OBO EDIT: A TOOL FOR CLASSIFING THE BIOLOGICAL DATA

SHEIKH KASHIF RAFFAT[1], MUHAMMAD SARIM[1], SALWA IQBAL[2], MUHAMMAD SHAHAB SIDDIQUI[3] AND MUHAMMAD ZAHID[4]

[1]Department of Computer Science, Federal Urdu University of Arts, Sciences and Technology, Karachi, Pakistan.
[2]Department of Computer Science, Dadabhoy Institute of Higher Education, Karachi, Pakistan.
[3]Computer Science and Information Technology Department, Jinnah University for Women, Karachi, Pakistan.
[4]Department of Zoology, Federal Urdu University of Arts, Sciences and Technology, Karachi, Pakistan.
Corresponding author e-mail: kashifraffat@fuuast.edu.pk

Abstract

Classification of the huge amount of information got more attention since the concept semantic web developed. The sharing of information between people and application efficiently is an important concern of last few decade. Many classification techniques like taxonomy, keyword, thesaurus, data model and ontology exist to classify and defining different biological concepts. To avoid replication and provide conceptual, syntactic and semantic view of data computer scientists adopt 'ontology' for classification. Open Biomedical Ontology Editor (OBO-Editor) is a tool that provides a common platform to share different biomedical and biological information under an umbrella.

Introduction

Every type of information needs to be classified, the classification based on the characteristics of data which is the core principle of any classification technique. All the existing classification techniques follow the core principle but the prime objective is to resolve duplication and give semantic to information.

Over the last few decades, ontology became most popular classification technique due to its object oriented approach and semantic reasoning (Raffat *et al.*, 2012a). The semantic nature of future web will also require all data in ontological content. Ontology based on the philosophy of "existence of object" and defines the "common features of an object". According to Gruber (2008, 1995), "ontology provides a controlled vocabulary of a domain, specify the relationship and attributes to describe the concepts of that domain". Ontology follows the taxonomy, hierarchical list of concepts from abstract to specific but defines semantic relationship between these hierarchical/tree structure. The main concepts of ontologies are; individuals, concepts, properties, relationships between concepts and individuals and most powerful rules and axioms for semantic querying.

There are many tools available to develop the ontologies like OBO-Edit (Wächter and Schroeder, 2010; Day-Richter *et al.*, 2007), Protégé (developed by Stanford University) (Gennari *et al*., 2003), web-based ontology development and editing (WODE) tool (Raffat *et al*., 2012b), first ever web based tool build in LAMP developed by Federal Urdu University, and TODE (Islam *et al*., 2010) developed by National University of Computer and Emerging Science.

OBO-Edit is one of the most popular tool to classify the biological and biomedical concepts into ontological content. OBO-Edit is free desktop based application that aims to unite all biological and biomedical ontologies under an umbrella. The major achievement of this group is Gene Ontology (GO), which is the most mature ontology of biological domain and provides the guideline for developing new ontologies.

In this paper, we will discuss the ontology development tool OBO-Edit and the existing biological ontologies those are developed in OBO-Edit as it now became the unclaimed standard tool for development of biological ontologies.

OBO-Edit and Existing Ontologies: OBO-Edit (Wächter and Schroeder, 2010; Day-Richter *et al*., 2007) is an open source platform independent ontology editor can be download from (http://oboedit.org/), developed in Java and provides a common format to communicate and link the existing ontologies. It also supports OWL format to share the concepts of other ontologies that were not developed in OBO-Edit. It provides different predefined relationships to relate concepts and individuals, and also welcome new relationships based on the requirement of the concepts few used relationships in different ontologies can be seen in table 1, the study of this table will also help to the researchers in the biological domain for using these and defining the new relationships for any new ontology. The sequences ontology defined and uses the more relationships than the any other biological domain ontology. OBO-Edit aims to unite biological ontologies on a single platform, more than 100 different ontologies have been developed in OBO-Edit (http://www.obofoundry.org). Gene Ontology, Protein Ontology, Chemical Interest Biological Entity, Sequence Ontology, Ribonucleic Acid Ontology, Plant Ontology, Disease Ontology are few popular biological ontologies developed in OBO-Edit tool.

