CIDO: The Community-based Coronavirus Infectious Disease Ontology

Yongqun He1, Hong Yu2, Edison Ong3, Yang Wang4, Yingtong Liu, Anthony Huffman, Hsin-hui Huang5, John Beverley5, Asiyah Yu Lin5, William D. Duncan5, Sivaram Arabandi6, Jiangan Xie6, Junguk Hur7, Xiaolin Yang8, Luonan Chen, Gilbert S. Omenn9, Brian Athey9, Barry Smith1

1University of Michigan Medical School, Ann Arbor, MI, USA. 2 People’s Hospital of Guizhou University, Guiyang, Guizhou, China. 3 National Yang-Ming University, Taipei, Taiwan. 4 Northwestern University, Evanston, IL, USA. 5 National Center for Ontological Research, Buffalo, NY, USA. 6 Lawrence Berkeley National Laboratory, Berkeley, CA, USA. 7 OntoPro LLC, Houston, TX, USA. 8 School of Bioinformatics, Chongqing University of Posts and Telecommunications, Chongqing, China. 9 University of North Dakota School of Medicine and Health Sciences, Grand Forks, ND, USA. 1 Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & School of Basic Medicine, Peking Union Medical College, Beijing, China. 2 Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China. 3 University at Buffalo, Buffalo, NY 14260, USA.

Abstract. Current COVID-19 pandemic and previous SARS/MERS outbreaks have caused a series of major crises to global public health. We must integrate the large and exponentially growing amount of heterogeneous coronavirus data to better understand coronaviruses and associated disease mechanisms, in the interest of developing effective and safe vaccines and drugs. Ontologies have emerged to play an important role in standard knowledge and data representation, integration, sharing, and analysis. We have initiated the development of the community-based Coronavirus Infectious Disease Ontology (CIDO). As an Open Biomedical Ontology (OBO) library ontology, CIDO is an open source and interoperable with other existing OBO ontologies. In this article, the general architecture and the design patterns of the CIDO are introduced, CIDO representation of coronaviruses, phenotypes, anti-coronavirus drugs and medical devices (e.g., ventilators) are illustrated, and an application of CIDO implemented to identify repurposable drug candidates for effective and safe COVID-19 treatment is presented.

Keywords. Coronavirus, COVID-19, ontology, drug repurposing, ventilator.

1 Introduction

Coronavirus diseases pose major crises to public health. In addition to the current COVID-19 pandemic, Severe Acute Respiratory Syndrome (SARS) [1] and Middle East respiratory syndrome (MERS) [2] are two other severe HCoV diseases that have

---

1 Yongqun Oliver He, Corresponding author. E-mail: yongqunh@med.umich.edu.
2 Hong Yu, Co-corresponding author. E-mail: yuhong20040416@sina.com.
occurred in the past two decades. More recently, the WHO declared the Coronavirus Disease 2019 (COVID-19) outbreak as a pandemic on March 11, 2020, when there were 118,326 confirmed cases and 4,292 deaths. As of May 25, there have been over 5.5 million confirmed cases and approximately 350,000 deaths globally.

A major bottleneck for coronavirus disease research, however, is that valuable research data and knowledge are siloed in non-integratable and non-interoperable data repositories. Big data has the characteristics of 5 V’s: volume, veracity (i.e., quality of data), velocity, variety, and value [3]. In the era of Information Technology and big data, biomedical research has become data-intensive with increasingly large, complex, multidimensional, and diverse datasets generated. Knowledge is a special type of data that embodies awareness and understanding. Non-integrated and non-interoperable data and knowledge inhibit computer-assisted reasoning, which is the essence of Artificial Intelligence. To address this bottleneck, we can rely on computer-interpretable integrative and interoperable ontology.

With that in mind, to study coronaviruses efficiently we have initiated the development of a community-based interoperable Coronavirus Infectious Disease Ontology (CIDO) for standardized and integrative coronavirus disease representation, integration, and analysis. This manuscript provides details on current status of CIDO development and applications.

