SOFTWARE TOOL ARTICLE

BED: a Biological Entity Dictionary based on a graph data model [version 1; peer review: 2 approved with reservations]

Patrice Godard ^{1,2}, Jonathan van Eyll²

¹Clarivate Analytics, Carlsbad, CA, 92008, USA

²UCB Pharma, Braine-l'Alleud, 1420, Belgium

V1 First published: 15 Feb 2018, 7:195
<https://doi.org/10.12688/f1000research.13925.1>
 Second version: 16 May 2018, 7:195
<https://doi.org/10.12688/f1000research.13925.2>
 Latest published: 19 Jul 2018, 7:195
<https://doi.org/10.12688/f1000research.13925.3>

Abstract

The understanding of molecular processes involved in a specific biological system can be significantly improved by combining and comparing different data set and knowledge resources. However these information sources often use different identification systems and an identifier conversion step is required before any integration effort. Mapping between identifiers is often provided by the reference information resources and several tools have been implemented to simplify their use. However these tools cannot be easily customized and optimized for any specific use. Also the information provided by different resources is not combined to increase the efficiency of the mapping process and deprecated identifiers from former version of databases are not taken into account. Finally finding automatically the most relevant path to map identifiers from one scope to the other is often not trivial. The Biological Entity Dictionary (BED) addresses these challenges by relying on a graph data model describing possible relationships between entities and their identifiers. This model has been implemented using Neo4j and an R package provides functions to query the graph but also to create and feed a custom instance of the database.

Keywords

genomics, transcriptomics, proteomics, RNA-seq, microarray, database, identifiers

Open Peer Review

Approval Status

	1	2
version 3		
(revision)		view
19 Jul 2018		
version 2		
(revision)	view	view
16 May 2018		
version 1		
15 Feb 2018	view	view

1. **Denise Slenter** , Maastricht University, Maastricht, The Netherlands

Martina M. Summer-Kutmon , Maastricht University, Maastricht, The Netherlands

2. **T. Ian Simpson** , University of Edinburgh, Edinburgh, UK

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the RPackage gateway.

Corresponding author: Patrice Godard (patrice.godard@ucb.com)

Author roles: **Godard P:** Conceptualization, Methodology, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **van Eyll J:** Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was entirely supported by UCB Pharma. The authors declared that no grants were involved in supporting this work.

Copyright: © 2018 Godard P and van Eyll J. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Godard P and van Eyll J. **BED: a Biological Entity Dictionary based on a graph data model [version 1; peer review: 2 approved with reservations]** F1000Research 2018, 7:195 <https://doi.org/10.12688/f1000research.13925.1>

First published: 15 Feb 2018, 7:195 <https://doi.org/10.12688/f1000research.13925.1>

Introduction

Since the advent of genome sequencing projects, many technologies have been developed to get access to different molecular information on a large scale and with high throughput. DNA micro-arrays are probably the archetype of such technology because of their historical impact on gathering data related to nucleic acids: genomic DNA and RNA. They triggered the emergence of “omics” fields of research such as genomics, epigenomics or transcriptomics. Lately massive parallel sequencing further increased the throughput of data generation related to nucleic acids by several orders of magnitude. In a different way, mass spectrometry-related technologies allow the identification and the quantification of many kinds of molecular entities such as metabolites and proteins. Many information systems have been developed to manage the exploding amount of data and knowledge related to biological molecular entities. These resources manage different aspects of the data. For example some are genome or proteome centered, whereas others are focused on molecular interactions and pathways. Thus all these resources rely on different identifier systems to organize the concepts of interest. The value of all the experimental data and all the knowledge collected in public or private resources is very high as such but is also often synergistically leveraged by their cross comparison in a dedicated manner. Indeed many datasets can be relevant when addressing the understanding of a specific biological system, a phenotypic trait or a disease for example. These datasets can focus on different biological entities such as transcripts or proteins in different tissues, conditions or organisms. Comparing all these data and integrating them with available knowledge requires the ability to map the identifiers on which each resource relies.

To achieve this task public and proprietary information systems provide mapping tables between their own identifiers and those from other resources. Furthermore many tools have been developed to facilitate the access to this information. Ensembl BioMarts (Kinsella *et al.*, 2011), mygene (Wu *et al.*, 2013), and g:Profiler (Reimand *et al.*, 2016a) are popular examples among many others. However, as pointed out by van Iersel *et al.* (2010), these tools are generally dedicated to a particular domain and not necessarily relevant or complete for all research projects, and keeping them up-to-date can also be an issue. Recognizing these challenges van Iersel *et al.* (2010) proposed the BridgeDb framework providing to bioinformatics developers a standard interface between tools and mapping services and also allowing the easy integration of custom data by a transitivity mechanism.

Here we present BED: a biological entity dictionary. BED has been developed to address three main challenges. The first one is related to the completeness of identifier mappings. Indeed direct mapping information provided by the different systems are not always complete and can be enriched by mappings provided by other resources. More interestingly direct mappings not identified by any of these resources can be indirectly inferred by using mappings to a third reference. For example, many human Ensembl gene identifiers are not directly mapped to any Entrez gene identifiers but such mapping can be inferred using respective mappings to HGNC identifiers. The second challenge is related to the mapping of deprecated identifiers. Indeed entity identifiers can change from one resource release to another. The identifier history is provided by some resources, such as Ensembl or the NCBI, but it is generally not used by mapping tools. The third challenge is related to the automation of the mapping process according to the relationships between the biological entities of interest. Indeed mapping between gene and protein identifier scopes should not be done the same way than two scopes of gene identifiers. Also converting identifiers from different organism should be possible using gene ortholog information.

To meet these challenges we designed a graph data model describing possible relationships between different biological entities and their identifiers. This data model has been implemented with the Neo4j® graph database (Neo4j inc, 2017) and conversion rules have been defined and coded in an R (R Core Team, 2017) package. We provide an instance of the BED database focused on human, mouse and rat organism but many functions are available to construct other instances tailored to other needs.

Methods

Data model

The BED (Biological Entity Dictionary) system relies on a data model inspired by the central dogma of molecular biology (Crick, 1970) and describing relationships between molecular concepts usually manipulated in the frame of genomics studies (Figure 1). A biological entity identifier (*BEID*) can identify either a *Gene* (*GeneID*), a *Transcript* (*TranscriptID*), a *Peptide* (*PeptideID*) or an *Object* (*ObjectID*). *Object* entities can correspond to complex concepts coded by any number of genes (i.e. a protein complex or a molecular function). *BEID* are extracted from public or private databases (*BEDB*). *BEDB* can provide an *Attribute* related to each *BEID*. For example it can be the sequencing region provided by the Ensembl database (Zerbino *et al.*, 2018) or the identifier status provided by Uniprot (The UniProt Consortium, 2017). *BEID* can have one or several associated names (*BENames*) and symbols (*BESymbol*). *GeneID* can have one or several homologs in other organisms belonging to the same *GeneIDFamily*. Many genomics platforms, such as micro-array, allow the identification of biological entity by using probes identified

Figure 1. The BED graph data model. The model is shown as an Entity/Relationship (ER) diagram: entities correspond to graph nodes and relationships to graph edges. "ID" and "idx" indicate if the corresponding entity property is unique or indexed respectively. Some redundancies occur in this data model. Indeed some "value" properties are duplicated in upper case ("value_up") in order to improve the performance of case-insensitive searches. Also the database of a BEID node is provided as a property to ensure uniqueness of the couples of "database" and "value" properties. The same approach has been applied for the "platform" property of ProbelID nodes.

by *ProbeID*. In general, *BEID* can be targeted by several probes belonging to a *Platform* which is focused on one, and only one, type of entity (BEType) among those described above: *Gene*, *Transcript*, *Peptide* or *Object*. A BEType can have several BEType products but can be the product of at most one BEType. This constraint allows the unambiguous identification of the most relevant path to convert identifiers from one scope to another and is fulfilled by the current data model: peptides are only produced from transcripts, which are only produced from genes, which can also code for objects.

BEID identifying the same biological entity are related through three different kinds of relationship according to the information available in the source databases, and to the decision made by the database administrator about how to use them. Two *BEID* which *corresponds_to* each other both *identify* the same biological entity. A *BEID* which *is_associated_to* or which *is_replaced_by* another *BEID* does not directly identify any biological entity: the link is always indirect through one or several other *BEID*. Therefore, by design a *BEID* which *is_associated_to* or which *is_replaced_by* another *BEID* can be related to several different biological entities. It is not the case for other *BEID* which identify one and only one biological entity. This set of possible relationship allows the indirect mapping of different identifiers not necessarily provided by any integrated resource.

In order to efficiently leverage an indirect path through these different relationships the data model has been implemented in a Neo4j® graph database (Neo4j inc, 2017).

Feeding the database

Two R (R Core Team, 2017) packages have been developed to feed and query the database. The first one, neo2R, provides low level functions to interact with Neo4j®. The second R package, BED, provides functions to feed and query the BED Neo4j® graph database according to the data model described above.

Many functions are provided within the package to build a tailored BED database instance. These functions are not exported in order not to mislead the user when querying the database (which is the expected most frequent usage of the system). An R markdown document showing how to build a BED database instance for human, mouse and rat organisms is provided within the package. It can be adapted to other organisms or needs.

Briefly these functions can be divided according to three main levels:

- The lowest level function is the `bedImport` function which loads a table in the Neo4j® database according to a Cypher® query.
- Functions of the second level allow loading identifiers and relationships tables ensuring the integrity of the data model.
- Highest level functions are helpers for loading information provided by some public resources in different specific formats.

Querying the database

The *BED* R package provides several functions to retrieve identifiers from different resources, and also to convert identifiers from one reference to another. These functions generate and call Cypher® queries on the Neo4j® database. Converting thousands of identifiers can take some time (generally a few seconds). Also such conversions are often recurrent and redundant. In order to improve the performance for such recurrent and redundant queries, a cache system has been implemented. The first time, the query is run on Neo4j® for all the relevant ID related to user input and the result is saved in a local file. Next time similar queries are requested, the system does not call Neo4j® but loads the cached results and filters it according to user input. By default the cache is flushed when the system detects inconsistencies with the BED database. It can also be manually flushed if needed.

Operation

Minimal system requirements for running BED and neo2R R packages:

- **R** ≥ 3.4
- **Operating system:** Linux, macOS, Windows
- **Memory** $\geq 4GB$ RAM

The graph database has been implemented with Neo4j® version 3 (Neo4j inc, 2017). The BED R package depends on the following packages available in the Comprehensive R Archive Network (CRAN):

- *visNetwork* (Almende *et al.*, 2017)
- *dplyr* (Wickham *et al.*, 2017)
- *htmltools* (RStudio inc, 2017)
- *DT* (Xie, 2016)
- *shiny* (Chang *et al.*, 2017)
- *miniUI* (Cheng, 2016)
- *rstudioapi* (Allaire *et al.*, 2017)

Use cases

Available database instance

An instance of the BED database (UCB-Human) has been built using the script provided in the BED R package and made available in a Docker® image (Docker inc, 2017) available here: <https://hub.docker.com/r/patzaw/bed-ucb-human/>

This instance used to exemplify the following use cases is focused on *Homo sapiens*, *Mus musculus* and *Rattus norvegicus* organisms and it has been built from the following resources:

- Ensembl (Zerbino *et al.*, 2018)
- NCBI (NCBI Resource Coordinators, 2017)
- Uniprot (The UniProt Consortium, 2017)
- biomaRt (Durinck *et al.*, 2009)
- GEOquery (Davis & Meltzer, 2007)
- Clarivate Analytics MetaBase® (Clarivate Analytics, 2017)

The numbers of biological entity (BE) identifiers (BEID) available in this BED database instance and which can be mapped to each other are shown in Table 1. In total, 3,519,181 BEID are available in this BED instance. This number includes deprecated identifiers without successor and which therefore cannot be mapped to any other identifier. All the genomics platforms included in this BED database instance are shown in Table 2. They provide mapping to BEID from 354,205 ProbeID in total.

Table 1. Numbers of BEID available in the BED UCB-Human database instance. Numbers have been split according to the BE type and the organism. Only BEID which can be mapped to each other are taken into account (e.g. excluding deprecated identifiers without successor).

BE	Organism	Database	BEID	URL
Gene	Homo sapiens	MIM_GENE	17,146	http://www.omim.org
Gene	Homo sapiens	miRBase	1,881	http://www.mirbase.org
Gene	Homo sapiens	UniGene	23,012	https://www.ncbi.nlm.nih.gov
Gene	Homo sapiens	Ens_gene	68,460	http://www.ensembl.org
Gene	Homo sapiens	HGNC	41,195	http://www.genenames.org
Gene	Homo sapiens	EntrezGene	81,761	https://www.ncbi.nlm.nih.gov
Gene	Homo sapiens	Vega_gene	19,141	http://vega.sanger.ac.uk
Gene	Homo sapiens	MetaBase_gene	23,377	https://portal.genego.com
Gene	Mus musculus	miRBase	1,193	http://www.mirbase.org
Gene	Mus musculus	UniGene	21,576	https://www.ncbi.nlm.nih.gov
Gene	Mus musculus	Ens_gene	56,954	http://www.ensembl.org
Gene	Mus musculus	MGI	78,547	http://www.informatics.jax.org
Gene	Mus musculus	EntrezGene	103,555	https://www.ncbi.nlm.nih.gov
Gene	Mus musculus	Vega_gene	45,237	http://vega.sanger.ac.uk
Gene	Mus musculus	MetaBase_gene	20,628	https://portal.genego.com
Gene	Rattus norvegicus	miRBase	495	http://www.mirbase.org
Gene	Rattus norvegicus	UniGene	12,613	https://www.ncbi.nlm.nih.gov
Gene	Rattus norvegicus	Ens_gene	34,963	http://www.ensembl.org
Gene	Rattus norvegicus	RGD	46,976	https://rgd.mcw.edu
Gene	Rattus norvegicus	EntrezGene	57,026	https://www.ncbi.nlm.nih.gov
Gene	Rattus norvegicus	Vega_gene	1,146	http://vega.sanger.ac.uk
Gene	Rattus norvegicus	MetaBase_gene	17,505	https://portal.genego.com
Transcript	Homo sapiens	Ens_transcript	228,389	http://www.ensembl.org
Transcript	Homo sapiens	Vega_transcript	37,017	http://vega.sanger.ac.uk
Transcript	Homo sapiens	RefSeq	189,384	https://www.ncbi.nlm.nih.gov

BE	Organism	Database	BEID	URL
Transcript	Mus musculus	Ens_transcript	136,967	http://www.ensembl.org
Transcript	Mus musculus	Vega_transcript	120,271	http://vega.sanger.ac.uk
Transcript	Mus musculus	RefSeq	112,390	https://www.ncbi.nlm.nih.gov
Transcript	Rattus norvegicus	Ens_transcript	42,393	http://www.ensembl.org
Transcript	Rattus norvegicus	Vega_transcript	1,271	http://vega.sanger.ac.uk
Transcript	Rattus norvegicus	RefSeq	98,431	https://www.ncbi.nlm.nih.gov
Peptide	Homo sapiens	Ens_translation	109,643	http://www.ensembl.org
Peptide	Homo sapiens	Vega_translation	36,460	http://vega.sanger.ac.uk
Peptide	Homo sapiens	RefSeq_peptide	117,465	https://www.ncbi.nlm.nih.gov
Peptide	Homo sapiens	Uniprot	232,130	http://www.uniprot.org
Peptide	Mus musculus	Ens_translation	65,406	http://www.ensembl.org
Peptide	Mus musculus	Vega_translation	57,318	http://vega.sanger.ac.uk
Peptide	Mus musculus	RefSeq_peptide	79,418	https://www.ncbi.nlm.nih.gov
Peptide	Mus musculus	Uniprot	114,825	http://www.uniprot.org
Peptide	Rattus norvegicus	Ens_translation	30,245	http://www.ensembl.org
Peptide	Rattus norvegicus	Vega_translation	1,260	http://vega.sanger.ac.uk
Peptide	Rattus norvegicus	RefSeq_peptide	68,716	https://www.ncbi.nlm.nih.gov
Peptide	Rattus norvegicus	Uniprot	40,786	http://www.uniprot.org
Object	Homo sapiens	MetaBase_object	24,748	https://portal.genego.com
Object	Homo sapiens	GO_function	4,104	http://amigo.geneontology.org
Object	Mus musculus	MetaBase_object	22,000	https://portal.genego.com
Object	Mus musculus	GO_function	4,081	http://amigo.geneontology.org
Object	Rattus norvegicus	MetaBase_object	18,648	https://portal.genego.com
Object	Rattus norvegicus	GO_function	4,001	http://amigo.geneontology.org

Table 2. Genomics platforms available in the BED UCB-Human database instance.

Name	Description	BE
GPL6101	Illumina ratRef-12 v1.0 expression beadchip	Gene
GPL6947	Illumina HumanHT-12 V3.0 expression beadchip	Gene
GPL10558	Illumina HumanHT-12 V4.0 expression beadchip	Gene
GPL1355	[Rat230_2] Affymetrix Rat Genome 230 2.0 Array	Gene
GPL1261	[Mouse430_2] Affymetrix Mouse Genome 430 2.0 Array	Gene
GPL96	[HG-U133A] Affymetrix Human Genome U133A Array	Gene
GPL13158	[HT_HG-U133_Plus_PM] Affymetrix HT HG-U133+ PM Array Plate	Gene
GPL571	[HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array	Gene
GPL570	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	Gene
GPL6480	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	Gene
GPL6885	Illumina MouseRef-8 v2.0 expression beadchip	Transcript

Exploring identifiers of biological entities

The `getBeIds` function returns all BE identifiers from a specific scope. A scope is defined by the type of BE or probe, the source of the identifiers (database or platform) and the organism. For example, the following code returns all the Ensembl identifiers of human genes.

```
beids <- getBeIds(
  be="Gene", source="Ens_gene", organism="human",
  restricted=FALSE
)
head(beids)
```

##	id	preferred	Gene	db.version	db.deprecated
## 82643	ENSG00000283891	TRUE	64781	91	FALSE
## 82642	ENSG00000207766	TRUE	64783	91	FALSE
## 82645	ENSG00000276678	TRUE	64785	91	FALSE
## 82644	ENSG00000265993	TRUE	64787	91	FALSE
## 82647	ENSG00000283793	TRUE	64789	91	FALSE
## 82646	ENSG00000283621	TRUE	64791	91	FALSE

The *id* column corresponds to the BEID from the source of interest. The column named according to the BE type (in this case *Gene*) corresponds to the internal identifiers of the related BE. This internal identifier is not a stable reference that can be used as such. Nevertheless, it is useful to identify BEID identifying the same BE. In the example above even if most of Gene BE are identified by only one Ensembl gene BEID, many of them are identified by two or more (5,809 / 59,515 = 10%); 277 BE are even identified by more than 10 Ensembl BEID (Figure 2.a). In this case, most of these redundancies come from deprecated ID from former versions of the Ensembl database (version in use here: 91) and can be excluded by setting the *restricted* parameter to *TRUE* when calling the `getBeIds` function (Figure 2.b). However many BE are still identified by two or more current Ensembl BEID (2,715 / 59,515 = 5%). This result comes from the way the BED database is constructed: When two identifiers from the same resource correspond to the same identifier in another resource (*correspond_to* relationship in the data model), all these BEID are considered to identify the same BE.

A complex example of such mapping is shown in Figure 3 mapping all the BEID of the human TAS2R8 gene which codes for a protein of the family of candidate taste receptors. There are three identifiers corresponding to this gene

Figure 2. Barplots showing the number of gene BE (log scale) identified by one or more Ensembl gene BEID. a) All Ensembl gene ID. b) Current Ensembl gene ID (version 91).

Figure 3. BED relationships between all the different identifiers of the human TAS2R8 gene recorded in the database. BEID are shown as circle and gene symbol in the rounded box. The color legend is shown to the left of the figure. BEID surrounded in bold correspond to *preferred* identifiers. Solid arrows represent *correspond_to* and *is_known_as* relationships. Dotted arrows represent *is_replaced_by* and *is_associated_to* relationships. This graph has been drawn with the `exploreBe` function.

symbol in Ensembl. All these three identifiers correspond to the same Entrez gene and the same HGNC identifiers. All these BEID are thus considered to identify the same gene. It turns out that the three Ensembl BEID correspond to the same gene mapped on different sequence version of the chromosome 12: the canonical (ENSG00000121314), CHR_HSCHR12_2_CTG2 (ENSG00000272712) and CHR_HSCHR12_3_CTG2 (ENSG00000277316). This information provided by Ensembl is encoded in the `seq_region` attribute for each Ensembl BEID (see data model) and is used to define *preferred* BEID which are mapped on canonical version of chromosome sequences. The ENSG00000272712 identifier shows also a complex history in former Ensembl versions.

Converting identifiers

The main goal of BED is to convert identifiers from one scope to another easily, rapidly and with high completeness. It has been thought in order to allow recurring comparisons to each other of many lists of biological entities from various origins.

The function `guessIdOrigin` can be used to guess the scope of any list of identifiers. A simple example regarding the conversion of human Ensembl gene to human Entrez gene identifiers is shown below and discussed hereafter. By setting the `restricted` parameter to `TRUE` the converted BEID are restricted to current - non-deprecated - version of Entrez gene identifiers. Nevertheless all the input BEID are taken into account, current and deprecated ones.

```
bedConv <- convBeIds (
  ids=beids$id, from="Gene", from.source="Ens_gene", from.org="human",
  to.source="EntrezGene", restricted=TRUE
)
```

Among all the 68,460 human Ensembl gene identifiers available in the database, 21,718 (32%) were not converted to any human Entrez gene identifier: 21,073 (33%) of the 64,661 non-deprecated and 645 (17%) of the 3,799 deprecated identifiers.

Three other tools were used on January 04, 2018 to perform the same conversion task: biomaRt (Durinck *et al.*, 2009; Kinsella *et al.*, 2011), mygene (Mark *et al.*, 2014; Wu *et al.*, 2013), and gProfileR (Reimand *et al.*, 2016a; Reimand, 2016b). At that time, biomaRt and mygene were based on the Ensembl 91 release whereas gProfileR was based on release 90.

The numbers of human Ensembl gene identifiers successfully converted by each method are compared in Figure 4. Five identifiers were only converted by gProfileR. They were provided by former versions of Ensembl or NCBI but are now deprecated in the current releases of these two resources. All the other gene identifiers converted by the different methods were also converted by BED. However, BED was able to map at least 17,912 more identifiers than all the other tools (Figure 4.a). A few of these mappings (3,154) are explained by the fact that BED is the only tool mapping deprecated identifiers to current versions. Nevertheless, even when focusing on the mapping of current versions of

Ensembl identifiers BED was able to map 14,758 more identifiers than all the other tools (Figure 4.b). A few of these mappings (627) are directly provided by the NCBI. But most of them (14,131) are inferred from a mapping of the Ensembl and Entrez gene identifiers to the same HGNC (Gray *et al.*, 2015) identifier.

A rough approximation of running times of the different methods is provided in Table 3. The aim of this table is to show that BED, as a dedicated and locally available tool, is a very efficient option to convert large lists of identifiers on the fly and recurrently. The aim of BED is to improve the efficiency of identifier conversion in a well defined context (organism, information resources of interest. . .) and not to replace biomaRt, mygene, gProfileR or other tools which provide many more features for many organisms and which should not be narrowed to this task for a complete comparison.

The BED `convBEIDs` function can be used to convert identifiers from any available scope to any other one. It automatically find the most relevant path according to the considered biological entities. It allows elaborate mapping such as the conversion between probe identifiers from a platform focused on mouse transcripts into human protein identifiers. Because such mappings can be intricate, BED also provides a function to show the shortest relevant path between two different identifiers (Figure 5).

Additional features

Some additional use cases and examples are provided in the BED R package vignette. Several functions are available for annotating BEID with symbols and names, again taking advantage of information related to connected identifiers. Other functions are also provided to seek relevant identifiers of a specific biological entity. These functions are used by a shiny (Chang *et al.*, 2017) gadget (Figure 6) providing an interactive dictionary of BEID which is also made available as an Rstudio add-in (Allaire *et al.*, 2017; Cheng, 2016).

Figure 4. Venn diagrams showing the number of human Ensembl gene identifiers mapped to at least one human Entrez gene identifier by the different tested tools when focusing (a) on all 68,460 or (b) on current 64,661 BEID (Ensembl 91 release).

Table 3. Rough approximation of running time of different methods to convert human Ensembl gene identifiers in human Entrez gene identifiers.

Method	Running time
BED (Not cached)	~9.9 secs
BED (Cached)	~2.5 secs
biomaRt	~40 secs
mygene	~3.9 mins
gProfileR	~1.2 mins

Figure 5. BED conversion shortest path between the ILMN_1220595 probe identifier targeting a transcript of the mouse Il17a gene and the Uniprot Q16552 identifier of the human IL17 protein. The legend is shown to the left of the figure. The red arrow represents the *is_homolog_of* relationship. This graph has been drawn with the `exploreConvPath` function.

Found	ID	Symbol	Name	Preferred	DB version
IL6 (Gene canonical Symbol in Homo sapiens)	0426291	IL6-207	interleukin 6	True	91
IL6 (Gene canonical Symbol in Homo sapiens)	ENST00000406575	IL6-205	interleukin 6	True	91
IL6 (Gene canonical Symbol in Homo sapiens)	ENST00000401630	IL6-202	interleukin 6	True	91
IL6 (Gene canonical Symbol in Homo sapiens)	ENST00000485300	IL6-209	interleukin 6	True	91
IL6 (Gene canonical Symbol in Homo sapiens)	ENST00000464710	IL6-208	interleukin 6	True	91
IL6 (Gene canonical Symbol in Homo sapiens)	ENST00000258743	IL6-201	interleukin 6	True	91

Showing 1 to 7 of 9 entries

☐ Cross species search (time consuming and not relevant for complex objects such as GO functions)

☒ Show gene annotation (not relevant for complex objects such as GO functions)

Figure 6. findBE Shiny gadget to seek relevant identifiers of a specific biological entity. In this example the user is looking after human Ensembl transcript identifiers corresponding to "il6".

Conclusions

BED is a system dedicated to the mapping between identifiers of molecular biological entities. It relies on a graph data model implemented with Neo4j® and on rules coded in an R package. BED leverages mapping information provided by different resources in order to increase the mapping efficiency between each of them. It also allows the mapping of deprecated identifiers. Rules are used to automatically convert identifiers from one scope to another using the most appropriate path.

The intent of BED is to be tailored to specific needs, and beside functions for querying the system, the BED R package provides functions to build custom instances of the database. Database instances can be locally installed or shared across a community. This design combined with a cache system makes BED efficient for converting large lists of identifiers from and to a large variety of scopes.

Because of our research field we provide an instance focused on human, mouse and rat organisms. This database instance can be directly used in relevant projects but it can also be enriched depending on user or community needs.

Software availability

Latest source code is available at:

<https://github.com/patzaw/BED>

<https://github.com/patzaw/neo2R>

Archived source code as at time of publication:

<https://zenodo.org/badge/latestdoi/119707445> (Godard, 2018a)

<https://zenodo.org/badge/latestdoi/119698430> (Godard, 2018b)

Software is available to use under a GPL-3 license

Competing interests

No competing interests were disclosed.

Grant information

This work was entirely supported by UCB Pharma. The authors declared that no grants were involved in supporting this work.

Acknowledgments

We are grateful to Frédéric Vanclef, Malte Lucken, Liesbeth François, Matthew Page, Massimo de Francesco, and Marina Bessarabova for fruitful discussions and constructive criticisms.

References

- Allaire JJ, Wickham H, Ushey K, *et al.*: **rstudioapi: Safely Access the RStudio API**. 2017.
[Reference Source](#)
- Almende BV, Thieurmél B, Robert T: **visNetwork: Network Visualization using 'vis.js' Library**. 2017.
[Reference Source](#)
- Chang W, Cheng J, Allaire JJ, *et al.*: **shiny: Web Application Framework for R**. 2017.
[Reference Source](#)
- Cheng J: **miniUI: Shiny UI Widgets for Small Screens**. 2016.
[Reference Source](#)
- Clarivate Analytics: **MetaCore delivers high-quality biological systems content in context**. 2017.
[Reference Source](#)
- CRAN: **The Comprehensive R Archive Network**.
[Reference Source](#)
- Crick F: **Central dogma of molecular biology**. *Nature*. 1970; 227(5258): 561–563.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Davis S, Meltzer PS: **GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor**. *Bioinformatics*. 2007; 23(14): 1846–1847.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Docker inc: **Docker Community Edition**. 2017.
[Reference Source](#)
- Durinck S, Spellman PT, Birney E, *et al.*: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt**. *Nat Protoc*. 2009; 4(8): 1184–1191.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Godard P: **patzaw/BED: Publication release (Version v1.0.0)**. *Zenodo*. 2018a.
[Data Source](#)
- Godard P: **patzaw/neo2R: Publication release (Version v1.0.0)**. *Zenodo*. 2018b.
[Data Source](#)
- Gray KA, Yates B, Seal RL, *et al.*: **Genenames.org: the HGNC resources in 2015**. *Nucleic Acids Res*. 2015; 43(Database issue): D1079–1085.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kinsella RJ, Kähäri A, Haider S, *et al.*: **Ensembl BioMart: a hub for data retrieval across taxonomic space**. *Database (Oxford)*. 2011; 2011: bar030.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mark A, Thompson R, Afrasiabi C, *et al.*: **mygene: Access MyGene. Info. services**. 2014.
[Publisher Full Text](#)
- NCBI Resource Coordinators: **Database Resources of the National Center for Biotechnology Information**. *Nucleic Acids Res*. 2017; 45(D1): D12–D17.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Neo4j inc: **Neo4j Community Edition**. 2017.

[Reference Source](#)

R Core Team: **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria; 2017.

[Reference Source](#)

Reimand J, Arak T, Adler P, *et al.*: **g:Profiler-a web server for functional interpretation of gene lists (2016 update)**. *Nucleic Acids Res.* 2016a; **44**(W1): W83–89.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Reimand J, Kolde R, Arak T: **gProfileR: Interface to the 'g:Profiler' Toolkit**. 2016b.

[Reference Source](#)

RStudio inc: **htmltools: Tools for HTML**. 2017.

[Reference Source](#)

The UniProt Consortium: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res.* 2017; **45**(D1): D158–D169.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

van Iersel MP, Pico AR, Kelder T, *et al.*: **The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services**. *BMC Bioinformatics.* 2010; **11**: 5.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Wickham H, Francois R, Henry L, *et al.*: **dplyr: A Grammar of Data Manipulation**. 2017.

[Reference Source](#)

Wu C, Macleod I, Su AI: **BioGPS and MyGene.info: organizing online, gene-centric information**. *Nucleic Acids Res.* 2013; **41**(Database issue): D561–565.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Xie Y: **DT: A Wrapper of the JavaScript Library 'DataTables'**. 2016.

[Reference Source](#)

Zerbino DR, Achuthan P, Akanni W, *et al.*: **Ensembl 2018**. *Nucleic Acids Res.* 2018; **46**(D1): D754–D761.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 26 March 2018

<https://doi.org/10.5256/f1000research.15138.r31928>

© 2018 Simpson T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

T. Ian Simpson

School of Informatics , University of Edinburgh, Edinburgh, UK, UK

In this article the authors present BED, a biological entity database implemented as a Neo4J labelled property graph. In addition, they provide two R-packages (BED & neo4J) for the construction and query of such graphs that adhere to their data model. These packages include utility functions to facilitate graph construction from a range of commonly used and publicly available data sources. The software and database are well documented and available through GitHub and Docker (as a Docker image) respectively and proved straight forward to install and run.

There are several elements of the current manuscript that warrant commentary:

1. **Motivation/Rationale.** The authors have correctly identified an important problem with the integration of biological data that has been addressed before, not least by the resources/tools mentioned in the manuscript (Biomart, my gene, g:Profiler amongst others). They have chosen a particularly good approach (labelled property graphs) to build the data architecture to address such a problem and one that has recently been used to great effect by the EMBL-EBI Reactome team to model data related to biological pathways. Currently this manuscript somewhat undersells the potential for the tools that have been developed. Whilst allusion is made at various points to the fact that the software developed could be used by others to develop custom resources very little is presented as to the suitability of their approach for such. The "Abstract" states that existing resources "cannot be customised and optimised for any specific use" which is not correct and should be removed or re-worded to clarify the author's meaning. Whilst the implementation presented here is focussed primarily on gene level mappings it should be made clear throughout the manuscript that the general approach used could be (and indeed has been, see citation) used in other really quite different biological data modelling scenarios.
2. **Introduction.** The issue of "transitivity" is raised here, this is a complex issue for many biological data types that are far removed from the rigid structures of ontologies that commonly enforce it by definition. The meaning of "transitivity" in the context used here is not clear and warrants further explanation. This is particularly important later in the article

where decisions are being made about inferring mappings where they don't exist in the data. Some such inferences are entirely logical (e.g. using HGNC ids to link gene_ids between two resources that don't map directly to each other) but others are far more complex (e.g. mapping between species). The inclusion of deprecated identifiers is excellent and will help to close a notable gap in many existing resources for which mapping older data into more recent datasets can be extremely time consuming and frustrating. The authors comment on "mapping between different scopes" is unclear and should be clarified.

3. **Methods.** The sections "Feeding the Database" and "Querying the Database" are very brief and would benefit from much more detail about the functionality of the database creation and query system. Whilst these are covered in detail in the various pieces of documentation (including some very nice working examples) there is not enough in the manuscript itself to allow the reader to assess the available functionality.

4. Use Cases.

- There appears to be a discrepancy in gene counts from the Ensembl examples used in this section; the first example calls human Ensembl genes and returns 59,515 genes the second states the total number of human Ensembl genes to be 68,460.
- Figure 3. illustrates a relationship graph including deprecated BEIDs. Whilst *is_replaced_by* is clear, it is not clear (or defined anywhere) what the meaning of *is_associated_to* is and how that differs from *corresponds_to*. This should be clarified in the text.
- The sentence "The function *guessIdOrigin...*" appears out of place, unconnected to the surrounding text.
- The statement "Five identifiers were only..." and the following sentence should be combined and re-worded so that the explanation as to why 5 BEIDs were uniquely found by gProfiler is clearer.
- No validation or commentary has been presented to test the efficacy of inferences made by the query system. I would like to have seen an attempt made to check the veracity of mappings made in this way especially when the majority (c.80%) of extra Ensembl->EntrezID mappings recovered via BED were inferred.
- A "rough approximation" of timings for queries within BED and across other systems is not particularly informative. It would have been straightforward to automate a sampling approach to generate a mean response time (and a variance) to a defined set of query sizes/complexities to give the user a better understanding of how variable these response times are between the systems in practice. In addition, it would have been nice to see some analysis/discussion about the "scalability" of the system as this is likely to be of particular interest to end-users considering a similar modelling approach in other domains.
- Figure 5. The meaning of directionality here is not clear. Whilst I can see the benefit for provenance reasons i.e. a mapping from EntrezGene **to** RefSeq its meaning here is somewhat moot.

The work presented in this manuscript promises to be very useful for researchers wanting to use LPGs for data integration. The implementation and deployment have been very well executed so that they can be readily adopted and modified by end-users.

References

1. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, et al.: The Reactome Pathway Knowledgebase.

Nucleic Acids Res. 2018; **46** (D1): D649-D655 [PubMed Abstract](#) | [Publisher Full Text](#)

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

Partly

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?

Partly

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

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biological informatics, computational biology, neuroscience, statistics, machine learning

I confirm that I have read this submission and 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 Apr 2018

Patrice Godard, UCB, Belgium

Thanks for having taken the time to review this article and for your constructive comments that will help us to improve its quality. We are working on a second version. In the mean time we would like to provide you some feedback about the different issues you arose and how we are going to take your comments into account in the second version of our manuscript. Also we would like to inform you that we are going to use an updated version of the BED instance based on version 92 of Ensembl (released in April). Thus numbers provided in the article will slightly change in the next version.

- **Motivation/Rationale.** The authors have correctly identified an important problem with the integration of biological data that has been addressed before, not least by the resources/tools mentioned in the manuscript (Biomart, my gene, g:Profiler amongst others). They have chosen a particularly good approach (labelled property graphs) to build the data architecture to address such a problem and one that has recently been used to great effect by the EMBL-EBI Reactome team to model data related to biological pathways. Currently this

manuscript somewhat undersells the potential for the tools that have been developed. Whilst allusion is made at various points to the fact that the software developed could be used by others to develop custom resources very little is presented as to the suitability of their approach for such.

We will address this point in the conclusion of the next version of the article by mentioning in which context BED can be used.

- **The "Abstract" states that existing resources "cannot be customised and optimised for any specific use" which is not correct and should be removed or re-worded to clarify the author's meaning.**

This statement has also been questioned by the other referee although slightly differently. We wanted to highlight the point that the way the mapping is done by most of these resources (excepted BridgeDB) cannot be customized, optimized or extended by the user according to his knowledge or to internal, non-public or non-standard information. These tools are dedicated to a particular domain: they are focused on species, type of identifiers, and update frequencies (as stated in BridgeDB publication by [van Iersel et al. \(2010\)](#)). It's convenient because ready to use but not flexible as BridgeDB or BED which allow an empowered user to focus on required information. We are going to modify this statement to make it less ambiguous in the next version.

- **Whilst the implementation presented here is focused primarily on gene level mappings it should be made clear throughout the manuscript that the general approach used could be (and indeed has been, see citation) used in other really quite different biological data modelling scenarios.**

We will mention in the introduction that graph databases, specially Neo4j, have been used to model different kind of biological data.

- **Introduction. The issue of "transitivity" is raised here, this is a complex issue for many biological data types that are far removed from the rigid structures of ontologies that commonly enforce it by definition. The meaning of "transitivity" in the context used here is not clear and warrants further explanation. This is particularly important later in the article where decisions are being made about inferring mappings where they don't exist in the data. Some such inferences are entirely logical (e.g. using HGNC ids two link gene_ids between two resources that don't map directly to each other) but others are far more complex (e.g. mapping between species). The inclusion of deprecated identifiers is excellent and will help to close a notable gap in many existing resources for which mapping older data into more recent datasets can be extremely time consuming and frustrating. The authors comment on "mapping between different scopes" is unclear and should be clarified.**

The transitivity mechanism is managed by the 2 following relationships: "corresponds_to" and "is_associated_to". On one hand the "corresponds_to" relationships make the mapping transitive since 2 BEIDs which are connected through this kind of relationship are considered to identify the same BE through an "identifies" relationship. On the other hand a BEID which "is_associated_to" to another one does not automatically "identify" the same BE making this kind of relationship not available for indirect mappings. When the BED database is fed, the user chooses which relationship should be of type "corresponds_to" or of type "is_associated_to". For example, in the instance we provide, cross references provided by Ensembl from Ensembl gene ID to Entrez, HGNC and Vega gene ID are considered as "corresponds_to" relationships whereas cross references to miRbase,

Unigene and OMIM are considered as "is_associated_to" relationship. In Ensembl the Hs.745351 Unigene ID is mapped to ENSG00000184033 and to ENSG00000268651 Ensembl gene IDs which correspond to 2 different genes in Ensembl but also in Entrez and in HGNC and these genes are located on the same chromosome but at different positions. This Unigene identifier will be mapped to both Ensembl gene IDs but another external identifier mapped to only one of these 2 Ensembl gene ID won't be mapped to the other (the association to Hs.745351 won't be used indirectly).

An identifier scope is defined by the type of BE or probe, the source of the identifiers (database or platform) and the organism. Two scopes are different when at least one of these three elements is different. Mapping is the process to identify equivalent identifiers in two different scopes. This definition comes too late and is spread in the current version of the article. We will improve it in the next version.

- **Methods. The sections "Feeding the Database" and "Querying the Database" are very brief and would benefit from much more detail about the functionality of the database creation and query system. Whilst these are covered in detail in the various pieces of documentation (including some very nice working examples) there is not enough in the manuscript itself to allow the reader to assess the available functionality.**

We will list the available functions (at least the most relevant ones) in the next version of the article.

- **Use Cases.**
- **There appears to be a discrepancy in gene counts from the Ensembl examples used in this section; the first example calls human Ensembl genes and returns 59,515 genes the second states the total number of human Ensembl genes to be 68,460.**

59,515 corresponds to the number of BE (Gene in this case). 68,460 corresponds to the number of BEIDs (Ensembl gene IDs in this case). As explained multiple BEID can identify the same BE. In other words 59,515 BE are identified by 68,460 BEID.

- **Figure 3. illustrates a relationship graph including deprecated BEIDs. Whilst is_replaced_by is clear, it is not clear (or defined anywhere) what the meaning of is_associated_to is and how that differs from corresponds_to. This should be clarified in the text.**

See our answer about transitivity. We will make it clearer in the next version of the article.

- **The sentence "The function guessIdOrigin..." appears out of place, unconnected to the surrounding text.**

We agree and it will be moved to another place in the next version of the article (probably in the "Additional features" section).

- **The statement "Five identifiers were only..." and the following sentence should be combined and re-worded so that the explanation as to why 5 BEIDs were uniquely found by gProfiler is clearer.**

We will refine this part to address your comment and the similar one raised by the other reviewer.

- **No validation or commentary has been presented to test the efficacy of inferences made by the query system. I would like to have seen an attempt made to check the veracity of mappings made in this way especially when the majority (c.80%) of extra Ensembl->EntrezID mappings recovered via BED were**

inferred.

This kind of validation is quite difficult. We propose to use gene coordinates provided by the NCBI and Ensembl in order to compare the position on chromosomes of genes which are mapped by the different tool. Two mapped gene identifiers should have identical or similar locations. We will add these results in the next version of the article.