Table 1. Relationships in OBO-Edit

Relationship	Example	Definition
is_a	A is_a B	A is child (sub class) of B
part_of	A part_of B	All A are part of B, A is sub region of B
has_part	A has_part B	All A have B, has_part is inverse of part_of
disjoint_from	A disjoint_from B	No sub classes of A and B are same
integral_part_of	A integral_part_of B	if and only if: A part_of B and B has_part A
has_integral part	A has_integral_part B	if and only if: A has_part B and B part_of A
transcribed_from	A transcribed_from B	if A is synthesized from template B
processed_into	A processed_into B	if a region A is modified to create B
processed_from	A is processed_from B	Inverse of processed_into
contained_by	A contained_by B	iff A starts after start of B and B ends before end of A
Contains	A contains B	Inverse of contained_by
overlaps	A overlaps B	iff there exists some X such that X contained_by A and X contained_by B
disconnected_from	A is disconnected_from B	iff it is not the case that A overlaps B
adjacent_to	A adjacent_to B	iff A and B share a boundary but do not overlap

Gene Ontology (GO): One of the most mature ontology of biological domain among the all existing ontologies is GO (Ashburner *et al*., 2000; Gene Ontology Consortium, 2004). It played the major role to provide the guidelines for classifying the other biological domains. Go contains three ontologies, Cellular Component, Molecular Function and Biological Process with relationship of "disjoint_from" that have more than 170,690 terms those provide the information of gene and gene product attributes across all species. Fig. 1(a) and 1(b) shows the hierarchal/tree structure and graphical view of GO in OBO-Edit.

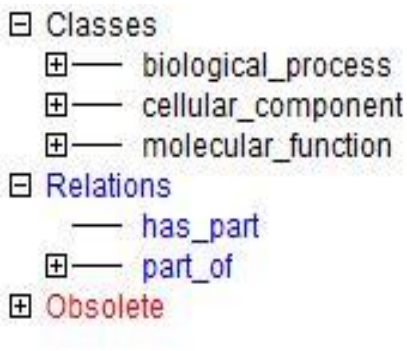

Fig. 1(a). Gene Ontology Tree View

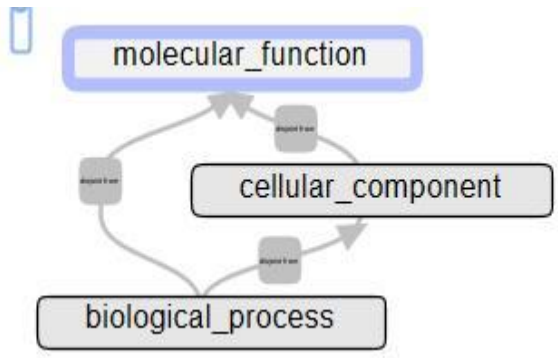

Fig. 1(b). Gene Ontology Graphical View

Protein Ontology (PRO): The PRO is a controlled vocabulary (ontology) that describes relationships of proteins and protein evolutionary classes (Natale *et al*., 2011, 2007). PRO has 83656 classes including 268 of Cell Ontology, 153 of GO and 9 of ChEBI etc. Fig. 2 shows the graphical view of PRO in OBO-Edit.

