SOFTWARE TOOL ARTICLE

Identifier Mapping in Cytoscape: idmapper [version 1; referees: 1 approved, 1 approved with reservations]

Adam Treister, Alexander R. Pico
Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, 94158, USA

Abstract
Identifier Mapping, the association of terms across disparate taxonomies and databases, is a common hurdle in bioinformatics workflows. The idmapper app for Cytoscape simplifies identifier mapping for genes and proteins in the context of common biological networks. This app provides a unified interface to different identifier resources accessible through a right-click on the table’s column header. It also provides an OSGi programming interface via Cytoscape Commands and CyREST that can be utilized for identifier mapping in scripts and other Cytoscape apps, and supports integrated Swagger documentation.

Keywords
Cytoscape, ID Mapping, Identifiers, BridgeDb

This article is included in the Cytoscape Apps gateway.

Corresponding author: Alexander R. Pico (apico@gladstone.ucsf.edu)

Author roles: Treister A: Methodology, Software, Writing – Original Draft Preparation; Pico AR: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Treister A and Pico AR. Identifier Mapping in Cytoscape: idmapper [version 1; referees: 1 approved, 1 approved with reservations] F1000Research 2018, 7:725 (doi: 10.12688/f1000research.14807.1)

Copyright: © 2018 Treister A and Pico AR. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: We would like to acknowledge funding from National Institute of General Medical Sciences [P41GM103504 (ARP, AT)] The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

Cytoscape is an integrated network visualization tool and analysis platform. Within its common workflows, identifier mapping remains a challenge when working with biological data from different sources. This problem has been addressed by the BridgeDB project, which created clients and services to translate between various identifiers. The original BridgeDb app for Cytoscape was written to provide an exhaustive set of functions to match the full capabilities of BridgeDb. Though this provided the needed functionality, its basic usage was unnecessarily complex. The idmapper app is a useful alternative, providing a subset of critical features with a simplified interface bundled into Cytoscape. Now, without any installation or configuration, Cytoscape users can right-click on a table header to map that column’s data to a different namespace (Figure 1). Although, the breadth of coverage is smaller than the full-featured BridgeDb app, it still covers over a dozen identifier data sources, including Ensembl, EntrezGene, HGNC, KEGG, Uniprot and various species-specific sources. Because idmapper supports Cytoscape’s new CyREST interface, identifier mapping can be included in scripted workflows, and driven from R or python programs.

Implementation

Inferring the data source

From within Cytoscape, a user initiates an ID mapping operation by right-clicking on the header of a column containing identifiers in the Table Panel. In the most common cases the type of identifier can be guessed by idmapper based on the its format. Table 1 shows the supported data sources and example identifier formats. The app looks at the first ten entries and choose the source from the option that matches corresponding regular expressions. This number of identifiers iteratively sampled is set by a static variable called N_Iterations. The algorithm for inferring the data source is implemented in IdGuess.java.

Cytoscape tasks

There are two different tasks supported by the idmapper app. ColumnMappingTask is activated by the right-click mouse event on a table header. It infers the current table and column from the information that comes from the mouse event. In order to support automation, we added MapColumnCommandTask as an analog that is exposed specifically for Commands and CyREST access. These tasks eventually result in the same algorithms being invoked.

Use cases

Cytoscape graphical user interface (GUI)

The idmapper app provides the same basic functionality of the BridgeDb app with less fuss. Users do not have to install it, launch it, make configuration decisions or think about which database they are accessing. The app comes bundled with every Cytoscape release. As such it usage in Cytoscape via the interactive GUI (graphical user interface) is documented in the Cytoscape manual, http://manual.cytoscape.org/en/stable/Node_and_Edge_Column_Data.html#mapping-identifiers.

To map an identifier from one source to another, right click on the column header of your identifier. Select the option to Map Column to bring up the idmapper dialog (Figure 1).

The idmapper dialog presents a few choices the user can override before performing ID mapping. The default Species is determined by the previous selection made per network, providing a “smart and sticky” behavior. The

![Figure 1. Simplified dialog for ID Mapping. Four options are presented to the user when accessing idmapper from within the Cytoscape GUI, each with common default or inferred values to reduce the number of steps required of the user.](image-url)
available choices for the identifier data sources are determined by the species. The **Map from** data source is automatically selected based on an inspection of the first ten identifiers found in the column clicked on by the user. This can easily be overridden by the pull down menu. The **To** data source must be selected by the user; Ensembl is presented by default. Finally, the **Force single** checkbox offers to simplify the results of ID mapping by ignoring one-to-many cases and only keeping the first result. If the option is off, a list of results will appear in the column. This can easily be overridden by clicking the toggled checkbox.

