, 12 properties, and 17 relationships (besides *isa*).

All 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.

### B. Backbone ontology

(1) Example concepts include *MiRNA*, *GeneExpression* (imported from GO), *Organ* (imported from FMA), *Tumor* (imported from HDO), *object*, *materail_entity*, *independent_continuant*, *continuant*, and *entity* (the last five were imported from BFO).

(2) Example properties include *cellLines*, *chromosomeLocation*, *directSupport*, *experimentSummary*, *miRNACompleteName*, *miRNASequenceLocation*, *targetGeneSymbol*, *targetPrimaryPeptideSequence*, and *targetTermAssociations*. These properties were all specifically designed for miRNA field.

(3) Example relationships include:

- *isa* (miRNAs are further divided into oncogenic miRNAs, tumor-suppressing miRNAs, and other miRNAs)
- *affectsTumor* (miRNAs affect numerous tumors, including cancers)
- *hasBinding* (each miRNA has some binding sites)
- *hasPrediction* (each miRNA has one or more computationally predicted target genes)
- *hasTarget* (each miRNA has one or more target genes)
- *hasValidation* (each miRNA has one or more biological validation for each of its target genes)
- *involvedInEvent* (miRNAs are involved in some pathological events)
- *regulateEvent* (miRNAs can down-regulate or up-regulate some pathological events)

Except for *isa*, all other example relationships listed above were specifically designed for miRNA field.

Fig. 2 demonstrates a subset of various relationships designed for *MiRNA*, the most important concept in the backbone ontology. Note that concepts inside dotted rectangles were imported from GO, FMA, and HDO, respectively.

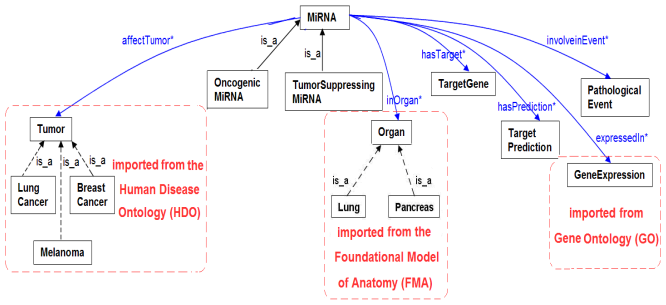


Fig. 2. A subset of relationships designed for concept *MiRNA*.

### C. Metadata/ontology alignment results

We chose three metadata/ontologies to evaluate the alignment algorithm: System Biology Ontology (SBO) [25], Gene Regulation Ontology (GRO) [26], and TarBase. The alignment algorithm was performed between pariwise metadata/ontologies among SBO, GRO, TarBase, and the backbone ontology, resulting in a total of six sets of experiments. Experimental results are analyzed below.

(1) Each of the four semantic weights was initialized to 0.25 in all six sets, and all weights converged to certain values in each set.

(2) Different pairs of metadata/ontologies had different learned weights because weights reflected intended meanings encoded by original metadata/ontology developers.

(3) Four commonly adopted measures were utilized to evaluate equivalent concept pairs output from the algorithm:

- Precision ($p$): the percentage of correct output equivalent concept pairs (those agreed by experimental biologists) over all output pairs, representing the correctness aspect of the alignment algorithm.
- Recall ($r$): the percentage of correct output equivalent concept pairs over actually equivalent pairs, estimating the completeness aspect of the alignment algorithm.
- F-Measure ($f$): also known as *Harmonic Mean* and calculated as $f = \frac{2rp}{r+p}$, aiming to consider both Precision and Recall measures.
- Overall ($o$): a measure calculated as $o = r(2 - \frac{1}{p})$, focusing on the post-alignment effort.

Table I demonstrates our experimental results in these four measures.

(4) Human efforts were significantly reduced. The percentage of training examples provided by biologists over actually equivalent concept pairs was 9%, 28%, 25%, 12%, 15%, and 5% in each of six sets, respectively. In other words, human labor only played a small portion during the semi-automated ontology development.

TABLE I. EVALUATATION OF EQUIVALENT CONCEPT PAIRS

| Experiment Set | Precision | Recall | F-Measure | Overall |
|---|---|---|---|---|
| GRO vs. SBO | 80.39% | 78.85% | 79.61% | 59.62% |
| GRO vs. TarBase | 83.33% | 71.43% | 76.92% | 57.14% |
| SBO vs. TarBase | 71.43% | 62.50% | 66.67% | 37.50% |
| GRO vs. Backbone | 84.62% | 80.49% | 82.50% | 65.85% |
| SBO vs. Backbone | 77.78% | 80.77% | 79.25% | 57.69% |
| TarBase vs. Backbone | 87.50% | 84.48% | 85.96% | 72.41% |
| Average | 80.84% | 76.42% | 78.49% | 58.37% |

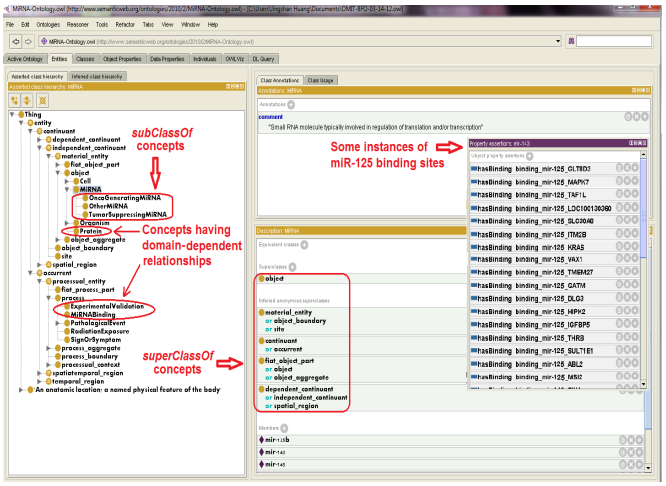


Fig. 3. The resultant miRNA ontology exhibited in Protégé, when concept *MiRNA* was selected.

### D. Resultant miRNA ontology

The resultant miRNA ontology contains a total of 517 concepts, 69 properties, and 93 relationships (besides *isa*). Compared with the backbone ontology, 464 concepts, 57 properties, and 76 relationships were added, all of which were augmented through the proposed algorithm. Note that the number of newly added concepts was much larger than that of equivalent concept pairs output from the alignment algorithm because as discussed earlier in Section III, descendant concepts were added along with identified equivalent concepts. This way, human efforts in developing domain ontologies can be significantly reduced. Additionally, the ontology also contains 237 instances as well. Fig. 3 exhibits a screen shot from Protégé graphical user interface (GUI), with concept *MiRNA* selected. Besides the ontology structure such as concept names and *isa* (i.e., *superClass/subClass*) relationships, some instances of miR-125 binding sites are also shown in the figure. The ontology is includeded in National Center for Biomedical Ontology (NCBO) BioPortal and can be downloaded from http://bioportal.bioontology.org/ontologies/OMIT.

## V. CONCLUSION

Biologists' conventional miRNA knowledge discovery is challenging: time-consuming, labor-intensive, and error-prone. Semantic technologies can significantly help in this regard but no specific miRNA domain ontologies currently exist. Moreover, a manual ontology development has many drawbacks. We presented in this paper a semi-automated ontology development methodology. Experiments have been conducted to evaluate our methodology. We not only developed the very first miRNA-specific ontology but also explored an effective machine-learning alignment algorithm to reduce human efforts in the ontology development. Future research includes to refine the ontology and to use the ontology to annotate data sources related to miRNAs and their biological functions.

## REFERENCES


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

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

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

[4] TarBase. [Online]. Available: http://diana.cslab.ece.ntua.gr/tarbase/

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

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

[7] GO. [Online]. Available: http://www.geneontology.org

[8] UniProt. [Online]. Available: http://www.uniprot.org/

[9] E. Ratsch, J. Schultz, J. Saric, P. C. Lavin, U. Wittig, U. Reyle, and I. Rojas, "Developing a proteininteractions ontology," *Comp Funct Genomics*, vol. 4, no. 1, pp. 85–89, 2003.

[10] H. Pinto and J. Martins, "Ontologies: How can they be built?" *Knowledge and Information Systems*, vol. 6, no. 4, pp. 441–464, 2004.

[11] D. Gasevic, D. Djuric, V. Devedzic, and V. Damjanovic, "From UML to ready-to-use OWL ontologies," in *Proc. 2nd International IEEE Conference Intelligent Systems, ICIS 04*, June 2004, pp. 485–490.

[12] A. Pivk, "Automatic ontology generation from Web tabular structures," *AI Communications*, vol. 19, no. 1, pp. 83–85, January 2006.

[13] B. Biebow and S. Szulman, "TERMINAE: a linguistics-based tool for the building of a domain ontology," in *Proc. 11th European Workshop on Knowledge Acquisition, Modeling and Management, EKAW 99*, Dagstuhl Castle, Germany, May 1999, pp. 49–66.

[14] T. Wachter and M. Schroeder, "Semi-automated ontology generation within OBO-Edit," *Bioinformatics*, vol. 26, no. 12, pp. 88–96, 2010.

[15] D. Moldovan and R. Girju, "Domain-specific knowledge acquisition and classification using Wordnet," in *Proc. 13th Intl'l Florida Artificial Intelligence Research Society Conference*, Orlando, 2000, pp. 224–228.

[16] M. Cho, H. Kim, and P. Kim, "A new method for ontology merging based on concept using Wordnet," in *Proc. 8th International Conference on Advanced Communication Technology, ICACT 06*, Phoenix Park, Korea, February 2006, pp. 1573–1576.

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

[18] FMA. [Online]. Available: http://sig.biostr.washington.edu/projects/fm/

[19] HDO. [Online]. Available: http://purl.bioontology.org/ontology/DOID

[20] OBO Library. [Online]. Available: http://www.obo.sourceforge.net/

[21] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. Goldberg, K. Eilbeck, A. Ireland, C. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. Sansone, R. Scheuermann, N. Shah, P. Whetzel, and S. Lewis, "The OBO foundry: coordinated evolution of Ontologies to support biomedical data integration," *Nat Biotechnol*, vol. 25, no. 11, pp. 1251–1255, November 2007.

[22] OWL. [Online]. Available: http://www.w3.org/2004/OWL/

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

[24] WordNet. [Online]. Available: http://wordnet.princeton.edu/

[25] SBO. [Online]. Available: http://www.ebi.ac.uk/sbo/main/

[26] GRO. [Online]. Available: http://www.ebi.ac.uk/Rebholz-srv/GRO