A domain ontology for the non-coding RNA field

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

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

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

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

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

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

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

Michael T. Zimmermann
Mayo Clinic College of Medicine
Rochester, Minnesota 55905, U.S.A.

Guoqian Jiang
Mayo Clinic College of Medicine
Rochester, Minnesota 55905, U.S.A.

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

Bin Wu
First Affiliated Hospital, Kunming Medical University
Kunming, Yunnan 650032, China

Yongqun He
University of Michigan
Ann Arbor, Michigan 48109, U.S.A.

Shaojie Zhang
University of Central Florida
Orlando, Florida 32816, U.S.A.

Xiaowei Wang
Washington University School of Medicine
St. Louis, Missouri 63108, U.S.A.

He Zhang
University of South Alabama
Mobile, Alabama 36688, 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.

***Abstract*—Identification of non-coding RNAs (ncRNAs) has been significantly enhanced due to the rapid advancement in sequencing technologies. On the other hand, semantic annotation of ncRNA data lag behind their identification, and there is a great need to effectively integrate discovery from relevant communities. To this end, the Non-Coding RNA Ontology (NCRO) is being developed to provide a precisely defined ncRNA controlled vocabulary, which can fill a specific and highly needed niche in unification of ncRNA biology.**

***Keywords*—non-coding RNA, biological and biomedical ontologies, domain ontology, ontology development, semantic data annotation and integration.**

I. Introduction

Non-coding RNAs (ncRNAs) are special functional RNA molecules that are not translated into protein, including transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), long ncRNAs (lncRNAs), and microRNAs (miRNAs), etc. Research interest in ncRNA biology has significantly grown, and a large amount of information has been continuously obtained thanks to rapidly developed sequencing technologies in recent years. Unfortunately, semantic annotation and integration of data about ncRNAs lag behind identification of ncRNAs; therefore, effective methodologies are needed to bring together discovery made by the ncRNA research community.

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

Emerging semantic technologies have been successfully applied to promote more precise communication among scientists in biological, biomedical, and clinical domains [1–6]. In particular, the Open Biological and Biomedical Ontologies (OBO) Library [7] has served as an umbrella for different bio-ontologies shared across various domains. However, the OBO Library does not include comprehensive ontologies targeted for the ncRNA domain. Likewise, no such ncRNA ontologies are found in the National Center for Biomedical Ontology (NCBO) BioPortal [8] either.

The Non-Coding RNA Ontology (NCRO), to be described in this paper, is the *very first* comprehensive domain ontology specifically designed for the ncRNA field. The controlled vocabulary that is precisely defined in the NCRO can be utilized as a resource to annotate and integrate ncRNA data generated by relevant communities. In the sense of semantic data annotation and integration, the NCRO is meant to fill a specific and highly needed niche in comprehensive unification of ncRNA biology.

The rest of this paper is organized as follows. Section II briefly summarizes state-of-the-art research work in both ncRNA biology and bio-ontologies; Section III introduces design principles, languages, and tools for the NCRO ontology development; Section IV describes greater details of NCRO terms and relations; and finally, Section V concludes with future research directions.

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

TABLE I. A SUBSET OF TERMS IN THE NCRO

Term	Explanation
IRES	A sequence element that recruits a ribosomal subunit to internal mRNA for translation initiation.
lncRNA	An non-protein coding transcript (longer than 200 nt).
miRNA	A small (22nt) RNA molecule that is the endogenous transcript of a miRNA gene.
promoter_of_miRNA	The promoter to start the miRNA transcription.
riboswitch	Part of mRNA, acting as a direct sensor of small molecules to control their own expression.
ribozyme	An RNA molecule with catalytic activity.
rRNA	Part of ribosome, providing structural scaffolding and catalytic activity.
snRNA	A small nuclear RNA that is involved in pre-mRNA splicing and processing.
sRNA	A small ncRNA molecule of 50-250 nt.
transcription_of_miRNA	The transcription process of a miRNA.

II. RELATED WORK

A. Research in ncRNA biology

Abnormal expression of ncRNAs is involved in many human diseases [9] [10]. When differentially expressed ncRNAs play regulatory roles in altering target gene expression, further phenotypic effects can be realized. Differential expression of ncRNAs in malignant tissues compared with normal tissues can be exploited as potential therapeutic targets for cancer therapy or as biomarkers used for diagnosis, prediction of patient outcome, or monitoring the effectiveness of cancer therapeutics [11]. In recent years serious attempts have been made to effectively deliver ncRNA into tumors in animal models [12–14].