### Cytoscape command line interface

The command interface does not use the same tasks as the GUI. In the GUI use case, the app knows the current context of where the command was activated, i.e., the network, table and column. This information must explicitly be provided as parameters to the command interface to perform the same operation. Thus, in addition to species, mapFrom, mapTo and forceSingle, the command line operation of idmapper also requires networkName, table and columnName (see next section for more details).

### Cytoscape automation

In the scripting environment, idmapper provides all of its functionality in a single call (**Figure 2**). This means that identifier mapping can be incorporated into Cytoscape automation workflows with a single additional command.

The **map column** function takes the following parameters:

- **columnName** (string): Specifies the column name where the source identifiers are located
- **forceSingle** (string, optional): When multiple identifiers can be mapped from a single term, this forces a singular result
- **mapFrom** (string): Specifies the data source describing the existing identifiers
- **mapTo** (string): Specifies the data source identifiers to be returned as a result in a new column
- **networkName** (string, optional): Which network is used in the mapping.

• **table** (string, optional): Which table is used as the source of the identifiers, e.g., “node” for the default node table

With Cytoscape running, the **map column** function can be called from any scripting environment or programming language that supports REST calls. In the case of R and Python scripts, there are dedicated packages to make this even easier. The RCy3 package wraps this command in an R function called `mapTableColumn` to conform to other table functions ([https://www.bioconductor.org/packages/release/bioc/html/RCy3.html](https://www.bioconductor.org/packages/release/bioc/html/RCy3.html)). The `py2cytoscape` library similarly provides this command as a python function, `cyclient.idmapper.map_column` ([https://github.com/cytoscape/py2cytoscape](https://github.com/cytoscape/py2cytoscape)).


**Case 1: Species-specific considerations**

The Yeast Perturbation sample network provided with Cytoscape can be loaded from the Starter Panel and provides gene identifiers of the form “YDL194W”. These are actually Ensembl-supported identifiers for Yeast, distinct from the typical “ENSXXXG00000123456” form as presented in Table 1. This presents a special case that users will need to be aware of when selecting species and source database or `mapFrom` in the GUI. In terms of automation, you could generate a new column of Entrez Gene IDs in this network with these calls:

```r
(RCy3): mapTableColumn(column='name', species='Yeast',
                       map.from='Ensembl', map.to='Entrez_Gene')
```

```python
(py2cytoscape): cyclient.idmapper.map_column(source_column='name', species='Yeast',
                                             source_selection='Ensembl', target_selection='Entrez_Gene')
```

**Case 2: From proteins to genes**

When working with protein interaction networks, for example those from the STRING database (see [https://apps.cytoscape.org/apps/stringapp](https://apps.cytoscape.org/apps/stringapp)), you may want to translate to gene identifiers. The idmapper app supports this case as well, but one should be aware of the assumptions involved when making this translation. Since most genes encode for many proteins, you may have many-to-one mappings in your results. For all human networks imported from STRING using the `StringApp`, the following commands will perform an ID mapping from Uniprot-TrEMBL (proteins) to Ensembl (genes):

```r
(RCy3): mapTableColumn(column='canonical_name', species='Human',
                       map.from='Uniprot-TrEMBL', map.to='Ensembl')
```

```python
(py2cytoscape): cyclient.idmapper.map_column(source_column='canonical_name',
                                             species='Human', source_selection='Uniprot-TrEMBL',
                                             target_selection='Ensembl')
```

**Limitations**

The idmapper app provides easy access to a critical subset of ID mapping functionality originally covered by the BridgeDb app. When users run into the limitations of idmapper, they still have the option of installing and using the full-featured BridgeDb app from [https://apps.cytoscape.org/apps/bridgedb](https://apps.cytoscape.org/apps/bridgedb). Examples of limitations include support for additional species or data sources. The BridgeDb app includes more of both as well as means to access custom data sources.
Data and software availability

2. Latest source code: https://github.com/cytoscape/idmapper
3. Archived source code as at the time of publication: https://doi.org/10.5281/zenodo.1246814
4. License: Apache License, Version 2.0

Author contributions
AT and ARP participated in the design of the described software. AT implemented the software. AT and ARP contributed to the writing of this article.

Competing interests
No competing interests were disclosed.