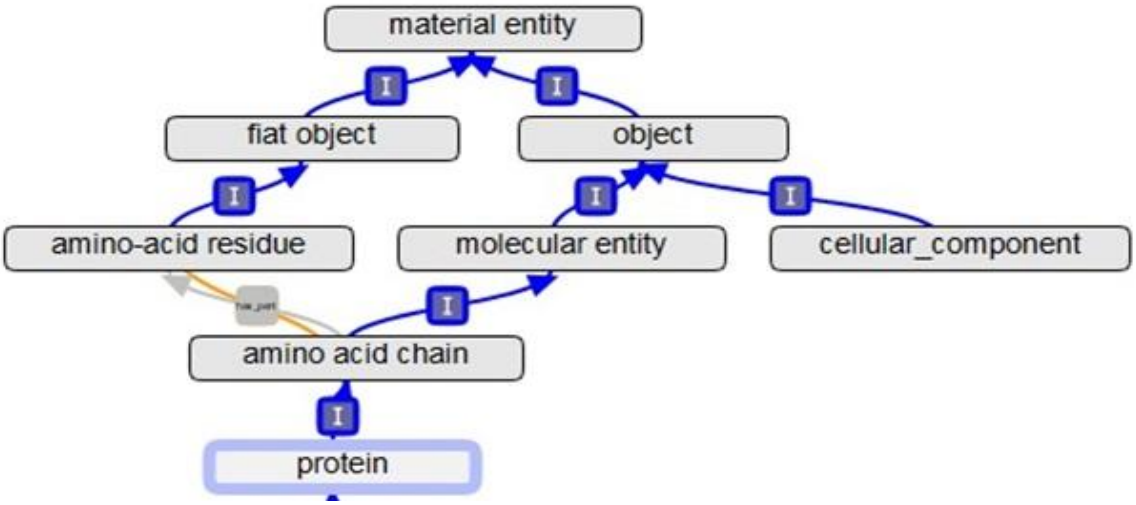

Fig. 2. Protein Ontology Graphical View

Sequence Ontology (SO): The SO is responsible to describe key features for genomic and other structured sequence (Eilbeck *et al*., 2005; Mungall *et al*., 2011). SO defined and uses more relationships than the any other biological domain ontology. The relationships used in SO can be seen in Fig. 3.

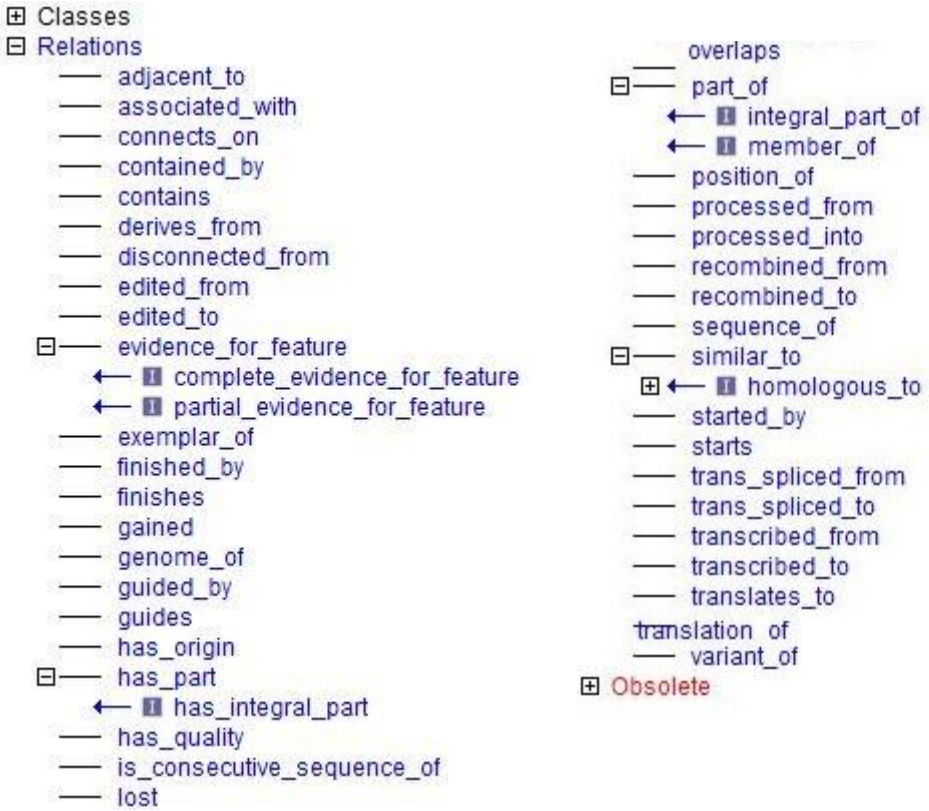

Fig. 3. OBO-Edit View of SO Relationships

Ribonucleic Acid Ontology (RNAO): RNAO is a structured vocabulary of RNA sequences, secondary, three-dimensional structures and dynamics pertaining to RNA function can be seen in Fig. 4. RNAO used has functional parent relation to link it with Chemical Entities of Biological Interest (ChEBI) ontology. RNAO also linked with Gene, Sequence and Multiple Alignment ontologies. RNAO developed in Protégé by using the guidelines of Ontology Development 101 and also available in OBO format (Hoehndorf *et al*., 2011; Leontis *et al*., 2006).

