ICEO: a biological ontology for representing and analyzing the bacterial integrative and conjugative element

Meng Liu[a,b], Hong-Yu Ou[a], Yongqun He[b]

[a] *State Key Laboratory of Microbial Metabolism, School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai, China*

[b] *University of Michigan Medical School, Ann Arbor, MI 48109, USA*

Abstract

Bacterial integrative and conjugative elements (ICEs) are mobile genetic elements critical to horizontal gene transfer and organism evolution. To better understand and analyze ICEs, it is critical to systematically represent, integrate and classify gene components, functional modules and related information of available ICEs archived in the ICEberg database. Toward this goal, we developed a community-driven ICE ontology (ICEO). ICEO is aligned with the Basic Formal Ontology (BFO) to allow the integration with other ontologies. ICEO reused the existing reliable ontologies, such as Ontology of Gene and Genome, Protein Ontology and NCBITaxon. ICEO now represents the information about over 270 experimentally verified ICEs from 235 bacterial strains. Two query use cases were provided, including a DL query of ICE-contained genes that are also virulence factors and a SPARQL query of ICEs under an upper level taxonomy type of Gammaproteobacteria. Our study demonstrated that ICEO supports computer-assisted reasoning and efficient SPARQL query.

Keywords:

ICE; ontology; mobile genetic elements

Introduction

Integrative and conjugative elements (ICEs), also called conjugative transposons before, are a large family of the bacterial mobile genetic elements (MGEs) (1). ICEs are integrative to the bacterial chromosome and encode functional conjugation machinery for the self-transmission between bacterial cells. Similar to all other bacterial MGEs, typically, ICEs have a highly modular structure with three core genetic modules: (i) recombination (integration and excision) module; (ii) conjugation module; and (iii) regulation module. The recombination module refers to those genes and sequence within the ICE responsible for the site-specific integration and excision of the element from the host chromosome, including genes encoding the integrase and or recombination directionality factor (also known as excisionase, which influences the direction of recombination mediated by the integrase to favor excision). The conjugation module denotes those gene and sequence involved in the conjugal process, such as genes encoding relaxase and the type IV secretion system (T4SS). The regulation module refers to those genes and sequence contributing to stabilization and maintenance of ICEs. Besides, virulence factors (VFs) and acquired antibiotic resistance genes (ARGs) often exist inside ICEs as the cargo genes (also called the accessory module of ICEs) and can confer the hosts with selective advantages, which make ICEs a vital role in the process of bacterial adaptation and evolution (2). Based on two different conjugal manners, ICEs can be categorized as T4SS-type ICEs and actinomycete ICEs (AICEs). T4SS-type ICEs are widely distributed both in Gram-negative and Gram-positive bacteria, while actinomycete ICEs (AICEs) only have been found in *Actinobacteria*, mainly in *Streptomyces*. And T4SS-type ICEs are transferred as linear single-stranded DNA (ssDNA) typically depended on a relaxase and a conjugative type IV secretion system (T4SS). AICEs are delivered as double-stranded DNA (dsDNA) relied on proteins for replications and translocation.

The information about thousands of experimentally validated or computationally predicted bacterial ICEs is freely accessible in ICEberg (http://db-mml.sjtu.edu.cn/ICEberg/), a comprehensive web-based ICE database that is developed by our group at the Shanghai Jiao Tong University (3). However, to make the best use of these available data and the ongoing increase of information, and to facilitate more effective and accurate identification and annotation of ICEs from single bacterial genomes or even metagenomes, a knowledge base about available bacterial ICEs in a format compliant for computer analysis is desired (4, 5). Ontology, a hierarchical and interconnected controlled vocabulary that emphasizes the logical organization and representation of complex data and knowledge, provides such a platform to achieve this goal. In this big data and IT era, structured ontology has been widely used in biological data and metadata standardization, integration, sharing, and analysis (6). For example, one of the most successful and widely-used ontology, Gene Ontology (GO; http://www.geneontology.org/) (7), which represents the information of cellular components, biological processes, and molecular functions, is often used as the standard to describe the function of gene and gene products across different databases and to conduct various gene expression analyses. The usage of ontology supports better representation, integration, and analysis of big data.

In this study, we report the development strategy of a community-driven Ontology of the Integrative and Conjugative Element

(ICEO), which is aimed to ontologically represent and integrate the ICE gene information and functional modules to support computer-assisted reasoning. There are 260 experimentally verified T4SS-type ICEs (113 with the entire nucleotide sequences) and 11 experimentally validated AICEs (7 with the entire nucleotide sequences) in ICEberg. The current ICEO is focused on ontological organization of the information about these 271 experimentally validated ICEs for now here. ICEO is developed by reusing many terms from existing ontologies and using the state-of-the-art ontology engineering technologies (8, 9). A systematic analysis of the ICEO-represented knowledge base allows us to generate new insights about these widely distributed integrative genetic elements.

Methods

1. ICEO ontology development strategy

The development of ICEO follows the Open Biological and Biomedical Ontologies (OBO) Foundry principles (10), such as openness, collaboration, use of a common shared syntax and so on. Therefore, the ICEO information can be easily integrated and processed with other ontologies in the OBO library. To support the data FAIRness (Findable, Accessible, Interoperable and Reusable) (11), the eXtensible Ontology Development (XOD) strategy (9) was also applied for the ontology development of ICEO. Basically, the XOD strategy recommends the reuse of existing terms and semantic relations from reliable ontologies, development and application of well-established ontology design patterns (ODPs), and involvement of community efforts for new ontology development (9).

2. ICE-related ontology term reuse

To support ontology interoperability and avoid reinventing the wheel, related existing terms from reliable ontologies were imported into ICEO via an Ontofox (http://ontofox.hegroup.org) import strategy (12). The external ontologies used here include Ontology of Genes and Genomes (OGG) (13), PRotein Ontology (PR) (14), Gene Ontology (GO) (7) and a taxonomy ontology of NCBI organismal classification (NCBITaxon) (15).

3. New ICEO term generation

Based on the available information, an ontology design pattern (ODP) was developed. Many new annotations and relations between different entities were added by utilizing the Ontorat (http://ontorat.hegroup.org/) (16), an online program designed to support ODP-based creation of new ontology terms, hierarchies, annotations, and logical axioms.

