The Cell Line Ontology integration and analysis of the knowledge of LINCS cell lines

Edison Ong[1], Jiangan Xie[1], Zhaohui Ni[1], Qingping Liu[1], Yu Lin[2], Vasileios Stathias[2], Caty Chung[2], Stephan Schurer*[2], Yongqun He*[1]
[1] University of Michigan, Ann Arbor, Michigan, USA; [2] University of Miami, Coral Gables, FL, USA

***Abstract*— Cell lines are crucial to study molecular signatures and pathways, and are widely used in the NIH Common Fund LINCS project. The Cell Line Ontology (CLO) is a community-based ontology representing and classifying cell lines from different resources. To better serve the LINCS research community, from the LINCS Data Portal and ChEMBL, we identified 1,097 LINCS cell lines, among which 717 cell lines were associated with 121 cancer types, and 352 cell line terms did not exist in CLO. To harmonize LINCS cell line representation and CLO, CLO design patterns were slightly updated to add new information of the LINCS cell lines including different database cross-reference IDs. A new shortcut relation was generated to directly link a cell line to the disease of the patient from whom the cell line was originated. After new LINCS cell lines and related information were added to CLO, a CLO subset/view (LINCS-CLOview) of LINCS cell lines was generated and analyzed to identify scientific insights into these LINCS cell lines. This study provides a first time use case on how CLO can be updated and applied to support cell line research from a specific research community or project initiative.**

Keywords*— Cell line, cell, ontology, CLO, LINCS, ChEMBL

I. Introduction

The NIH Common Fund Library of Integrated Network-based Cellular Signatures (LINCS) program aims to create a network-based biological understanding of gene expression and cellular processes when cells are exposed to various perturbing agents (http://www.lincsproject.org/). Over 1000 cell lines have been used in LINCS and play a critical role as disease model systems to produce molecular and cellular signatures and networks.

The Cell Line Ontology (CLO) is a community-based ontology system for representing cell lines [1]. The overall goal of this study is to use CLO to represent and integrate the knowledge of LINCS cell lines in order to power LINCS cell lines' integrity across multiple resources.

II. Methods

A. Information extraction and data mapping

Two sources, including the LINCS Data Portal (http://lincsportal.ccs.miami.edu/entities/) and ChEMBL [2], were used to obtain LINCS cell line information. The data in these two sources were compared and mapped to the CLO knowledge base, and new information was identified.

B. CLO modeling nd design pattern generation

Based on the data types obtained from the mapping process, an updated CLO design pattern model was generated in order to accommodate new LINCS cell line data attributes.

C. New information incorporation into CLO

Based on the new design patterns, Ontorat (http://ontorat.hegroup.org) was used to incorporate LINCS cell line data from different data sources to CLO. Manual checking was performed to ensure correctness.

D. Generation and analysis of a LINCS cell line set of CLO

OntoFox (http://ontofox.hegroup.org) was used to generate a CLO subset (LINCS-CLOview) that includes all LINCS cell lines, as shown here: https://raw.githubusercontent.com/CLO-ontology/CLO/master/src/ontology/LINCS-CLOview.owl. The LINCS CLO subset was also submitted to Ontobee (http://www.ontobee.org). The information of the subset was visualized using Protégé OWL editor, queried using Ontobee SPARQL web program, and further analyzed.

III. Results

A. LINCS cell line information extraction and mapping from different resources

As of April 15, 2016, 1,097 cell lines were extracted from the LINCS Data Portal. Among these LINCS cell lines, 794 cell lines could be directly mapped to CLO. Meanwhile, ChEMBL included 637 cell line entries with LINCS IDs. Among these, 451 cell lines have CLO_IDs, and 51 out of the remaining 186 cell lines could be mapped to CLO using name matching. The data types available related to these cell lines in the LINCS Portal and ChEMBL are shown in Fig. 1.

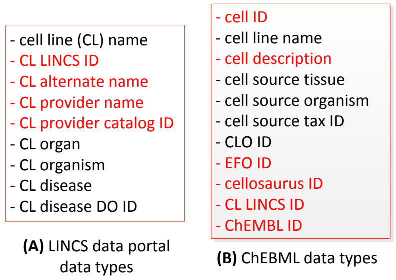

Fig. 1. Cell line-related data types of the data downloaded from LINCS Data Portal and ChEMBL. (A) Data types from LINCS Data Portal. (B) Data types from ChEMBL. Red-highlighted items (e.g., ChEMBL ID) were not covered in CLO, which were added later to CLO in this study.

*: corresponding authors: stephan.schurer@gmail.com; and yongqunh@med.umich.edu

Among the total of 1097 LINCS cell lines each with a unique LINCS cell line ID (e.g., LCL-1512 for HeLa cell), 466 have ChEMBL, LINCS, and CLO IDs, 279 have LINCS and CLO IDs, and 352 LINCS cell lines do not have any CLO IDs.

B. *CLO modeling and design pattern generation*

To represent the new database information to a specific cell line (Fig. 1), we used 'seeAlso' relation. For example, for the HeLa cell (CLO_0003684), we added to CLO: '*Cell line LINCS ID*: LCL-1512' and '*seeAlso*: EFO: EFO_0001185; CHEMBL: CHEMBL3308376; CVCL: CVCL_0030'.

To more conveniently link a specific cell line and a disease, we have also generated a new shortcut relation 'derived originally from patient with disease' (Fig. 2).

