Semantic Integration of Intrinsically Disordered Proteins and Existing DBs

Atsuko Yamaguchi[1], Hideki Hatanaka[1], Satoshi Fukuchi[2], and Motonori Ota[3]

[1] Database Center for Life Science (DBCLS),
Research Organization of Information and Systems (ROIS).
178-4-4 Wakashiba, Kashiwa, Chiba, 277-0871 Japan
{atsuko,hideki}@dbcls.rois.ac.jp
[2] Faculty of Engineering, Maebashi Institute of Technology.
Maebashi 371-0816, Japan
sfukuchi@maebashi-it.ac.jp
[3] Graduate School of Informatics, Nagoya University.
Nagoya 464-8601, Japan
mota@i.nagoya-u.ac.jp

Abstract. IDEAL (http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/) is one of the largest collections of experimentally verified intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs). It is known that IDPs and IDRs play important roles in crucial biological processes such as signal transaction and transcription control. Essentially, knowledge on IDPs and IDRs is complementary information of 3D structure of proteins stored in the wwPDB. In addition, comparison of regions of intrinsically disordered proteins with annotated regions stored in other biological databases such as UniProt may help a user to develop new biological knowledge. Therefore, we constructed the RDF version of IDEAL to facilitate semantic comparison with the existing DBs such as the wwPDB and UniProt.

Keywords: Linked Open Data, database integration, intrinsically disordered proteins

1 Background

Intrinsically disordered proteins (IDPs) are proteins that do not adopt unique 3D structures under physiological conditions. IDEAL (Intrinsically Disordered proteins with Extensive Annotations and Literature) is a database that provides a collection of knowledge on experimentally verified intrinsically disordered proteins or intrinsically disordered regions (IDRs)[1, 2]. IDEAL contains manually curated annotations on IDPs in locations, structures, and functional sites such as protein binding regions and post-translational modification sites together with references and structural domain assignments. As of October 2019, IDEAL contains 11643 non-redundant IDRs and is the largest database of IDRs. All data of IDEAL is available under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

CEUR-WS.org/Vol-2849/paper-24.pdf

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

For each protein in IDEAL, order and disorder regions are annotated. Disorder regions manually derived from literature are complementary information of the wwPDB because the wwPDB provides 3D structures of proteins. On the other hand, the reference of start and end positions of an IDR is an amino acid sequence in UniProt. Therefore, it should be useful to integrate IDEAL with the existing DBs such as the wwPDB and UniProt. Fortunately, the wwPDB and UniProt have already been published as Linked Open Data (LOD). Therefore, we constructed an RDF model of IDEAL to publish it as LOD.

2 Result

To facilitate comparison with the existing databases, we designed an RDF model to use ontologies and vocabularies used in the wwPDB and UniProt if possible. For example, to describe a start and end position of a region in IDEAL, we employed FALDO[3] that is used by UniProt. By using FALDO for regions in IDEAL, knowledge in IDEAL can be seamlessly merged to annotations from the existing databases including UniProt. For an identifier for IDEAL, we used URLs provided by identifiers.org[4]. Registered information for IDEAL is available at https://registry.identifiers.org/registry/ideal. Additionally, we defined 21 classes and 65 properties that are necessary to describe annotations for IDRs.

A file of IDEAL in RDF is downloadable in Turtle format from the IDEAL site. A SPARQL endpoint for IDEAL, which uses Virtuoso 7, is accessible at https://ideal-rdf.dbcls.jp/sparql. The numbers of triples is 2.7M. By providing IDEAL in the RDF model, knowledge in IDEAL is integrated to the existing databases in LOD through the wwPDB and UniProt. We believe that it helps a user to find new biological knowledge by combining annotations in IDEAL and information from other biological databases in LOD.

Acknowledgments This work was supported by JSPS KAKENHI grant numbers 17K00434 and by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

References

1. Fukuchi, S., Amemiya, T., Sakamoto, S. , Nobe, Y., Hosoda, K., Kado, Y., Murakami, S. D., Koike, R., Hiroaki H., Ota, M.: IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. Nucleic Acids Res., 42(Database issue), D320-D325 (2014).
2. IDEAL: Disordered Proteins and Annotations. http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/
3. Bolleman, J. T., Mungall, C. J., Strozzi, F., Baran, J., Dumontier, M., Bonnal, R. J. P., Buels, R., Hoehndorf, R., Fujisawa, T., Katayama, T., Cock, P. J. A.: FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. Journal of Biomedical Semantics, 7, 39 (2016).
4. Juty, N., Novère, N. L., Laibe, C: Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res., 40(Database issue), D580–D586 (2012).