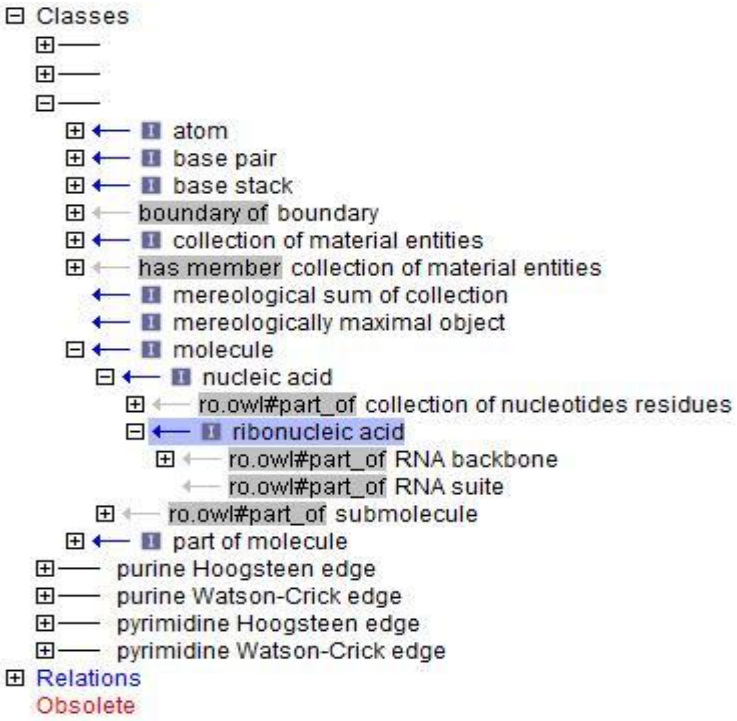

Fig. 4. Tree structure of RNAO

Chemical Interest Biological Entity (ChEBI): It is the ontology of chemical entities focus on molecule, atom and ion (Degtyarenoko *et al*., 2008; Hastings *et al*., 2013). The ChEBI consist of three sub ontologies; Chemical Entity, Role (application, biological and chemical) and Subatomic Particle can be seen in Fig. 5(a). ChEBI developed in Protégé and then converted in OBO format to link with other biological ontologies. Protein ontology, Sequence ontology, Ribonucleic acid ontology, and Biological viruses community ontology (Raffat *et al*., 2011) linked ChEBI with their terms.

Disease Ontology (DO): DO is a controlled vocabulary of human diseases, map the medical code of ICD 10 and SNOMED CT (Schriml *et al*., 2012). It categorized the diseases by environmental origin, infectious agent, anatomical entity, biological process, mental health, disorder, hereditary disease and syndrome as shown in Fig. 5(b). DO developed in OBO-Edit by using the using the principles of OBO.

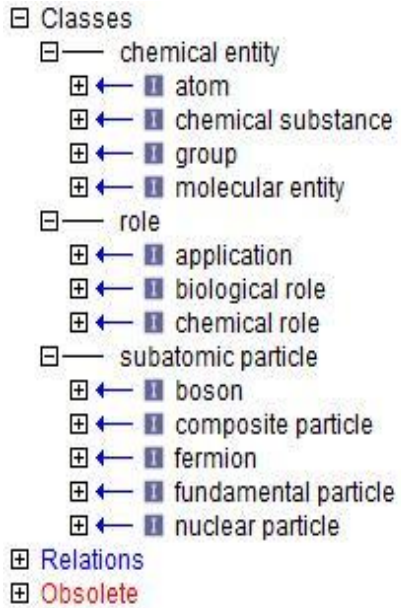

Fig. 5(a). Tree structure of ChEBI

Fig. 5(b). Tree structure of DO

Plant Ontology (PO): The PO is a controlled vocabulary that describes the anatomy of plant, morphology and stages of development for all plants (Avraham *et al*., 2008; Cooper *et al*., 2013; Jaiswal *et al*., 2005). Plant ontology contains 1592 terms including 1296 terms of plant anatomical entity and 296 terms of plant structure development stage. The hierarchical view of plant ontology can be seen in Fig. 6.

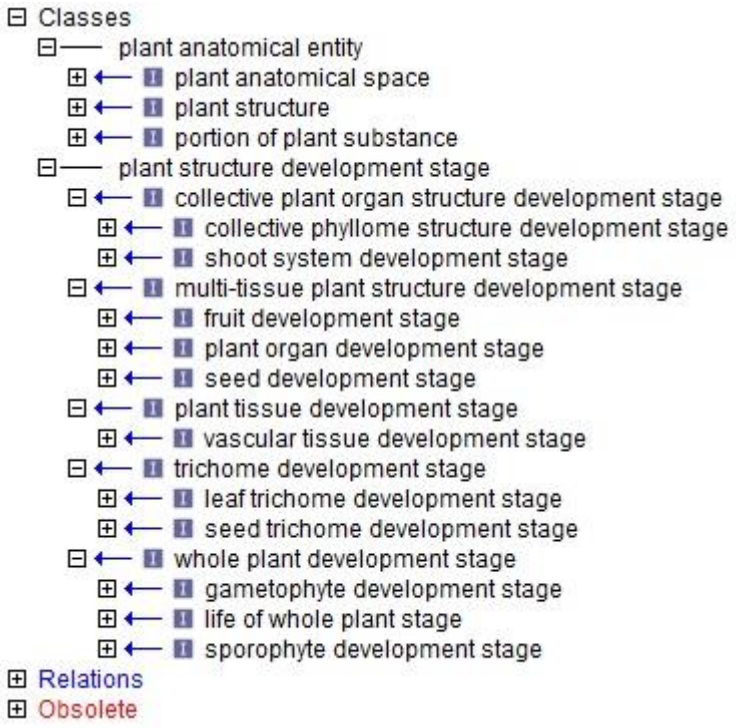

Fig. 6. Tree structure of Plant Ontology

Plant Trait Ontology (PTO): It is a complete and controlled vocabulary and defines each trait as a distinguishable feature, quality and characteristics of individual plant (Plant Ontology Consortium, 2002). It includes the aspects of growth and development trait, quality trait, sterility trait, morphology trait, yield trait, stress trait, biochemical trait, vigor trait can be seen in Fig. 7. Plant trait ontology contains 3067 terms with maximum 159 number of children of a class and maximum depth of 12. 25 classes have more than 4 children and 259 classes have only one child.

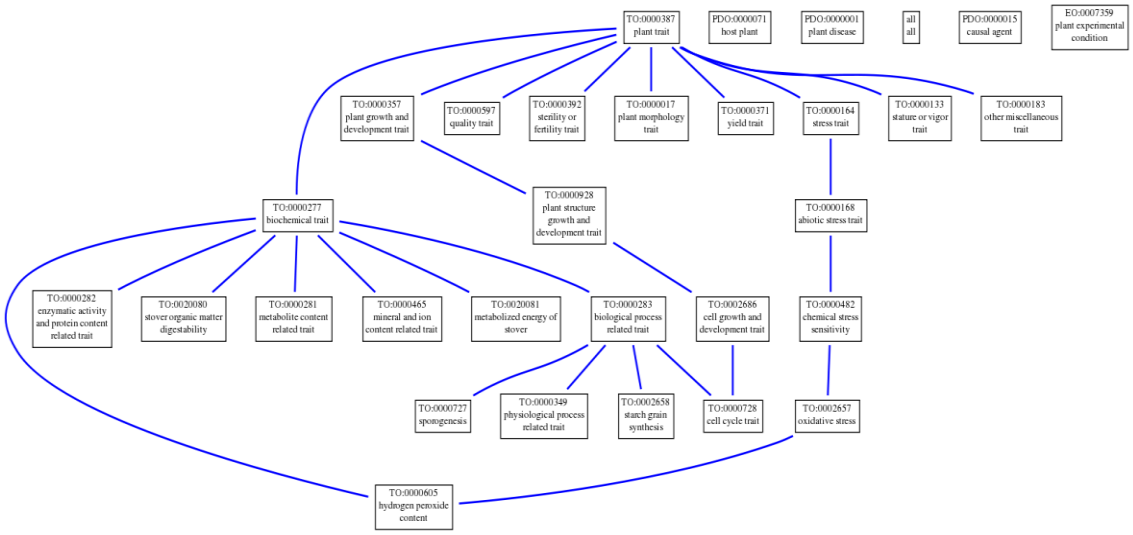

Fig. 7. Graphical View of Plant Trait Ontology

Conclusion

The aim of this study was to highlight the importance of ontology especially in biological domain, give the awareness about the most popular ontology development tool OBO-Edit and few existing biological ontologies. The teachers of our local environment should provide the existing ontologies of their domain to the students for the better understanding the concepts.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J. M., ... and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1): 25-29.

