

Advanced data curation in GTRD database: hierarchical dictionaries of cell types and experimental factors

Mikhail A. Kulyashov
 BIOSOFT.RU, LLC
 Novosibirsk State University
 Institute of Cytology and Genetics
 SB RAS, Novosibirsk, Russia
 Institute of Computational
 Technologies SB RAS
 Novosibirsk, Russia
 m.kulyashov@mail.ru

Semyon K. Kolmykov
 BIOSOFT.RU, LLC
 Institute of Computational
 Technologies SB RAS
 Institute of Cytology and Genetics
 SB RAS
 Novosibirsk, Russia
 kolmykovsk@gmail.com

Ivan S. Yevshin
 BIOSOFT.RU, LLC

Fedor A. Kolpakov
 Institute of Computational
 Technologies SB RAS
 BIOSOFT.RU, LLC
 Novosibirsk, Russia
 fkolpakov@gmail.com

Abstract — GEO database contains a lot of different experiments about transcription regulation. Most part of such experiments (ChIP-seq, DNase-seq and Histone-ChIP-seq) for 9 species (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Arabidopsis thaliana*) were annotated and uniformly processed in GTRD database. Here we are describing the latest advances in these data annotations for GTRD database to formalise experimental data: hierarchical dictionaries of cell types and experimental factors. This approach helps us to integrate experimental data by cell lines (cell types and tissues) and experimental conditions as well as simplifies searching for relevant information for a user.

Keywords — *GTRD*, *database structure*, *experiments annotation*, *experiments comparison*

Introduction

GTRD database contains a large number of ChIP-seq, DNase-seq and Histone-ChIP-seq experiments which were manually annotated [1].

ChIP-seq experiment shows binding regions for transcription factors. DNase-seq experiment gives an opportunity to identify regions of open chromatin. Histone ChIP-seq shows a presence of different types of histones modifications, such as methylation, acetylation and e.t.c. It should be noted that results of such experiments are specific for cell type and experimental conditions (different treatments, conditions of experiment, characteristics of donors such as sex or age and many other).

An important task is integration of such data for a given cell type and experimental condition to get an integrated view on main events related with transcription regulation. This is one of the main goals of GTRD database.

To solve this task we are building a hierarchical dictionary which contains all cell lines, tissues and cell types

which were described in GTRD as well as a hierarchical dictionary of different experimental conditions. Using this approach we also can compare experiments on the same cell lines, cell types or tissues in different databases. This is further development of our previous work [2].

Materials and Methods

Dictionary of cell types and tissues:

All cell lines, tissues and cell types which were used in experiments, described in GTRD, contain 2 fields that were used for building the dictionary [2]:

Source - tissues or cell types from which cells were isolated in experiments or system, which includes this cell, if isolation region was not indicated;

Cell type - cells on which experiment was made.

First step in process of creation a hierarchical dictionary was to separate sources to *main clusters* - a global groups which unify different sources, such organs, embryonic parts, cell lines and some other in union groups:

Cell line - this group includes cell lines, which (permanently established cultures) like MCF-7 cell line, or cell types, which were derived from cell lines, for example WA09 derived fibroblasts.

Embryo - are all cell types, which were described as whole embryos and this group includes a lower cluster embryonic cells in which presented all cell types and tissues which were derived from embryos, like mesendoderm, embryonic organs e.t.c.

Stem cell - stem cell of different stages (totipotent, pluripotent, multipotent, e.t.c.), and all cell types which were derived from stem cells, for example embryonic stem cells derived epithelial cells.

Multicellular organism - this cluster includes all tissues and cell types in which were used whole multicellular

organism and all smaller clusters which describe different tissues of the organism, for example brain, spleen, bone-marrow and many others.

Whole organism - whole unicellular organisms, which are included in GTRD database, in different strains and developmental stages.

Other - this cluster includes all cell types which now are not classified to any cluster.

Dictionary of experiments description:

Dictionary of *experimental factors (ExpFactors)* includes all factors which can influence an experiment. This dictionary is creating simultaneously with the process of manual annotation and all these factors are adding by annotator.