2. Methods

2.1. Coronavirus disease-related data collection

The coronavirus disease related data were collected from the literature and openly available databases. We have focused on coronavirus and host taxonomy data, phenotype, drug, and vaccine data.

2.2. Ontology development

CIDO development followed the OBO Foundry ontology development principles [4] and the eXtensible Ontology Development (XOD) strategy [5]. The CIDO development started with the reuse and alignment of terms and relations from existing ontologies using the Ontofox tool [6]. We used the Basic Formal Ontology (BFO) [7], Chemical Entities of Biological Interest (ChEBI) [8], Human Disease Ontology (DOID) [9], Human Phenotype Ontology (HPO) [10], Infectious Disease Ontology (IDO) [11], etc. The Protégé-OWL editor was used for ontology editing.

2.3. CIDO status, source code, deposition, and license

CIDO has been accepted as an OBO library ontology (http://www.obofoundry.org/ontology/cido.html). CIDO source code is freely available on the GitHub website https://github.com/CIDO-ontology/cido. The source code uses the license CC-BY. CIDO has been deposited to the Ontobee ontology repository (http://www.ontobee.org/ontology/CIDO) the BioPortal repository (https://bioportal.bioontology.org/ontologies/CIDO), and the OLS repository (https://www.ebi.ac.uk/ols/ontologies/cido).
2.4. **CIDO representation demonstrations and drug repurposing use case**

In this manuscript, we provide demonstrations of different CIDO ontological representations of coronaviruses, drugs, vaccines, and medical devices. We also show how CIDO can be used to support rational design for drug repurposing. Description Logic (DL) and SPARQL queries are used in Protégé–OWL editor for ontology content query.

3. **Results**

3.1. **High level structure**

Figure 1 lays out the high-level hierarchical structure of CIDO, as well as the various ontologies which it imports. Abbreviations in parentheses indicate an entity source ontology, and red text indicates terms introduced by CIDO:

![Figure 1. Top level hierarchical structure of CIDO. Abbreviations not defined earlier: GO – Gene Ontology, NCBITaxon – NCBI Taxonomy Ontology, OBI – Ontology for Biomedical Investigations, OGMS – Ontology for General Medical Science, OHPI – Ontology of Host-Pathogen Interaction, OPMI – Ontology of Precision Medicine and Investigation.](image)

As illustrated above, CIDO uses the Basic Formal Ontology (BFO) as the top-level ontology. BFO has been approved to become ISO/IEC standard 21838. BFO contains two branches, ‘continuant’ and ‘occurrent’. The ‘continuant’ represents time-independent entities such as material entity, and the ‘occurrent’ represents time-related entities such as process. BFO has been adopted by >300 ontologies. BFO provides us a mechanism to reuse and integrate with these many ontologies without interoperability issues, and will ultimately facilitate CIDO becoming a standard widely-used ontology for coordinating and integrating coronavirus research.

Figure 2 describes the general CIDO design pattern that lays out the relations among selected major entities modeled in the ontology.
One central term in CIDO is ‘COVID-19 disease process’ (Figure 2), which is defined as a subclass of ‘disease process’, which is defined as a subclass of ‘pathological bodily process’ (OGMS_0000061). The ‘COVID-19 disease process’ has many features including those defined using the following axioms:

- ‘occurs in’ some animal
- ‘occurs in’ some lung, where lung is part of animal.
- ‘caused by infection with’ some SARS-CoV-2
- realizes some ‘COVID-19 disease’, which is a disposition defined in Human Disease Ontology (DOID).

Based on the “all-some rule” in ontology [12], the axiom (‘occurs in’ some lung) means that every COVID-19 disease process occurs in some lung. We know that the disease can also occur in other organs such as kidney and brain; however, the occurrence in kidney or brain does not always happen. In this case, we can use the relation ‘susceptibly occurs in’, which represents a susceptibility of the occurrence rather than an assurance.

In addition, ‘COVID-19 vaccine’ ‘immunizes against disease process’ the ‘COVID-19 disease process’, where the relation is a Vaccine Ontology (VO) relation (Figure 2). Furthermore, ‘COVID-19 drug’ ‘is substance that treats’ the ‘COVID-19 disease process’, and here the relation is a Relation Ontology (RO) term.