The Protégé-OWL editor (version 5.2) (http://protege.stanford.edu/) was used for the ICEO manual processing and editing, ontology term merging and visualization. ICEO-specific terms were generated using new ICEO identifiers with the prefix "ICEO_" followed by auto-generated 7 digits. The Hermit reasoner (http://hermit-reasoner.com/) was applied for semantic consistency checking and inferencing.

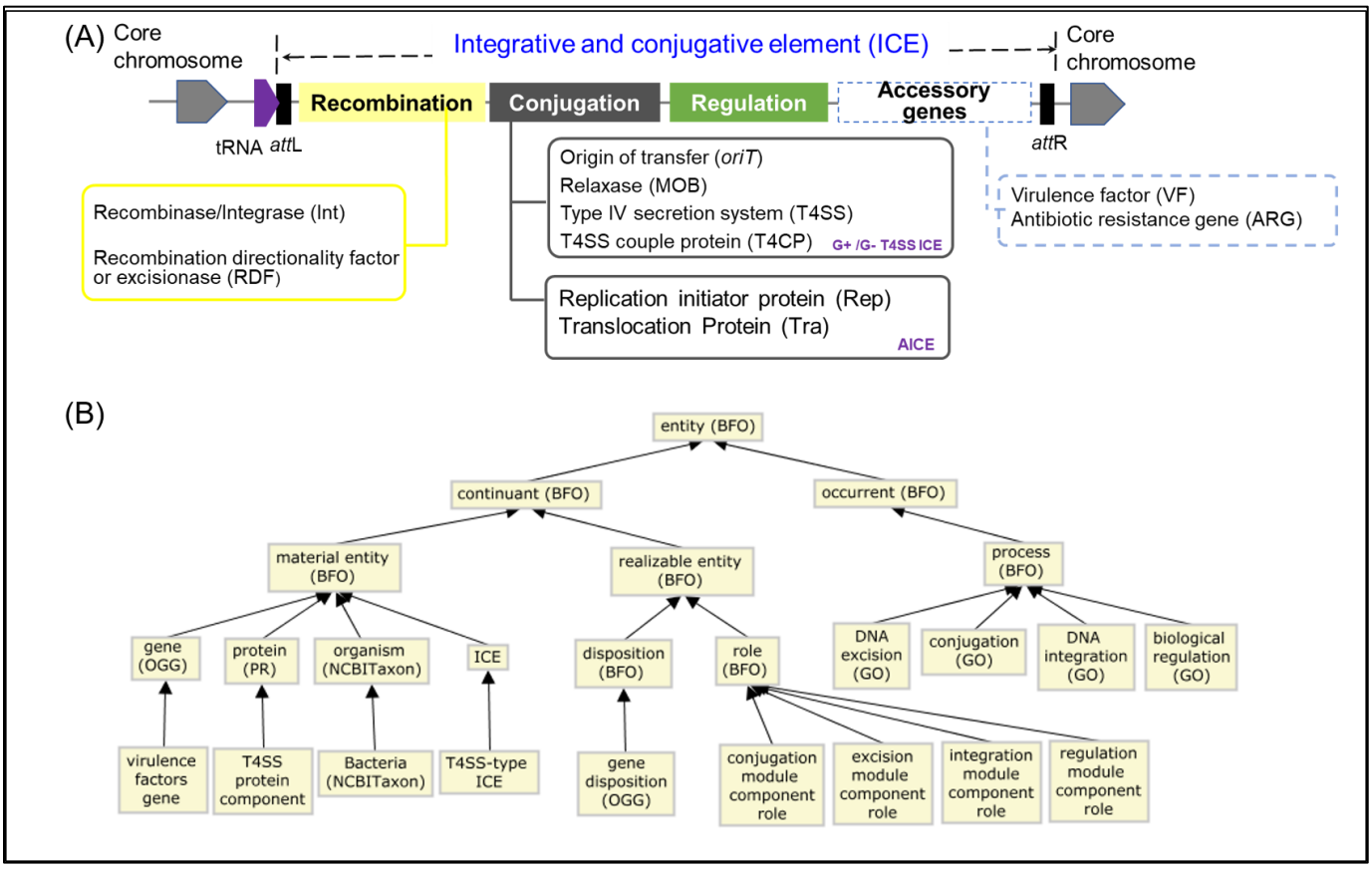

Fig. 1. (A) Classical ICE conserved modules. *ICEs typically contain three core modules: (i) a recombination (integration and excision) module, (ii) a conjugation module and (iii) a regulation module. In addition to three core modules, most ICEs possess conserved accessory regions.* ***(B) ICEO top-level hierarchy****. Terms with ontology abbreviations inside parentheses are imported from external ontologies, while terms without an identified source are ICEO terms. Some intermediate terms such as those terms in between different layers are not shown to make the relations simple and clear. All the arrows indicate the 'is a' relation.*

4. ICEO format, source code, and access

The ICEO is developed using the format of W3C standard Web Ontology Language (OWL2) (https://www.w3.org/TR/owl-guide/) (10). The source code of ICEO is open and available for public view and download on the GitHub website: https://github.com/ontoice/ICEO. The ICEO source code is freely available under the Creative Commons 4.0 License (http://creativecommons.org/licenses/by/4.0/), which allows ICEO users to freely distribute and use ICEO. The latest version of ICEO is also accessible for visualization and downloading from Ontobee (17, 18) ontology repository website: http://www.ontobee.org/ontology/ICEO, or NCBO's BioPortal website: https://bioportal.bioontology.org/ontologies/ICEO.

5. ICEO knowledge query and analysis

The knowledge stored in the ICEO ontology can be queried through different approaches. In this study, we used the Description Logic (DL) query and SPARQL (a recursive acronym for SPARQL Protocol and RDF Query Language) query. The DL query was performed using the Protégé OWL editor. For the SPARQL query, ICEO was stored in the Resource Description Framework (RDF; https://www.w3.org/RDF/) triples in the Ontobee RDF triple store (17, 18). The Ontobee SPARQL query interface (http://www.ontobee.org/sparql) was then used for ICEO specific SPARQL query.