B. Research in bio-ontologies

RNA Ontology (RNAO) [15]: 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.

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.

Ontology for microRNA Target (OMIT): OMIT [17] [18] is a miRNA domain ontology being developed as part of the NIH OmniSearch project. The purpose is to establish standard metadata in miRNA domain for more effective identification of miRNAs' roles in various human diseases.

Sequence Ontology (SO) [19]: 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) [20]: 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.

III. THE DEVELOPMENT OF THE NCRO

A. Development principles

In the development pipeline for the NCRO, we have observed a set of practices proposed by the OBO Foundry Initiative [21] [22]. For example, the ontology should be: freely available; expressed in a standard language; documented for successive versions; orthogonal to existing ontologies; including natural language specifications; developed collaboratively; and used by multiple researchers.

B. BFO-compliant ontology

All NCRO terms descend from terms defined in the Basic Formal Ontology (BFO) [23]. The BFO is a small, upper-level ontology that is designed for use in supporting information retrieval, analysis, and integration in scientific and other domains. Because the BFO is a well-established upper ontology adopted by all OBO ontologies, our strategy to make the NCRO a BFO-compliant ontology will set the stage for interoperability between the NCRO and other OBO ontologies.

C. Ontology languages and development tools

Both the OBO and Web Ontology Language (OWL) [24] formats have been chosen to describe the ontology. We used OBO-Edit [25] to generate an OBO version of the ontology file in the first place; we then utilized the obo-release-manager (OORT) tool [26] to convert the ontology file into an equivalent OWL version; finally we verified the converted ontology in Protégé [27].

IV. NCRO TERMS AND RELATIONS

While greater details of all terms and relations in the NCRO are publicly available [28] [29], the ontology design is exhibited as follows.

- Table I presents a subset of representative terms defined in the NCRO.

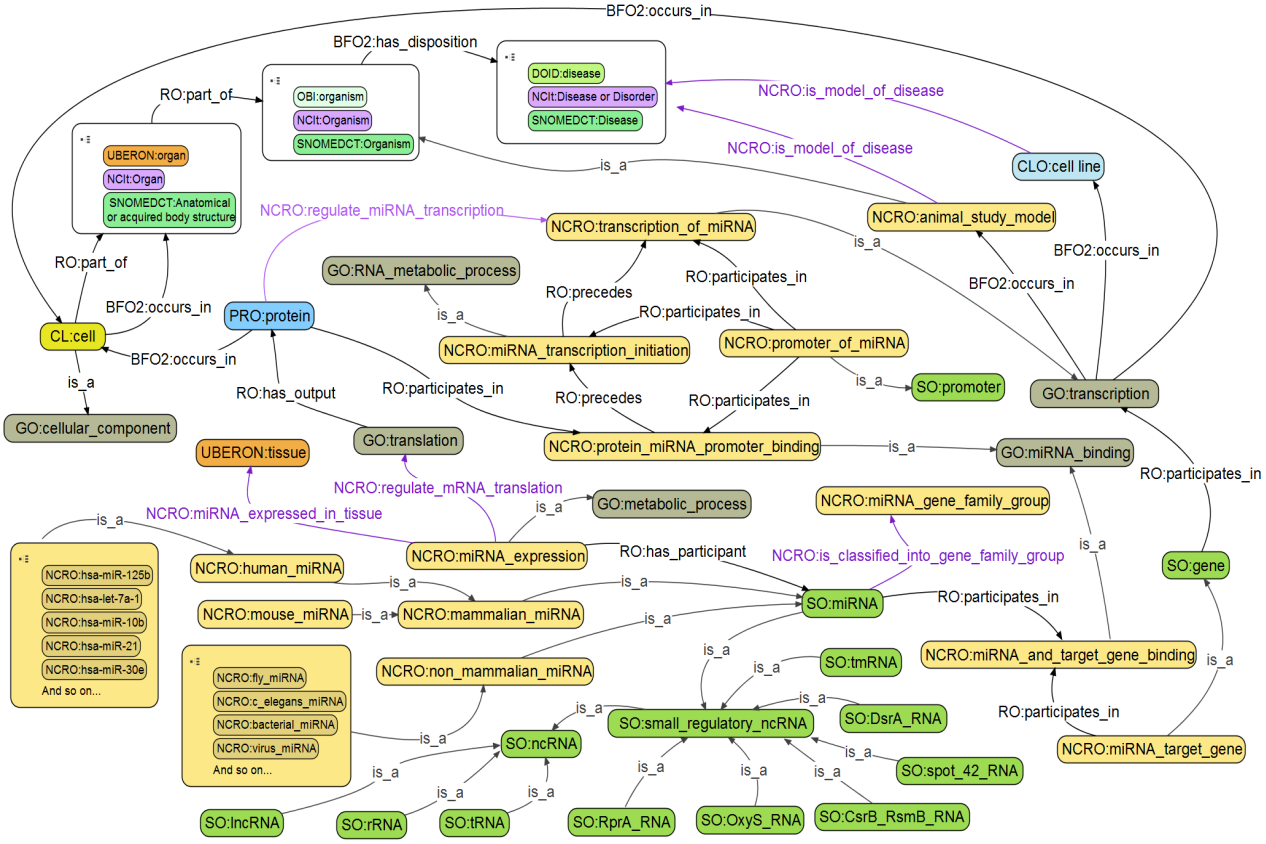

Fig. 1. Core terms and relations in the NCRO ontology.

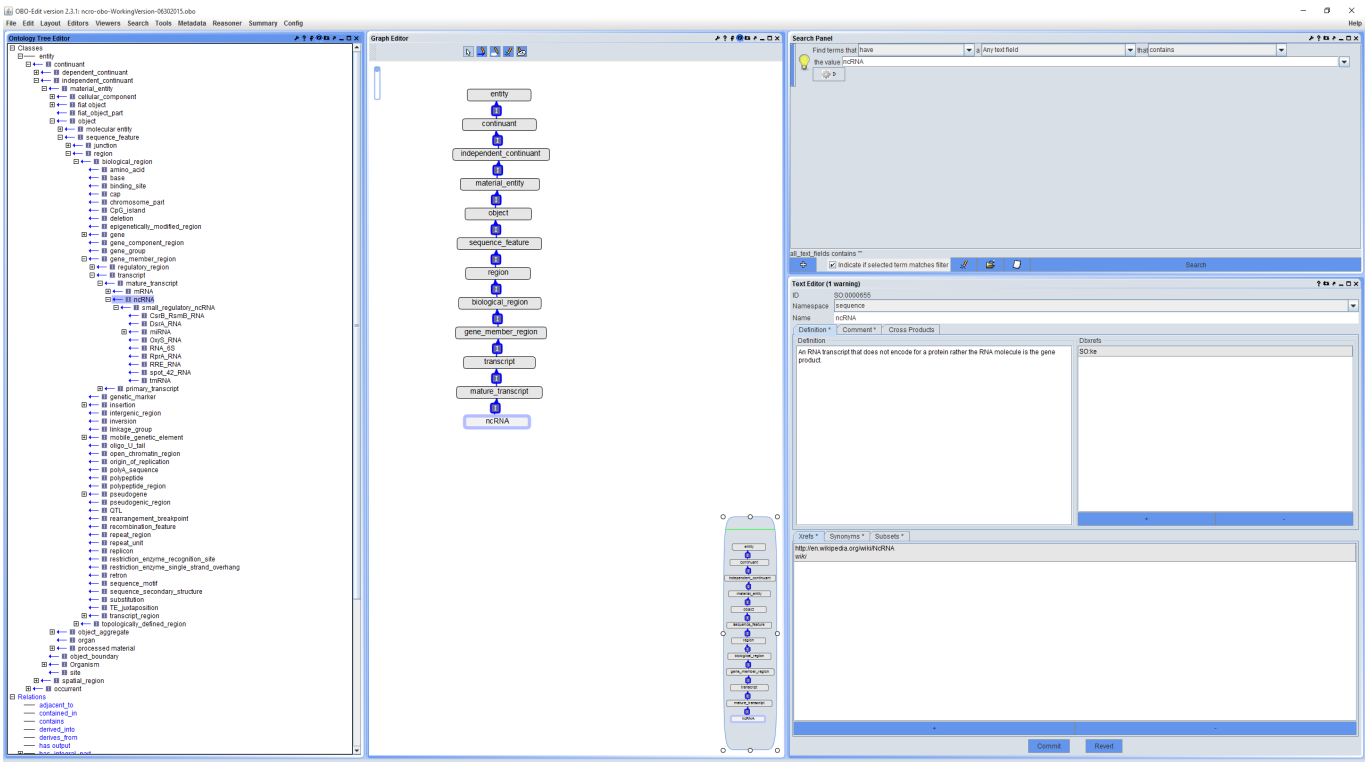

Fig. 2. The NCRO ontology in the OBO-Edit GUI.