Instead of directly using the disposition term ‘COVID-19 disease’, we chose to use ‘COVID-19 disease process’ due to multiple reasons. First, the relation ‘occurs in’ only links a process to an anatomic entity (e.g., lung), and it cannot link a disposition to an anatomic entity. The relation ‘caused by infection with’ is also defined to link a pathological process (not a disposition) with a pathogen. Furthermore, the vaccines and drugs are used for an existing disease process instead of a disease disposition. Meanwhile, the ‘COVID-19 disease’ disposition is realized in the ‘COVID-19 disease process’, which closely links the two terms together.

Figure 2 also lays out the genes, cells, quality types (e.g., phenotypes, age) and their relations with each other and other entities in CIDO. For example, human ACE2 gene, which encodes for the Angiotensin I Converting Enzyme 2 (ACE2), is a receptor of the SARS-CoV-2 S protein. The ACE2-S binding in the epithelial cells of the lung would
stimulate a list of host genes up- or down-regulated in the cells, which may further contribute to the disease development.

3.2. **Classification of coronaviruses, hosts, and host phenotypes**

CIDO includes taxonomic representations of coronaviruses and hosts, and the phenotypes shown in coronavirus patients. Figure 2A represents the taxonomic hierarchy of representative coronaviruses. SARS-CoV and SAR-CoV-2 belong to the Sarbecovirus, a subclass of Betacoronavirus genus. MERS-CoV belongs to Merbecovirus subgenus under the same Betacoronavirus genus. Four human coronavirus strains (229E, NL63, HKU1, and OC43) cause mild common colds in humans, and they belong to different taxonomic locations compared to SARS-CoV, SARS-CoV-2, and MERS-CoV. Figure 2B shows the common phenotypes shown in COVID-19 patients, including fever, chills, cough, shortness of breath (dyspnea), loss of smells and taste, etc. These terms have all been extracted from NCBITaxon and HPO and imported to CIDO.

![Figure 3. Ontological representation of representative coronaviruses and COVID-19 phenotypes.](image)

(A) NCBITaxon representation of the hierarchy of representative coronaviruses. (B) HP representation of common phenotypes shown in COVID-19 patients. These hierarchical terms are imported to CIDO.

3.3. **Anti-coronavirus drug classification**

Extending our previous preprint work [13], we collected 136 anti-coronavirus drugs, including 26 coronavirus protein-specific monoclonal or polyclonal antibody drugs and 110 general purpose drugs. These drugs have been experimentally verified *in vivo* or *in vitro* to be effective against the infections of various human coronaviruses including SARS-CoV, MERS-CoV, and SARS-CoV-2.

We have mapped all these drugs to three ontologies: ChEBI [8], National Drug File – Reference Terminology (NDF-RT) [14], and the Drug Ontology (DrON) [15]. The subsets of these ontologies that contain the mapped drugs and their related characteristics were then extracted and imported to CIDO using the Ontofox tool [6].
Through the mechanism of Emergency Use Authorization (EUA), FDA has authorized two COVID-19 drugs for the emergency treatment of COVID-19 patients: Remdesivir drug (https://www.fda.gov/media/141477/download) and COVID-19 Convalescent Plasma (CCP) (https://www.fda.gov/media/137566/download). The following axiom defines the usage of Remdesivir and CCP for treating COVID-19:

\[ \text{is substance that treatments some 'COVID-19 disease process'} \]

The following axiom represents the relation between the drug ‘Remdesivir drug’ and its active ingredient:

\[ \text{has active ingredient some remdesivir} \]

As defined in ChEBI, remdesivir (http://purl.obolibrary.org/obo/CHEBI_145994) has many properties such as the antiviral role and anticonvoviral agent role.