Results

1. ICEO top-level design and development

a. ICEO is aligned with BFO and OBO foundry ontologies

Fig. 1A illustrates the genetic functional modules of the ICEs abundant in both Gram-positive and Gram-negative bacteria. Basically, a typical ICE includes three core modules: recombination (integration and excision) module, conjugation module, and regulation module, which relate to the ICE life cycle. In addition to these three core modules, most ICEs carry some cargo genes (also called accessory genes) like virulence factors (VFs) genes and acquired antibiotic resistance genes (ARGs) conferring adaptive phenotypes to the hosts of ICEs.

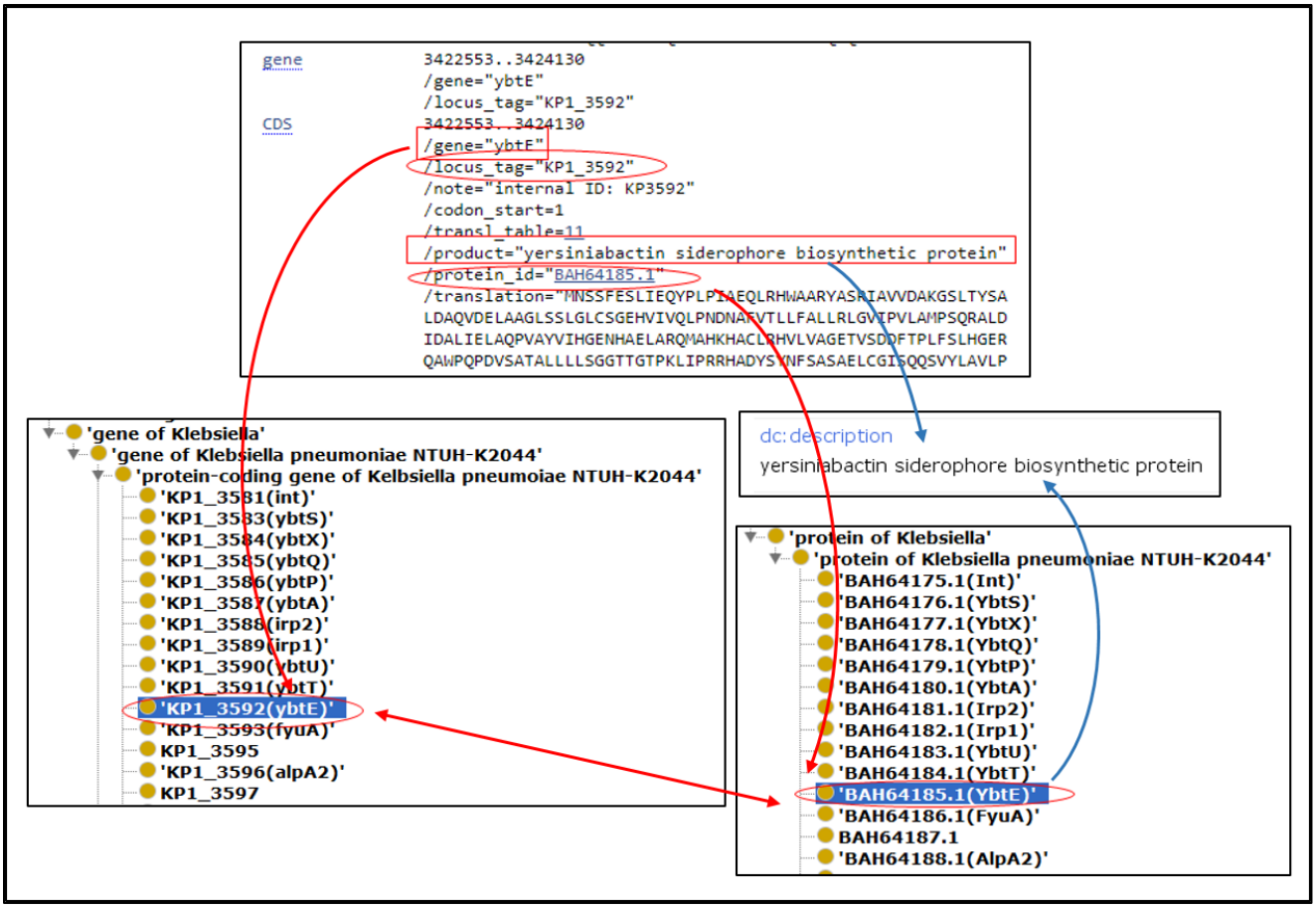

Fig. 2. Gene and protein label naming strategy of ICEO. *The 'ybtE' gene archived in the NCBI GenBank database is assigned with gene label 'KP1_3592(ybtE)' in ICEO, and its corresponding protein label is 'BAH64185.1(YbtE)'.*

Fig. 1B represents the basic top-level ICEO hierarchical structure in accordance with the ICE genetic functional modules (Fig. 1A). Specifically, ICEO is aligned to the upper-level Basic Formal Ontology (BFO) 2.0 version (19, 20). BFO consists of 'continuant' and 'occurrent' branches. The 'continuant' branch stands for time-independent entities (e.g., material entity and their quality and roles), while the 'occurrent' branch represents time-related entities (e.g. process and time). Since BFO has been used as the upper-level ontology by over 100 ontologies, the alignment

of ICEO with BFO facilitates the effective integration of ICEO with many other ontologies.

To enable the reusability of existing ontologies, ICEO imports many related terms and relations from OBO library ontologies. As shown in Fig. 1B, ICEO imports OGG and PR to represent the genes and proteins of ICEs. NCBITaxon terms are imported to represent various ICE-containing organisms in the taxonomic organism hierarchy. GO terms are imported to represent the processes in the whole life cycle of ICEs.

b. Modified and extensive gene and protein ID assignments and label naming strategy

Given the ever-growing number of genes sequenced and annotated, the phenomena of having genes or proteins in different organisms but with the same names archived in the NCBI GenBank database is inevitable. To avoid name conflicts, the original OGG designed a special scheme to automatically assign gene IDs by mapping ontology ID with NCBITaxon IDs and NCBI Gene IDs (13). However, the integer sequence identifiers known as "GIs" and Gene ID are no longer provided and used by NCBI for the sequence records in non-reference strains since September 2016 (21). Furthermore, for many organisms harboring ICEs, for example, *Escherichia coli* strain ECOR31 that carries ICE gene components, there are no available NCBITaxon IDs. In addition, OGG still faces the gene label redundancy since OGG only used the gene name or locus tag as the ontology label of genes.

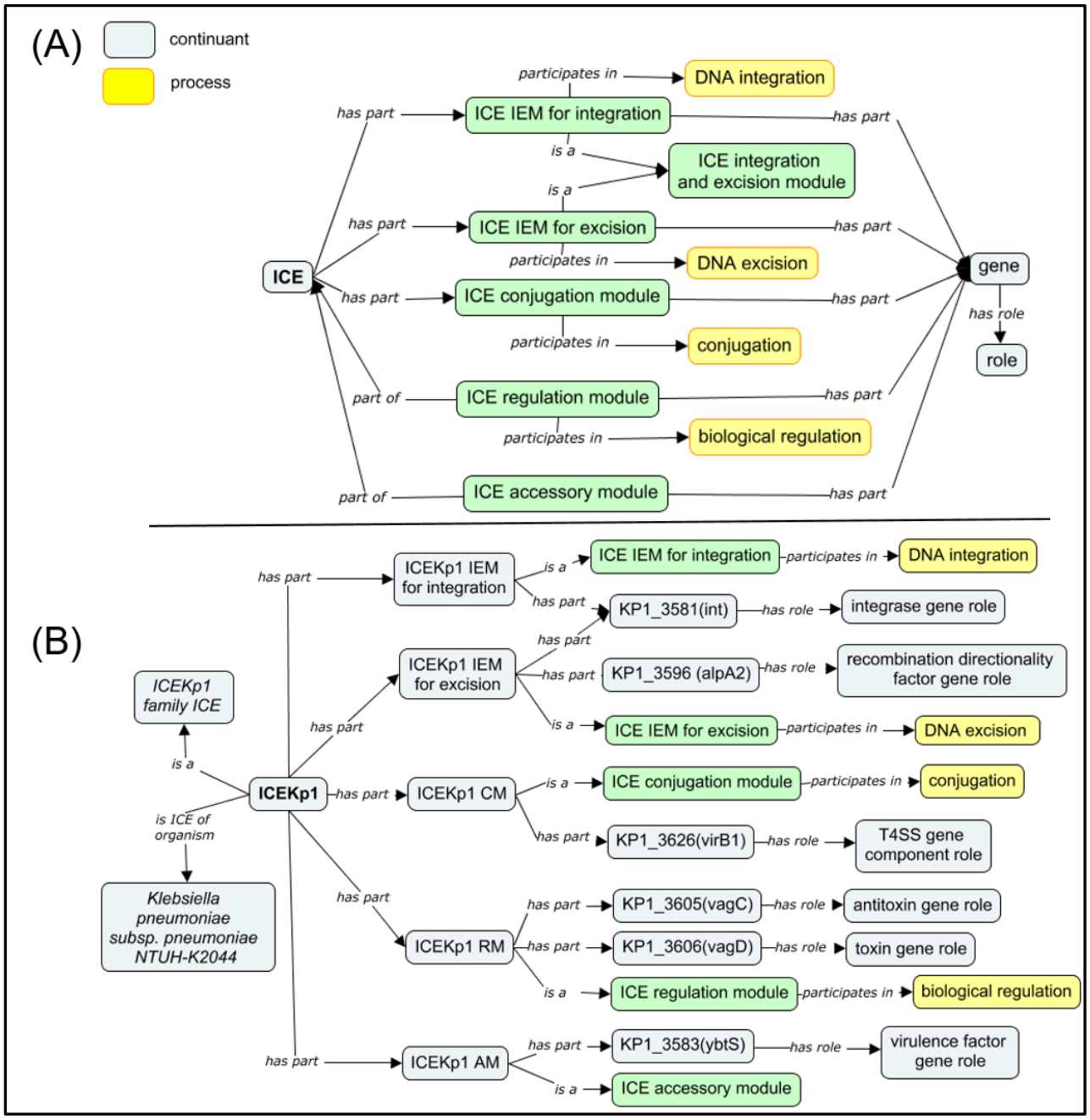

Fig. 3. ICEO design pattern and an example. *(A) Generic ontology design pattern for relations among terms in ICEO. (B) An example of ICEKp1 representation by applying ICEO design pattern with extended information obtained from the ICEberg database.*

Due to these reasons, we have worked with the OGG development team and developed an OGG-based extended strategy of generating new OGG IDs for gene assignments for ICE-related genes. Simply put, this strategy assigns OGG gene IDs using NCBI locus_tag identifiers commonly seen in GenBank gene records. Generally, if gene name is available for a gene, then it's gene label will be assigned as 'locus_tag(gene_name)'; if not, the gene label will be 'locus_tag'. In addition, 'product' information will be added to 'dc:description' property of the corresponding protein. Such a naming strategy allows us to develop and design computer programs to automatically generate readable and nonredundant ICEO gene label. For example, the *yetE* gene in *Klebsiella pneumoniae* strain NTUH-K2044 has a locus tag of KP1_5092. Accordingly, we assign this yetE gene as 'KP1_5092(yetE)' and assign its ID as "OGG_KP1_5092" (Fig. 2).

ICE is essentially a genetic feature so that the gene representation is our priority. Since the Protein Ontology (PR) does not include all the proteins included in ICEO, we applied a similar strategy to represent protein names in ICEO (Fig. 2). We have also contacted the Protein Ontology (PR) team, and will request new classes in the PR to improve ICEO.

2. ICEO ontology design pattern

Fig. 3 illustrates the ICEO ontology design pattern which logically links different types of entities. Compared to the ICEO top-level hierarchy (Fig. 1) which shows the hierarchical relationships among different terms, Fig. 3 shows the logical relations of related terms across different hierarchical structures in ICEO. Together, the combination of Fig. 1 and Fig. 3 presents us a general framework of the ontological design of ICEO.

As shown in Fig. 3A, the basic ICEO design pattern is to represent ICE from the view of typical function modules. ICE '*has part*' integration, conjugation, regulation and accessory module components, and these components and or their encoding proteins '*participants in'* specific ICE life process.

An example of applying this general design pattern to a concrete example is shown in Fig. 3B where the design pattern is used to represent ICE*Kp*1, a virulence-associated ICE found in *Klebsiella pneumoniae subsp. pneumoniae* NTUH-K2044 causing a primary liver abscess (22, 23). Basically, ICE*Kp*1 contains all essential genes necessary to the whole life cycle of ICE*Kp*1 (that is, the process of excision, conjugation, regulation, and integration) (Fig. 3B). Besides, a virulence factor gene cluster within ICE*Kp*1 responsible for the synthesis, regulation, and transport of the siderophore yersiniabactin confers the high virulence to the *K. pneumoniae* NTUH-K2044 (22, 23). Fig. 4 is the more specific demonstration of such an example visualized by Protégé.

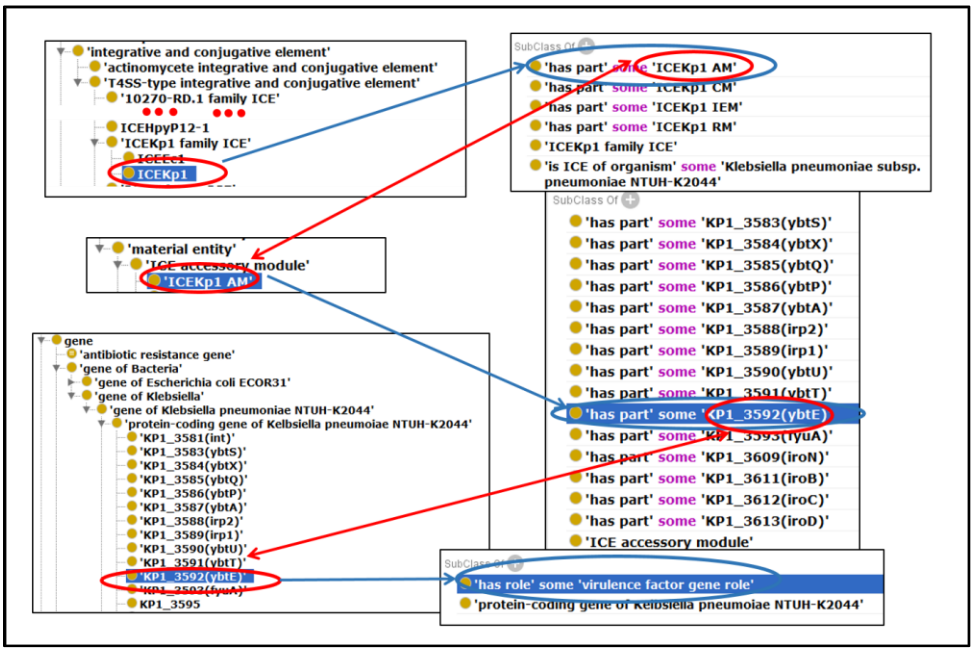

Fig. 4*. Demonstration of ICEO linkage of different entities and representation of the virulence factor gene ybtE of the accessory module in ICEKp1.*

3. ICEO statistics

The latest release of ICEO contains a total of 7738 terms, including 7604 classes, 52 object properties, and 78 annotation properties. Among these terms, 1713 terms have ICEO_ namespace. The full ontology statistics of ICEO is accessible in the Ontobee ICEO statistics page (http://www.ontobee.org/ontostat/ICEO).

4.ICEO applications

ICEO is formatted in the machine-interpretable OWL format, which is easily understood by computer programs and can support various advanced queries and analyses. Therefore, ICEO can be used for various applications, such as DL query and SPARQL query, which is designed for RDF triple and cannot be done directly in ICEberg. Two ICEO use cases are provided in the study as follows:

Use Case 1: Use DL Query to query the specified group of genes of an ICE

ICEO supports OWL-based automated reasoning using reasoning programs within OWL editors. As an example, we designed the following question for query the ontology:

What genes in the ICEKp1 encode virulence factors?

In ICEO, a virulence factor gene is logically defined as "something that *has role* some *virulence factor gene role*". To answer this question, the following query was simply performed using the DL Query function in Protégé OWL editor version 5.2 (Fig. 5):

('has role' some 'virulence factor gene role') and ('part of' some ICEKp1)

As shown in Fig. 5, our query identified 15 genes that are part of the specific ICE*Kp*1 and also encode for virulence factors in the host bacterium of the ICE*Kp*1. The result indicates that ICE*Kp*1 is responsible for transporting this set of virulence factor genes to *Klebsiella pneumoniae* strain NTUH-K2044, the bacterium that hosts the ICE. ICE*Kp*1 is indeed critical to make the bacterium virulent (22, 23).

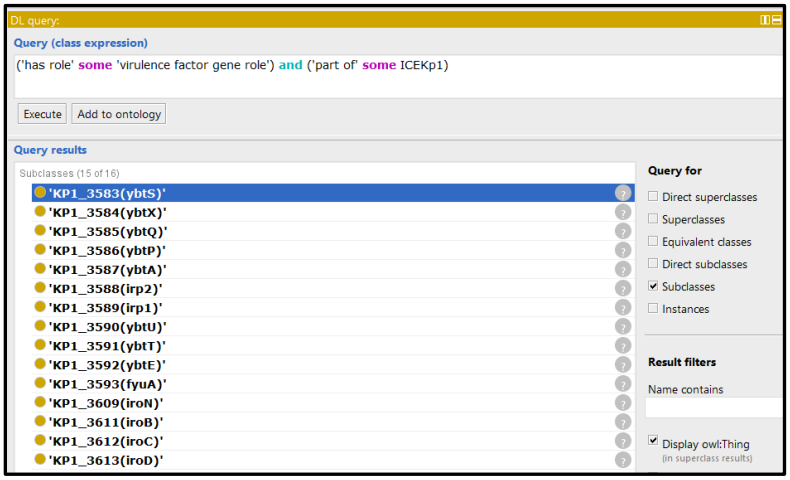

Fig. 5. DL Query window in Protégé 5. *The query of all the virulence factor genes of ICEKp1 was performed using the DL Query of OWL editor Protégé 5.2. The query code is shown on the top, and the query results are displayed at the bottom.*

Use Case 2: Use SPARQL query for advanced analysis

As the OWL-formatted ICEO is stored in the Ontobee RDF triple store (17, 18), the ICEO information can be also queried and analyzed using the RDF query language, SPARQL (https://www.w3.org/TR/rdf-sparql-query-protocol/).