Avraham, S., Tung, C.W., Ilic, K., Jaiswal, P., Kellogg, E.A., McCouch, S., ... and Ware, D. (2008). The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Research*, 36(suppl 1), D449-D454.

Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.A., Stevenson, D.W., Smith, B., ... and Jaiswal, P. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology*, 54(2), e1-e1.

Day-Richter, J., Harris, M. A., Haendel, M. and Lewis, S. (2007). OBO-Edit—an ontology editor for biologists. *Bioinformatics*, 23(16): 2198-2200.

Degtyarenoko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., ... and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl 1), D344-D350.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5), R44.

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(suppl 1), D258-D261.

Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubézy, M., Eriksson, H., ... and Tu, S.W. (2003). The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-computer Studies*, 58(1): 89-123.

Gruber, T. (2008). Ontology. *Entry in the Encyclopedia of Database Systems*, Ling Liu and M. Tomer Ozsu (Eds.), *Springer-Verlag*.

Gruber, T.R. (1995). Toward principles for the design of ontologies used for knowledge sharing?. *Int. J of Human-computer Studies*, 43(5): 907-928.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., ... and Steinbeck, C. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, 41(D1): D456-D463.

Hoehndorf, R., Batchelor, C., Bittner, T., Dumontier, M., Eilbeck, K., Knight, R., ... and Leontis, N.B. (2011). The RNA Ontology (RNAO): an ontology for integrating RNA sequence and structure data. *Applied Ontology*, 6(1), 53-89.

Islam, N., Siddiqui, M. S., and Shaikh, Z. (2010). TODE: A dot net based tool for ontology development and editing. *2nd International Conference on in Computer Engineering and Technology (ICCET), IEEE*.

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., ... and Zapata, F. (2005). Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comparative and Functional Genomics*, 6(7-8), 388-397.

Leontis, N.B., Altman, R.B., Berman, H.M., Brenner, S.E., Brown, J.W., Engelke, D.R., ... and Westhof, E. (2006). The RNA Ontology Consortium: an open invitation to the RNA community. *RNA*, 12(4), 533-541.

Mungall, C.J., Batchelor, C. and Eilbeck, K. (2011). Evolution of the Sequence Ontology terms and relationships. *Journal of Biomedical Informatics*, 44(1): 87-93.

Natale, D.A., Arighi, C.N., Barker, W.C., Blake, J., Chang, T.C., Hu, Z., ... and Wu, C.H. (2007). Framework for a protein ontology. *BMC Bioinformatics*, 8(Suppl 9), S1.

Natale, D.A., Arighi, C.N., Barker, W.C., Blake, J.A., Bult, C.J., Caudy, M., ... and Wu, C.H. (2011). The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Research*, 39(suppl 1): D539-D545.

Plant Ontology Consortium. (2002). The Plant Ontology™ consortium and plant ontologies. *International Journal of Genomics*, 3(2): 137-142.

Raffat, S.K., Siddiqui, M.S., Shaikh, Z.A. and Memon, A.R. (2012a). Ontology: A Scientific Classification Technique. *Sindh University Research Journal*, 44(2AB): 63-68.

Raffat, S.K., Siddiqui, M.S., Siddiq, M. and Shafiq, F. (2012b). Towards the Development of Web-based Ontology Development and Editing (WODE) Tool. *Research Journal of Recent Sciences*, 1 (12): 67-69.

Raffat, S.K., Siddiqui, M.S., Shaikh, Z.A. and Memon, A.R. (2011). Towards the development of Biological Viruses Community Ontology (BVCO). *J. of Computing*, 3(4): 125-129.

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., ... and Kibbe, W.A. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1): D940-D946.

Wächter, T. and Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit. *Bioinformatics*, 26(12): i88-i96.