3.4. Anti-coronavirus vaccine representation in CIDO

In another recent study published in a journal article [17], we reported our COVID-19 vaccine design using reverse vaccinology and machine learning. This study collected 44 vaccine candidates experimentally verified to be effective against coronavirus challenges in laboratory animal models and/or used in clinical trials. The Vaccine Ontology (VO) [18] has been used to represent these vaccines. The VO-represented terms and relations were then imported to CIDO (data not shown). Furthermore, we have used CIDO and other ontologies including the Ontology of Adverse Events (OAE) to systematically examine the adverse events associated with SARS/MERS/COVID-19 vaccine candidates, and found differential profiles in different types of vaccines (data not shown).

3.5. Medical device for COVID-19 treatment

During this unprecedented time of COVID-19 pandemic, due to the rapid spread of SARS-CoV-2, the control of the pandemic requires large amounts of medical devices in all healthcare settings, especially the nursing homes and or long-term care facilities where the most vulnerable population lives in high density. A medical device is any device intended to be used for medical purposes. The medical devices used for the pandemic include diagnosis testing kits, ventilators, and personal protective equipment (PPE) for medical use such as surgical masks, face shields, respirators, gowns, gloves etc. In the USA, the FDA is responsible for regulating medical devices used to diagnose, prevent and treat COVID-19. The NIH has also developed COVID-19 treatment guidelines to inform clinicians how to care for patients with COVID-19 (https://www.covid19treatmentguidelines.nih.gov).

Some COVID-19 patients may develop acute respiratory failure and require ventilatory support. As a specific use case, we focus the ontological representation on ventilators. A ventilator is defined as a machine that supports breathing when a person is unable to breathe enough on their own. The modern ventilator technology has evolved to enable a single device to operate in numerous modes, from basic continuous positive pressure (CPAP and bi-level PAP) to traditional pressure and volume ventilator modes. As a complicated machine, the modern ventilator system includes an operator interface where parameters will be set for proper management of the automatic ventilation supported by a mechanical ventilator. The central part of the ventilator is like a “black box”. The output of the ventilator can be measured and monitored to fit the patient’s condition. We developed a framework to demonstrate how CIDO models the entities involved in a mechanical ventilation assistance and management process (Figure 5).
A ‘mechanical ventilator’ is a subclass of ‘ventilator’, which is a subclass of ‘medical device’. The function of the ‘ventilator’ is to ‘deliver mixed gas into lung’. Multiple entities involved in a ‘mechanical ventilator operation’ process: a ‘mechanical ventilator’, a clinical ‘ventilator operator’, a ‘patient’ and the preset parameter of ‘tidal volume’. As a planned process, the operation of the mechanical ventilator aims to achieve the objective of assisting patient’s breathing that ‘is about treating’ the ‘acute respiratory failure disease process’, a pathological process that a ‘COVID-19 patient’ may be susceptible to. Note that the low tidal volume (4–8 mL/kg of predicted body weight) is an important parameter for the mechanical ventilation operation in the case of caring acute respiratory failure without damaging the lung (https://www.covid19treatmentguidelines.nih.gov/critical-care/).

3.6. Statistics of CIDO

As of August 30, 2020, CIDO has 6,758 terms, including imported terms from over 40 existing ontologies. CIDO includes 190 CIDO-specific terms and hundreds of newly generated axioms that link different entities. The detailed statistics of CIDO is available at the Ontobee website: http://www.ontobee.org/ontostat/CIDO.

3.7. Use cases of CIDO

CIDO can be applied for many use cases. Below we will focus on drug repurposing demonstration and then introduce a few other use cases.

Drug repurposing is a strategy to identify new uses for approved drugs that are outside the scope of the original medical indication. We previously identified 110 ant coronovirus drugs [13], most of which were not originally targeted for COVID-19 treatment. We hypothesize that some of these drugs could be repurposed for treating COVID-19 patients. To address this hypothesis, we performed ontology-based bioinformatics analysis with the assumption that ontology-formatted computer-interpretable results provided us a new way to identify more scientific insights. As shown in Figure 5, a DL query of CIDO identified 6 chemical compounds that have the antiviral
agent role and antimalarial role and has been found effective against the infection of SARS-CoV, SARS-CoV-2, or MERS-CoV. Our further analysis found that 3 out of these 6 chemicals (amodiaquine, chloroquine, hydroxychloroquine) all have the role of inhibiting viral entry to host cells, and all these three chemicals are all under the class of quinolines. Furthermore, among these anticoronavirus chemicals, a query to find an entity having all of antiviral, antimalarial, and anti-inflammatory agent roles retrieved only one chemical compound - amodiaquine (Figure 6).