- Fig. 1 demonstrates a collection of core terms and relations in the NCRO.
- Fig. 2 is a screenshot of the NCRO ontology in the OBO-Edit graphic user interface (GUI).

V. Conclusions

Research interest in ncRNA biology has significantly grown in recent years, resulting in a great need to establish common data elements and data exchange standards for relevant communities to share their discovery. Neither the OBO Library nor the NCBO BioPort contains ontologies that are specifically designed for the ncRNA domain. It thus motivated us to develop the NCRO ontology to fill such an important knowledge gap. The NCRO aims to provide a set of standardized terms and relations to facilitate semantics-oriented knowledge capture out of large amounts of ncRNA data that are continuously generated from ncRNA and related communities.

The NCRO is an on-going research effort, and ontology files and design documentations are publicly available on a designated project website housed in the University of South Alabama domain [28], as well as on the GitHub project site [29]. In addition, the NCRO ontology is also included in both the OBO Library [30] and the NCBO BioPortal [31].

Acknowledgment

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

References

[1] J. Bard, "Ontologies: Formalising biological knowledge for bioinformatics," *Bioessays*, vol. 25, no. 5, pp. 501–506, May 2003.

[2] J. Blake, "Bio-ontologies-fast and furious," *Nat Biotechnol*, vol. 22, no. 6, pp. 773–774, June 2004.

[3] J. Blake and C. Bult, "Beyond the data deluge: data integration and bio-ontologies," *J Biomed Inform*, vol. 39, no. 3, pp. 314–320, 2006.

[4] J. Huang, D. Dou, L. He, J. Dang, and P. Hayes, "Ontology-Based Knowledge Discovery and Sharing in Bioinformatics and Medical Informatics: A Brief Survey," in *Proc. 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD-2010*, August 2010.

[5] J. Huang, D. Dou, J. Dang, J. Pardue, X. Qin, J. Huan, W. Gerthoffer, and M. Tan, "Knowledge Acquisition, Semantic Text Mining, and Security Risks in Health and Biomedical Informatics," *World J Biol Chem*, vol. 3, no. 2, pp. 27–33, February 2012.

[6] NeuroCommons Project. [Online]. Available: http://neurocommons.org/

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

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

[9] J. Mattick, "Non-coding RNAs: the architects of eukaryotic complexity," *EMBO Rep*, vol. 2, no. 11, pp. 986–991, November 2001.

[10] J. Mattick, "Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms," *Bioessays*, vol. 25, no. 10, pp. 930–939, October 2003.

[11] R. Fatima, V. Akhade, D. Pal, and S. Rao, "Long noncoding RNAs in development and cancer: potential biomarkers and therapeutic targets," *Mol Cell Ther*, vol. 3, June 2015.

[12] I. Babar, C. Cheng, C. Booth, X. Liang, J. Weidhaas, W. Saltzman, and F. Slack, "Nanoparticle-based therapy in an in vivo microrna-155 (mir-155)-dependent mouse model of lymphoma," *Proc Natl Acad Sci USA*, vol. 109, no. 26, pp. E1695–E1704, June 2012.

[13] C. Daige, J. Wiggins, L. Priddy, T. Nelligan-Davis, J. Zhao, and D. Brown, "Systemic delivery of a mir34a mimic as a potential therapeutic for liver cancer," *Mol Cancer Ther*, vol. 13, no. 10, pp. 2352–2360, October 2014.

[14] C. Cheng, R. Bahal, I. Babar, Z. Pincus, F. Barrera, C. Liu, A. Svoronos, D. Braddock, P. Glazer, D. Engelman, W. Saltzman, and F. Slack, "MicroRNA silencing for cancer therapy targeted to the tumour microenvironment," *Nature*, vol. 518, no. 7537, pp. 107–110, Feb. 2015.

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

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

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

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

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

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

[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] The OBO Foundry Principles. [Online]. Available: http://www.obofoundry.org/crit.shtml

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

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

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

[26] OORTl. [Online]. Available: https://code.google.com/p/owltools/

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

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

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

[30] NCRO ontology in the OBO Library. [Online]. Available: http://www.obofoundry.org/cgi-bin/detail.cgi?id=ncro

[31] NCRO ontology in the NCBO BioPortal. [Online]. Available: http://bioportal.bioontology.org/ontologies/NCRO