Fig. 6 demonstrates a SPARQL query over ICEO. This example includes only a few lines of SPARQL query code. However, it enabled the identification of the organisms under the taxonomic class of 'Gammaproteobacteria' (NCBITaxon_1236) that include experimentally verified ICE*Kp*1 family ICEs. A variety of queries can be achieved with new SPARQL query scripts for more practical and advanced analysis.

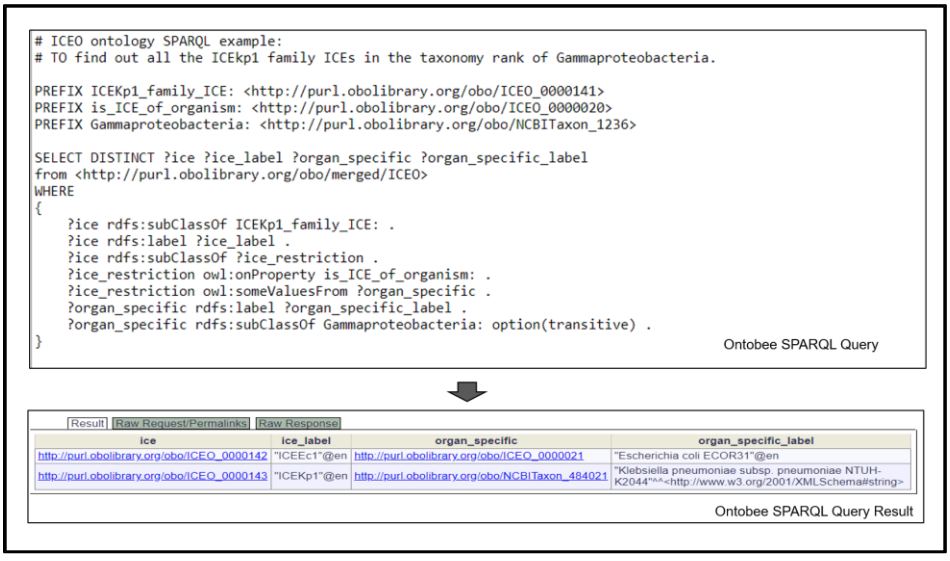

Fig. 6. SPARQL query of all the ICEKp1 family ICEs in the taxonomy rank of Gammaproteobacteria. *The ICEO term ICEO_0000141 is 'ICEKp1 family ICE' class, ICEO_0000020 refers to an object property 'is ICE of organism', and the NCBITaxon term NCBITaxon_1236 points to 'Gammaproteobacteria' class. The query was performed using the Ontobee SPARQL query interface (http://www.ontobee.org/sparql/).*

Discussion

In this study, we developed a community-driven Integrative and Conjugative Element Ontology (ICEO). ICEO ontologically represents the complex hierarchical structure of ICEs, ICE components, and the relations among ICE and ICE components. As demonstrated in two use cases, the ICEO representation of the experimentally verified ICE knowledge supports computer-assisted data integration, efficient query, and reasoning.

ICEO now is built by standardizing and integrating the rich information from ICEberg database. And ICEO can perform tasks that cannot be done in current ICEberg. For example, in our use case 1, we were able to easily query any virulence factors for any level of MGEs. Currently, ICEberg will label VF for those MGEs that are virulence factors. However, it is still impossible for users to query all VFs for a specific bacterial group. In our use case 2, we further illustrate an efficient way of using ICEO to query ICEs under any specific bacterial taxon level like class, species or family. ICEberg can only query based on species level. However, rather than being only a complement or translation of ICEberg database, ICEO and ICEO-based features will be explored to be integrated into ICEberg in the future.

ICEO is the first BFO-based ICE ontology. Toussaint et al. developed the MeGO, a Gene Ontology dedicated to the functions of mobile genetic elements, and used it in the ACALME database (A CLAssification of Mobile genetic Elements, http://aclame.ulb.ac.be/) (24–26). MeGO is a non-OBO ontology expanded from the Phage Ontology (PhiGO). MeGO contains 375 classes, a single object property (which is "*part of*"), and 22 annotation properties. Most of MeGO terms are related to phages, GO, and sequences. Only a few terms directly related to ICE are included in MeGO. MeGO does not include any specific ICEs and ICE gene components. In addition, MeGO terms are poorly aligned. It is also noted that the MeGO and ACALME database have not been updated in the past six years. In comparison, ICEO is systematically developed by aligning with the widely used BFO upper level ontology and following the OBO Foundry principles. ICEO represents the complicated gene components, functional modules and related information about T4SS-type ICEs and AICEs, making it possible to perform the automatically computer-assisted reasoning, query, and advanced analysis of ICE.

However, ICEO is currently at its early development stage and will be further developed in the future. We will represent more known information about T4SS-type ICEs. Due to differences in cell membrane structure, Gram-positive and Gram-negative bacteria have different T4SS organization and features and are associated with different T4SS-type ICEs information. Such differential characteristics will be further categorized, modeled, and represented in ICEO. Besides, only 11 experimentally verified AICE are included in ICEO. Compared with T4SS-type ICEs, AICE is less commonly seen in bacteria. However, AICE is also important in terms of developing useful tools for genetic engineering of *Actinobacteria* (27). In the future, we plan to more systematically represent and analyze AICE information in ICEO.

ICEO will be also be used for more applications. A major application is to apply ICEO as the base of the next version of the ICEberg database to standardize and integrate the rich ICE information for the advanced data sharing and organization. Second, we are going to combine and integrate the structured ICEO ontology into the ICEfinder (3), an ICE prediction tool developed also by our group with both web server and standalone versions available. With the use of ICEO, the enhanced ICEfinder will facilitate more effective, accurate prediction of ICE and powerful sequence analysis from the raw genome sequence. Furthermore, ICEO may also facilitate ontology-based literature mining, which has been shown in many other ontology-based research domains (28, 29).

Conclusions

To conclude, ICEO is a biological ontology and a knowledge-centric platform of the bacterial integrative and conjugative element. ICEO can serve as an ICE knowledgebase and facilitate the systematical representation, integration and automatical computer-assisted reasoning of ICE data.

Acknowledgments

National Key R&D Program of China [2017YFC1600100 to H.Y.O.]; ML was supported by a jointly funded Ph.D.-studentship of the China Scholarship Council and University of Michigan Medical School (Grant No. 201806230209).

Address for correspondence

YH and HYO are the co-corresponding authors. Their email addresses are yongqunh@med.umich.edu and hyou@sjtu.edu.cn respectively. Any suggestions or questions are highly welcomed.

References

1. Johnson,C.M. and Grossman,A.D. (2015) Integrative and Conjugative Elements (ICEs): What They Do and How They Work. Annu Rev Genet, 49, 577–601.

2. Delavat,F., Miyazaki,R., Carraro,N., Pradervand,N. and van der Meer,J.R. (2017) The hidden life of integrative and conjugative elements. FEMS microbiology reviews, 41, 512–537.

3. Liu,M., Li,X., Xie,Y., Bi,D., Sun,J., Li,J., Tai,C., Deng,Z. and Ou,H.-Y. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. Nucleic Acids Res, 10.1093/nar/gky1123.

4. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. Nature Reviews Microbiology, 3, 722–732.

5. Toussaint,A. and Chandler,M. (2012) Prokaryote Genome Fluidity: Toward a System Approach of the Mobilome. In van Helden,J., Toussaint,A., Thieffry,D. (eds), Bacterial Molecular Networks: Methods and Protocols, Methods in Molecular Biology. Springer New York, New York, NY, pp. 57–80.

6. Hoehndorf,R., Schofield,P.N. and Gkoutos,G.V. (2015) The role of ontologies in biological and biomedical research: a functional perspective. Brief Bioinform, 16, 1069–1080.

7. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research, 47, D330–D338.

8. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J., et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol, 25, 1251–5.

9. He,Y., Xiang,Z., Zheng,J., Lin,Y., Overton,J.A. and Ong,E. (2018) The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. J Biomed Semantics, 9, 3.

10. Group,O.W. and others (2009) OWL 2 Web Ontology Language Document Overview: W3C Recommendation 27 October 2009.

11. Wilkinson,M.D., Dumontier,M., Aalbersberg,Ij.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

12. Xiang,Z., Courtot,M., Brinkman,R.R., Ruttenberg,A. and He,Y. (2010) OntoFox: web-based support for ontology reuse. BMC research notes, 3, 175.

13. He,Y., Liu,Y. and Zhao,B. (2014) OGG: a Biological Ontology for Representing Genes and Genomes in Specific Organisms. In ICBO. Citeseer, pp. 13–20.

14. Natale,D.A., Arighi,C.N., Blake,J.A., Bult,C.J., Christie,K.R., Cowart,J., D'Eustachio,P., Diehl,A.D., Drabkin,H.J., Helfer,O., et al. (2014) Protein Ontology: a controlled structured network of protein entities. Nucleic Acids Res, 42, D415–D421.

15. NCBITaxon: An ontology representation of the NCBI organismal taxonomy. http://obofoundry.org/ontology/ncbitaxon.html.

16. Xiang,Z., Zheng,J., Lin,Y. and He,Y. (2015) Ontorat: automatic generation of new ontology terms, annotations, and axioms based on ontology design patterns. Journal of biomedical semantics, 6, 4.

17. Xiang,Z., Mungall,C., Ruttenberg,A. and He,Y. (2011) Ontobee: A linked data server and browser for ontology terms. In ICBO.

18. Ong,E., Xiang,Z., Zhao,B., Liu,Y., Lin,Y., Zheng,J., Mungall,C., Courtot,M., Ruttenberg,A. and He,Y. (2016) Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic acids research, 10.1093/nar/gkw918.

19. Grenon,P. (2003) Spatio-temporality in basic formal ontology Ifomis.

20. Arp,R., Smith,B. and Spear,A. (2015) Building Ontologies using basic formal ontology. Cambridge, MA, USA, 2015.

21. NCBI is phasing out sequence GIs - use Accession.Version instead! (2012) https://www.ncbi.nlm.nih.gov/books/NBK431010/#news_03-02-2016-phase-out-of-GI-numbers.

22. Lin,T.-L., Lee,C.-Z., Hsieh,P.-F., Tsai,S.-F. and Wang,J.-T. (2008) Characterization of Integrative and Conjugative Element ICEKp1-Associated Genomic Heterogeneity in a Klebsiella pneumoniae Strain Isolated from a Primary Liver Abscess. J Bacteriol, 190, 515–526.

23. Wu,K.-M., Li,L.-H., Yan,J.-J., Tsao,N., Liao,T.-L., Tsai,H.-C., Fung,C.-P., Chen,H.-J., Liu,Y.-M., Wang,J.-T., et al. (2009) Genome sequencing and comparative analysis of Klebsiella pneumoniae NTUH-K2044, a strain causing liver abscess and meningitis. J. Bacteriol., 191, 4492–4501.

24. Leplae,R., Hebrant,A., Wodak,S.J. and Toussaint,A. (2004) ACLAME: A CLAssification of Mobile genetic Elements. Nucleic Acids Res, 32, D45–D49.

25. Toussaint,A., Lima-Mendez,G. and Leplae,R. (2007) PhiGO, a phage ontology associated with the ACLAME database. Research in microbiology, 158, 567–71.

26. Leplae,R., Lima-Mendez,G. and Toussaint,A. (2010) ACLAME: A CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Research, 38, D57–D61.

27. Raynal,A., Karray,F., Tuphile,K., Darbon-Rongère,E. and Pernodet,J.-L. (2006) Excisable Cassettes: New Tools for Functional Analysis of Streptomyces Genomes. Appl. Environ. Microbiol., 72, 4839–4844.

28. Hur,J., Ozgur,A., Xiang,Z. and He,Y. (2012) Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining. Journal of biomedical semantics, 3, 18.

29. Hur,J., Özgür,A., Xiang,Z. and He,Y. (2015) Development and application of an interaction network ontology for literature mining of vaccine-associated gene-gene interactions. Journal of Biomedical Semantics, 6, 2.

Responses to the Reviewers:

We appreciate the time and efforts of the reviewers who reviewed our manuscript. Their suggestions and comments are constructive and helpful. We have revised our manuscript by incorporating these comments accordingly as described below. It is noted that the italicized paragraphs below are the reviewer's comments and our followed replies are in regular font style.

---------- REVIEW 1 ----------
SUBMISSION: 23
TITLE: ICEO: a biological ontology for representing and analyzing the bacterial integrative and conjugative element
AUTHORS: Meng Liu, Hong-Yu Ou and Yongqun He

---------- Overall evaluation ----------
SCORE: 2 (accept)

----- TEXT:
Nice, clearly written paper on the development of an application ontology for bacterial mobile genetic elements. Good strong use-case and I liked the inclusion of example queries that the ontology could be used for.

The gene-naming issue seems like it might be an issue long-term, but this is largely out of the authors control and they seem to address it in a pragmatic way.

Reply: Thanks for your kind and positive comments.

The authors mention their use of the "eXtensive Ontology Development (XOD) strategy" which includes the involvement of community efforts to develop ontologies, but how they in fact involve the community is not discussed. The paper would be improved by the authors discussing how they had approached their interaction with the community - e.g. by providing the ontology to experts for review.

Reply: We have submitted the ICEO to the OBO Foundry community and have been checking and refining the ICEO according to the community's comments.

---------- REVIEW 2 ----------
SUBMISSION: 23
TITLE: ICEO: a biological ontology for representing and analyzing the bacterial integrative and conjugative element
AUTHORS: Meng Liu, Hong-Yu Ou and Yongqun He

---------- Overall evaluation ----------
SCORE: 1 (weak accept)
----- TEXT:
This paper describes the translation of a database (ICEberg) of bacterial integrative and conjugative elements (ICEs) into an ontology (ICEO). ICEO builds on existing OBO ontologies such as the Basic Formal Ontology, the Protein Ontology, and the Ontology of Genes and Genomes. The authors explain the design patterns used to build ICEO and demonstrate how the results can be queried. The presentation is clear and the subject is in scope for ICBO, however the novelty is limited.

Reply: We appreciate the reviewer's summary and positive comments.

The greatest shortcoming of this paper is that it does not clearly demonstrate the benefits of the ontological translation. Are the DL and SPARQL queries doing work that could not be done directly in ICEberg?

Reply: This is a good point. We have added a new paragraph in the Discussion part to address the reviewer's comment:

> "ICEO now is built by standardizing and integrating the rich information from ICEberg database. And ICEO can perform tasks that cannot be done in current ICEberg. For example, in our use case 1, we were able to easily query any virulence factors for any level of MGEs. Currently, ICEberg will label VF for those MGEs that are virulence factors. However, it is still impossible for users to query all VFs for a specific bacterial group. In our use case 2, we further illustrate an efficient way of using ICEO to query ICEs under any specific bacterial taxon level like class, species or family. ICEberg can only query based on species level. However, rather than being only a complement or translation of ICEberg database, ICEO and ICEO-based features will be explored to be integrated into ICEberg in the future." (Discussion part, column 1, paragraph 2)

I reviewed the paper and looked at the project's GitHub repository and OWL files. The `iceo_merged.owl` file loaded in Protege and reasoned under HermiT. ICEO claims to be developed according to OBO principles. It uses an OBO namespace <http://purl.obolibrary.org/obo/ICEO_>, however no OBO ID has been requested for ICEO. The paper claims a CC-BY 4.0 license, but the GitHub repository and OWL files contain no license information. I did not see any textual definitions for ICEO terms.

Reply: We have submitted the OBO ID request for ICEO and have been refining the ICEO according to the community's comments. CC-BY 4.0 license and all the textual definitions have been added to the ICEO.

Minor points:

- p2 "eXtensive Ontology Development" should be "eXtensible Ontology Development"
- Figure 3 "particpants in" should be "participates in".

Reply: All the above points have been corrected in the latest version manuscript.

---------- REVIEW 3 ----------
SUBMISSION: 23
TITLE: ICEO: a biological ontology for representing and analyzing the bacterial integrative and conjugative element
AUTHORS: Meng Liu, Hong-Yu Ou and Yongqun He

---------- Overall evaluation ----------
SCORE: 2 (accept)
----- TEXT:
Liu et al. present the ICEO ontology for the representation of bacterial integrative and conjugative element, ICE. In general

the presentation is clear and the paper is fairly well written with just a few errors.

Positive points:
1) The ontology is grounded in BFO and OBO-Foundry principles quite well, and the authors make an appropriate nod to the FAIR principles.
2) The design patterns presented seem appropriate for representing the domain of ICE.
3) The authors present the use of ICEO for querying of ICE via DL and SPARQL, and presumably the results would be useful to researchers in this domain.

Reply: Thanks for your positive comments and advice.

Of interest:
The Authors state: "Since the Protein Ontology (PR) does not include all the proteins included in ICEO. We applied a similar strategy to represent protein names in ICEO (Fig. 2)." The should be a single sentence, but more importantly, the authors should request classes in the PR to cover the protein entities of interest, and then revise their ontology appropriately. This is part of working within the OBO Foundry community.

Reply: Thanks for your advice. We have added a new paragraph in the Results part to address the reviewer's comment:

> "ICE is essentially a genetic feature so that the gene representation is our priority. Since the Protein Ontology (PR) does not include all the proteins included in ICEO, we applied a similar strategy to represent protein names in ICEO (Fig. 2). We have also contacted the Protein Ontology (PR) team, and will request new classes in the PR to improve ICEO." (Results part, column 1, paragraph 4)