Grant information
We would like to acknowledge funding from National Institute of General Medical Sciences [P41GM103504 (ARP, AT)].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments
Jianjiong Gao and Chao Zhang for their work on the original BridgeDb app (https://f1000research.com/articles/3-148/v1). Nuno Nunes for his work on the BridgeDb web service.

References

Open Peer Review

Current Referee Status: ? ✓

Version 1

Referee Report 03 July 2018
doi:10.5256/f1000research.16116.r34906

Augustin Luna
cBio Center, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

This Cytoscape app provides functionality that is widely useful for Cytoscape users for converting network identifiers to different databases. Some points not completely clear:

1. The list of regular expressions used for inference, do they come from identifiers.org? What happens if the inference fails? Does the app try to pick the closest matching regular expression?
2. For use of RCy3, is RCy3 a generic package to interact with any REST function in any Cytoscape package? Or did the developers of RCy3 specially include the mapTableColumn function to access the idmapper app?

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
Augustin, thanks for the review.

1. The regular expressions come from BridgeDb (which in turn gets them from identifiers.org). If there isn't a match to the regular expressions (or if more than one system is matched), then it just picks the first option in the list. A few of the system types aren't well specified. It's a simple matter to override this in the UI. We will add a sentence or two in the next version of the paper in response to all reviewers.

2. Right. RCy3 supports both a generic function call and a specific mapTableColumn function call. Since this is a "core" app, the RCy3 package supports custom convenience functions to make this operation easier to use and better documented. We will add a more detailed description and contrasting example in the next version of the paper.

Competing Interests: No competing interests were disclosed.
Later on in the case 1 you mention an exception to the basic regular expression behaviour for yeast. Are there more exceptions that the app handles?

It is also unclear why the "Code" column is required.

Might be nice to separate the data sources that can handle any species and those that are species specific.

In the use cases section "The default Species is determined by the previous selection made per network, providing a "smart and sticky" behavior." It is unclear but the previous selection was?

In the use cases section in the specific cases two example use cases are presented, species specific and protein to gene conversions. It would be helpful to list these and other common use cases at the start of the use case section as well. One of the most common use cases being going from non-descriptive identifiers (like entrez gene ids, and ensembl ids) to something more understandable such as species specific IDs (HGNC or MGD, Is it possible to map to proper gene symbols?)

In the use case section it is stated that if the ID maps to multiple identifiers there is an option, "Force single", that when selected the app selects the first result. How are the returned IDs sorted? Is the first match the "best" match, alphabetical, random?

In the Cytoscape automation section in the parameters section for the species option all the available mappings are listed but for the mapFrom and mapTo no options are listed. (if the recognized data source name is add to Table 1 you can just reference the table here or if the first column of Table 1 are the recognized names it would be good to reference it here). Also, the parameters listed for Cytoscape automation section are very different from the parameters used in the use cases which can be very confusing. Maybe adding an example using the RCy3 commandsGet option under RCy3 and py2cytoscape examples just showing how the user can use all the parameters as specified using the command directly.

Minor comments/questions:

- In the introduction "Uniprot and various specied-specific sources" should be "Uniprot and various species-specific sources"
- In the implementation section "The app looks at the first ten entries and choose the source" would be better as "The app looks at the first ten entries and chooses the source"
- In the implementation section - "This number of identifiers iteratively sampled is set by a static variable called N_Iterations. The algorithm for inferring the data source is implemented in ldGuess.java." - This is a little confusing why this is needed. Is this a parameter the user can control or tweak?
- In Use cases section - "As such it usage in Cytoscape via" should be "As such its usage in Cytoscape via"
- In use case 2 "you may want to translate to gene identifiers" might be better as "you may want to translate protein identifiers (for example: Uniprot-TrEMBL) to gene identifiers"
- Can idmapper convert a list column? (in the example use case where the network is an enrichment map and each node contains a set of genes as opposed to each node being a gene)
- What is the resulting column name? What if a column with that name already exists?

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes
Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

**Author Response 26 Jun 2018**

**Alexander Pico,** The Gladstone Institutes, UCSF, USA

Ruth, thank you so much for your thorough review. These clarifications, fixes and additions have greatly improved the article. Version 2 should be released soon, addressing all the issues you raised. We decided not to include the regular expressions in Table 1, however. They are messy and are what you’d expect from the example identifiers provided, which we feel do a better job of communicating what to expect from each data source.

We hope you’ll have a chance to look over version 2 and find that it meets your expectations.

**Competing Interests:** No competing interests were disclosed.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage