An R package with BridgeDb for identifier mapping

Egon Willighagen

August 30, 2017

1 Introduction

BridgeDb `https://github.com/bridgedb/BridgeDb` is a combination of an application programming interface (API), library, and set of data files for mapping identifiers for identical objects [2]. Because BridgeDb is use by projects in bioinformatics, like WikiPathways and PathVisio [1, 3], identifier mapping databases are available for gene products and metabolites.

2 Concepts

BridgeDb has a few core concepts which are explained in this section. Much of the API requires one to be familiar with these concepts, though some are not always applicable. The first concept is an example of that: organisms, which do not apply to metabolites.

2.1 Organisms

However, for genes the organism is important: the same gene has different identifiers in different organisms. BridgeDb identifies organisms by their latin name and with a two character code. Because identifier mapping files provided by PathVisio have names with these short codes, it can be useful to have a conversion method:

```
> code = getOrganismCode("Rattus norvegicus")
> code

[1] "Rn"
```

2.2 Data Sources

Identifiers have a context and this context is often a database. For example, metabolite identfiers can be provided by the Human Metabolome Database (HMDB), ChemSpider, PubChem, and ChEBI. Similarly, gene product identifiers can be provided by databases like Ensemble. Such a database providing identifiers is in BridgeDb called a data source.

Importantly, each such data source is identified by a human readable long name and by a short system code. This package has methods to interconvert one into the other:

```
> fullName <- getFullName("Ce")
> fullName

[1] "ChEBI"

> code <- getSystemCode("ChEBI")
> code

[1] "Ce"
```

2.3 Identifier Patterns

Another useful aspect of BridgeDb is that it knows about the patterns of identifiers. If this pattern is unique enough, it can be used used to automatically find the data sources that match a particular identifier. For example:

```
> getMatchingSources("HMDB00555")

 [1] "Ensembl Plants"                    "LipidBank"
 [3] "KEGG Pathway"                      "Wikipedia"
 [5] "NCI Pathway Interaction Database" "SWISS-MODEL"
 [7] "NCBI Protein"                      "EMBL"
 [9] "HGNC"                              "SUPFAM"
[11] "HMDB"

> getMatchingSources("ENSG00000100030")

 [1] "Ensembl Plants"                    "Ensembl"
 [3] "LipidBank"                         "Ensembl Human"
 [5] "Wikipedia"                         "NCI Pathway Interaction Database"
 [7] "SWISS-MODEL"                       "NCBI Protein"
 [9] "EMBL"                              "HGNC"
[11] "SUPFAM"
```

3 Identifier Mapping Databases

The BridgeDb package primarily provides the software framework, and not identifier mapping data. Identifier Mapping databases can be downloaded from various websites. The package knows about the download location provided by PathVisio, and we can query for all gene product identifier mapping databases:

```
> getBridgeNames()

 [1] "Ag_Derby_Ensembl_Metazoa_32.bridge" "An_Derby_Ensembl_Fungi_32.bridge"
 [3] "At_Derby_Ensembl_Plant_32.bridge"   "Bs_Derby_Ensembl_85.bridge"
 [5] "Bt_Derby_Ensembl_85.bridge"         "Ce_Derby_Ensembl_85.bridge"
 [7] "Cf_Derby_Ensembl_85.bridge"         "Ci_Derby_Ensembl_85.bridge"
 [9] "Dm_Derby_Ensembl_85.bridge"         "Dr_Derby_Ensembl_85.bridge"
[11] "Ec_Derby_Ensembl_85.bridge"         "Gg_Derby_Ensembl_85.bridge"
[13] "Gm_Derby_Ensembl_Plant_32.bridge"   "Gz_Derby_Ensembl_Fungi_32.bridge"
[15] "Hs_Derby_Ensembl_85.bridge"         "Hv_Derby_Ensembl_Plant_32.bridge"
[17] "Ml_Derby_Ensembl_85.bridge"         "Mm_Derby_Ensembl_85.bridge"
[19] "Mx_Derby_Ensembl_85.bridge"         "Oa_Derby_Ensembl_85.bridge"
[21] "Oi_Derby_Ensembl_Plant_32.bridge"   "Oj_Derby_Ensembl_Plant_32.bridge"
[23] "Pi_Derby_Ensembl_Plant_32.bridge"   "Pt_Derby_Ensembl_85.bridge"
[25] "Qc_Derby_Ensembl_85.bridge"         "Rn_Derby_Ensembl_85.bridge"
[27] "Sc_Derby_Ensembl_85.bridge"         "Sl_Derby_Ensembl_Plant_32.bridge"
[29] "Ss_Derby_Ensembl_85.bridge"         "Vv_Derby_Ensembl_Plant_32.bridge"
[31] "Xt_Derby_Ensembl_85.bridge"         "Zm_Derby_Ensembl_Plant_32.bridge"
```

3.1 Downloading

The package provides a convenience method to download such identifier mapping databases. For example, we can save the identifier mapping database for rat to the current folder with:

```
> dbLocation <- getDatabase("Rattus norvegicus",location=getwd())
```

The dbLocation variable then contains the location of the identifier mapping file that was downloaded.

3.2 Loading Databases

Once you have downloaded an identifier mapping database, either manually or via the getDatabase() method, you need to load the database for the identifier mappings to become available.

```
> mapper <- loadDatabase(dbLocation)
```

4 Mapping Identifiers

With a loaded database, identifiers can be mapped. The mapping method uses system codes. So, to map the human Entrez Gene identifier (system code: L) 196410 to Affy identifiers (system code: X) we use:

```
> location <- getDatabase("Homo sapiens")
> mapper <- loadDatabase(location)
> map(mapper, "L", "196410", "X")
```

Mind you, this returns more than one identifier, as BridgeDb is generally a one to many mapping database.

References

[1] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo. WikiPathways: Pathway editing for the people. *PLoS Biol*, 6(7):e184+, July 2008.

[2] M. van Iersel, A. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B. Conklin, and C. Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11(1):5+, 2010.

[3] M. P. van Iersel, T. Kelder, A. R. Pico, K. Hanspers, S. Coort, B. R. Conklin, and C. Evelo. Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*, 9, SEP 25 2008.