![Figure 5. DL query of potential drugs for potential COVID-19 treatment. (A) DL query of chemicals that have two roles (antiviral agent role and antimalarial role). (B) DL query of chemicals that have three roles including the two roles above and the anti-inflammatory agent role.]

CIDO is also being used in many other applications. For example, CIDO is being used by OHNLP (http://www.ohnlp.org/), the Open Health Natural Language Processing (NLP) Consortium aimed to promote clinical COVID-19 data NLP. The OHNLP COVID-19 research is also an effort of the National COVID Cohort Collaborative (N3C) program (https://ncats.nih.gov/n3c). CIDO has also been applied to support COVID-19 vaccine study and semantic calculation and prediction.

4. Discussion
This manuscript introduces the initial development and applications of the community-based Coronavirus Infectious Disease Ontology (CIDO). Our study demonstrates that CIDO provides an ideal platform to integrate important data needed to research different coronavirus disease-related entities such as viruses, phenotypes, drugs, vaccines, medical devices (e.g. ventilators), supporting integrative analysis of COVID-19 and other human coronavirus diseases.

We provide one CIDO use case concerning drug repurposing for COVID-19 treatment. Despite extensive research and clinical trials, we as of now have found only one potentially effective therapy for COVID-19, Remdesivir [19]. Our ontology-based analysis provides a new strategy for rational drug repurposing design (Figure 4), complementing others [20, 21]. Eight active drug chemicals were found to have antiviral and antimalarial roles, and amodiaquine was found to not only have these two roles but also an additional anti-inflammatory role. Similar to chloroquine, Amodiaquine is a 4-aminoquinoline that has been used widely to treat and prevent uncomplicated malaria [22]. A comparative study showed that amodiaquine is more effective than chloroquine in the treatment therapy of falciparum malaria in Kenya [22]. It has also been recommended that combining amodiaquine with artesunate to reduce the risk of drug
resistance [23]. The performance of amodiaquine for possible COVID-19 treatment deserve further evaluation [24].

An ongoing CIDO development is to model and represent the mechanisms of the molecular and cellular interaction between the hosts and coronaviruses. Such modeling will provide the foundation for our rational drug repurposing and vaccine development. In our drug preprint study [13], we extracted and analyzed the interactions between anti-coronavirus drugs and their target proteins. These anti-coronavirus drugs were identified to be effective against coronavirus infections in vitro or in vivo. It is likely that some of the drug targets participate in active host-SARS-CoV-2 interactions leading to severe COVID-19 disease outcomes. Deeper modeling and representation of the intricate host-virus-drug interactions would help us in better drug repurposing analysis.

Being a “community-based” ontology, CIDO is committed to serve the community. CIDO is created to be open and freely available for the community to use. It is an interoperable ontology that reuses and interlinks to existing ontologies and resources. We also expect that the community gets involved in its development, and we are always ready to accept new ideas and critiques.

In future work we will use CIDO as a platform to standardize different coronavirus research-related metadata types and apply them for the standardization and enhanced analysis of specific conditions defined in different experimental and clinical studies, and how these conditions would affect the disease outcomes. We will also identify and develop more applications tools that implementing CIDO for different purposes. More researchers and developers are welcome to join our community-based effort to advance the CIDO ontology and its applications.

Acknowledgements
This project is supported by NIH grants U24CA210967 and P30ES017885 (to GSO); R01GM080646, 1UL1TR001412, 1U24CA199374, and 1T15LM012495 (to BS); the National Natural Science Foundation of China 61801067 (to JX); the Natural Science Foundation of Chongqing CSTC2018JCYJAX0243 (to JX); the non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences 2019PT320003 (to HY); and University of Michigan Medical School Global Reach award (to YH).

References