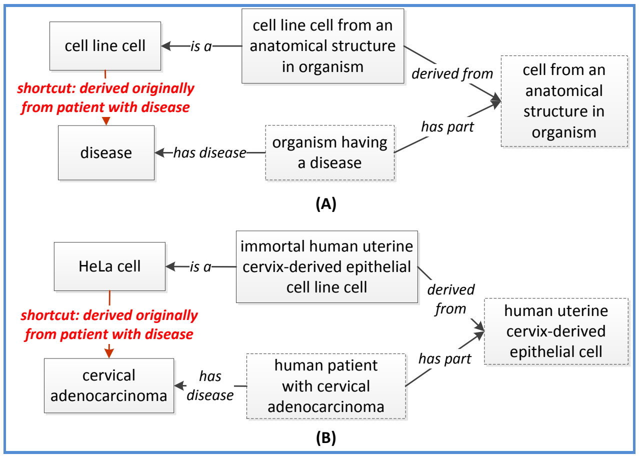
cell line cell
is a
cell line cell from an anatomical structure in organism
derived from
cell from an anatomical structure in organism
shortcut: derived originally from patient with disease
has part
disease
has disease
organism having a disease
(A)
HeLa cell
is a
immortal human uterine cervix-derived epithelial cell line cell
derived from
human uterine cervix-derived epithelial cell
shortcut: derived originally from patient with disease
has part
cervical adenocarcinoma
has disease
human patient with cervical adenocarcinoma
(B)

Fig. 2. CLO design pattern model for using the new shortcut relation 'derives originally from patient having disease'. (A) General design pattern; (B) an example to illustrate the design pattern. The shortcut relation makes it more efficient to represent the relation between a cell line cell and a disease when the parent term of the cell line cell includes sufficient information about the cell type and tissue/organ. In this illustration, the classes as shown in the dotted boxes are redundant and are not needed.

C. *New data integration to CLO and CLO subset generation*

Based on the mapping and the design pattern models (Fig. 1 and 2), extra data available in the LINCS Data Portal and ChEMBL were integrated into to CLO.

A CLO subset of LINCS cell lines (LINCS-CLOview) was also generated. LINCS-CLOview can be considered as a CLO "community view" [3] for the LINCS research community. As of May 1, 2016, LINCS-CLOview contained 1,924 terms, including 1,825 classes, 25 object properties, 61 annotation properties, and 13 instances. These terms include 1,315 terms with CLO IDs. The other terms were imported from 17 other ontologies. Detailed statistics of LINCS-CLOview is shown: http://www.ontobee.org/ontostat/LINCS-CLOview.

D. *Analysis of LINCS cell lines by querying LINCS-CLOview*

With the availability of LINCS-CLOview, we were able to analyze LINCS cell lines from different aspects.

Our study found that LINCS cell lines are associated with 121 diseases. These 121 diseases include three benign neoplasms, i.e., breast fibrocystic disease (associated with MCF 10A and MCF 10F cells). The other 118 diseases are various types of cancers. The hierarchical structure of these diseases under the Disease Ontology (DOID) also helped the understanding of all the diseases associated with LINCS cell lines. For example, 19 LINCS cell lines (e.g., HeLa cell) were derived from patients with cervical adenocarcinoma, 4 with cervical clear cell adenocarcinoma (a specific type of cervical adenocarcinoma), and 14 with cervical squamous cell carcinoma. These diseases all belong to cervix carcinoma.

We also examined the tissue and organ types from which the LINCS cell lines were derived. In CLO, the multi-species anatomy ontology UBERON is used to represent tissues and organs. In total 131 UBERON terms have been used in LINCS-CLOview to refer to various anatomic locations from which LINCS cell lines were derived.

The cell types of LINCS cell lines were analyzed. The Cell Type Ontology (CL) was used in CLO to demonstrate the cell types of different cell lines. In total, 43 CL cell types, such as epithelial cell, B cell, and T cell, are included in LINCS-CLOview. Each of these cell types is linked to different cell line cells. For a project to study cellular signatures related to a specific cell type, the LINCS-CLOview provides a feasible method to identify which cell line cells to use.

IV. Discussion

This article is the first report of developing a CLO community view to serve a specific community, in this case, the LINCS research community. Since tens of thousands of cell lines have been represented in CLO, it is inefficient to use the whole CLO for LINCS cell line related research. The generation of LINCS-CLOview allows standardization and modularization of the LINCS cell lines, which facilitates the better analysis and reuse of the LINCS cell line information.

Acknowledgment

This work was supported by grant U54HL127624 (BD2K LINCS Data Coordination and Integration Center, DCIC) awarded by the National Heart, Lung, and Blood Institute through funds provided by the trans-NIH LINCS Program and the trans-NIH Big Data to Knowledge (BD2K) initiative (http://www.bd2k.nih.gov). LINCS is an NIH Common Fund projects. This project was also supported by a BD2K-LINCS DCIC external data science research award.

References

[1] S. Sarntivijai, Y. Lin, Z. Xiang, T. F. Meehan, A. D. Diehl, U. D. Vempati, *et al.*, "CLO: The Cell Line Ontology," *J Biomed Semantics,* vol. 5, p. 37, 2014.

[2] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, *et al.*, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Res,* vol. 40, pp. D1100-7, Jan 2012.

[3] J. Zheng, Z. Xiang, C. J. Stoeckert, Jr., and Y. He, "Ontodog: a web-based ontology community view generation tool," *Bioinformatics,* vol. 30, pp. 1340-2, Feb 1 2014.