Similar with cell types we also can group experimental factors into clusters like treatment, conditions, age e.t.c., main of them were described in our previous work [2]. They are the starting point of corresponding classification trees. Then we can manually specify child-parent relationships, for example, "doxycycline" will be parent for "doxycycline for 45 mins".

Database structure:

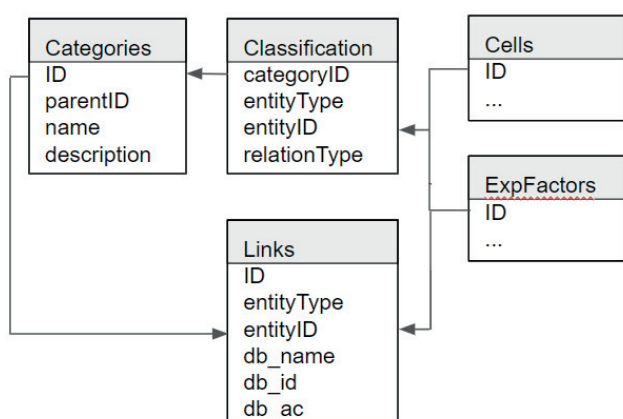


Fig. 1. Fragment of GTRD database structure for hierarchical classification of cell types and experiment conditions

We extended GTRD database scheme for hierarchical classification of data presented in tables Cells (cell types) and ExpFactors (experimental factors):

Categories - describes classification tree where the clusters described above are the roots (parentID is null).

Classification - links flat dictionary of cell types and experimental factors with categories tree.

Links - links Categories, Cells and ExpFactors with specialized ontologies: and linked it with specialized ontologies: BRENDA [3], Cell ontology [4], Uberon [5], Celosaurus [6], Experimental factor ontology [7].

Tables described above were filled semi-automatically. Part of information was extracted from mentioned above specialised ontologies which were downloaded in OBO format and translated into relational database. Then this information was checked and supplemented by annotator using specialized GTRD tools.

Results

Hierarchical dictionaries for all cell types and experimental conditions for annotated in GTRD database were developed. This approach helps us to integrate experimental data by cell types and experimental conditions as well as simplifies searching for relevant information for a user.

ACKNOWLEDGMENT

This work was supported by the Russian Science Foundation. (grant No. 19-14-00295).

REFERENCES

- [1] Yevshin I., Sharipov. R., Kolmykov S., Kondrakhin Y., Kolpakov F. "GTRD: a database on gene transcription regulation—2019 update", Nucleic Acids Research. 2019, Volume 47, Issue D1, Pages D100–D105.
- [2] Kulyashov M., Kolmykov S., Yevshin I., Kolpakov F. "Description, Characteristic And Algorithm For Creation Of A Dictionary Of Cell Types And Tissues In The Gtrd Database", CEUR-Workshop proceedings, 2020, Vol-2569, p.13-18.
- [3] Gremse M, Chang A, Schomburg I, et al. "The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources", Nucleic Acids Res. 2011, D507–D513.
- [4] Bard J., Rhee S.Y., Ashburner M. "An ontology for cell types", Genome Biol. 2005;6(2):R21.
- [5] Mungall C.J., Torniai C., Gkoutos G.V. et al. "Uberon, an integrative multi-species anatomy ontology", Genome Biol. 2012, 13, R5.
- [6] Bairoch A. "The Cellosaurus, a Cell-Line Knowledge Resource", J Biomol Tech. 2018, 29(2):page 25–38.
- [7] Malone J., Holloway E., Adamusiak T., et al. "Modeling Sample Variables with an Experimental Factor Ontology", Bioinformatics. 2010, 26(8):1112-1118.
- [8] Kolpakov F., Akberdin I., Kashapov T., Kiselev I., Kolmykov S., Kondrakhin Y. et al. "BioUML: an integrated environment for systems biology and collaborative analysis of biomedical data", Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W225–W233