- **A "rough approximation" of timings for queries within BED and across other systems is not particularly informative. It would have been straightforward to automate a sampling approach to generate a mean response time (and a variance) to a defined set of query sizes/complexities to give the user a better understanding of how variable these response times are between the systems in practice. In addition, it would have been nice to see some analysis/discussion about the "scalability" of the system as this is likely to be of particular interest to end-users considering a similar modelling approach in other domains.**

We will make the analysis of mean response time with different kinds of queries and we will incorporate the results in the next version of the article. Scalability is not discussed because it highly depends on the graph database system. Here we use Neo4j for which scalability depends on the edition, community or enterprise (<https://neo4j.com/subscriptions/>).

- **Figure5. The meaning of directionality here is not clear. Whilst I can see the benefit for provenance reasons i.e. a mapping from EntrezGene to RefSeq its meaning here is somewhat moot.**

This comment has also been made by the other reviewer. We will remove the arrows from the edges in figure 5 to avoid the confusion about the use of directionality to find a path between two identifiers.

Competing Interests: No competing interests were disclosed.

Reviewer Report 05 March 2018

<https://doi.org/10.5256/f1000research.15138.r31026>

© 2018 Summer-Kutmon M et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Denise Slenter

Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, The Netherlands

Martina M. Summer-Kutmon

Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, The Netherlands

The article introduces BED a new identifier mapping tool. Using a graph database like Neo4j provides a **fast** way to query relationships between the biological entities and retrieve mappings of interest. The available source code is nicely documented and for bioinformaticians, setting up the database and running queries should be straight-forward.

Nevertheless, there are several major issues that we would like to comment on:

- Already in the abstract, it is indicated that current tools cannot be easily customized and optimized for any specific use. It is unclear what the authors actually mean with this statement and how this is solved through BED. Further on it is also stated that current tools are generally dedicated to a particular domain, which is also true for BED. BED only focuses on gene related identifiers (genes, transcripts, proteins) similar to mygene, Ensembl BioMart and g:Profiler.
- In the introduction, three main challenges are mentioned which are addressed by BED.
 - (1) Integration of mappings from different resources - very relevant but the difficult question is if transitive mappings are always biological meaningful. They can also lead to conflicting statements when resources show inconsistent relationships (we have experienced this when comparing Ensembl → UniProt and UniProt → Ensembl mappings) - how are you dealing with that? We want to state that mygene is also integrating mappings from multiple resources.
 - (2) Mapping of deprecated identifiers - this is indeed an interesting problem when analysing older datasets and the visualization in Figure 3 can be very useful when running into such issues. While you mention that BED contains all deprecated identifiers, it is not discussed why g:Profiler has five deprecated identifiers that are not in BED (Figure 4).
 - (3) Mapping scope - It is not clear why the automation of mapping between different scopes needs to be done differently and how BED is solving this. Importantly, BioMarts and mygene also provide easy ways to map between the different scopes (gene - gene / gene - protein / gene - homolog).
- Figure 3 - we believe that it would make sense to use two different edge styles for `is_replaced_by` and `is_associated_to` since they have very different meaning. Also check the layout (in this example, it looks like the blue node is placed over the edge from the purple to the light-purple node).
- Figure 5 - what do the bold borders of nodes mean in the network? Preferred identifiers? How are those selected? Additionally, when talking about the shortest relevant path, the arrows on the edges might be misleading and confusing (since there is no path from ILMN_1220595 to Q16552 taking the directionality into account).
- The authors shortly mention the neo2R package to build the database. The functionality is not discussed in detail and it is unclear why the existing R package provided by Neo4j (<https://neo4j.com/developer/r/>) was not used. Neo4j can also be easily queried from other programming languages. Are you planning to provide APIs in other languages that would allow the integration in tools other than R?
- While the conversion rate from Ensembl to Entrez Gene is very interesting, we are missing a comparison between the tools for real research examples, e.g. selection of several datasets and mapping from probe to Ensembl identifier / Entrez Gene identifier (one of the most common use cases in R workflows). This is also mentioned under the criteria for a software tool article in F1000: "The article should provide examples of suitable input data sets and include an example of the output that can be expected from the tool and how this output should be interpreted."

- Is it possible to only include edges from certain resources when performing the identifier conversion? Or do the users need to build their own database with only those selected resources?

As a final comment, we think that structure of the article is sometimes hard to follow and paragraphs are often not linked to each other. In the section "Converting identifiers" you state the following: "The aim of BED is to improve the **efficiency of identifier conversion** in a well defined context (organism, information resources of interest. . .) and not to replace biomaRt, mygene, gProfileR or other tools which provide many more features for many organisms and which should not be narrowed to this task for a complete comparison." We believe that this efficiency, especially in the context of run time, is the key advantage of this tool and this should be made more clear in the article (abstract/intro/conclusion).

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

Partly

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?

Partly

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

Partly

Competing Interests: We would like to note that the in the article mentioned BridgeDb framework is developed within our group.

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

Author Response 26 Apr 2018

Patrice Godard, UCB, Belgium

Thanks for having taken the time to review this article and for your constructive comments that will help us to improve its quality. We are working on a second version. In the mean time we would like to provide you some feedback about the different issues you arose and how we are going to take your comments into account in the second version of our

manuscript. Also we would like to inform you that we are going to use an updated version of the BED instance based on version 92 of Ensembl (released in April). Thus numbers provided in the article will slightly change in the next version.

- **Already in the abstract, it is indicated that current tools cannot be easily customized and optimized for any specific use. It is unclear what the authors actually mean with this statement and how this is solved through BED. Further on it is also stated that current tools are generally dedicated to a particular domain, which is also true for BED. BED only focuses on gene related identifies (genes, transcripts, proteins) similar to mygene, Ensembl BioMart and g:Profiler.**

This statement has also been questioned by the other referee although slightly differently. We wanted to highlight the point that the way the mapping is done by most of these resources (excepted BridgeDB) cannot be customized, optimized or extended by the user according to his knowledge or to internal, non-public or non-standard information. These tools are dedicated to a particular domain: they are focused on species, type of identifiers, and update frequencies (as stated in BridgeDB publication by [van Iersel et al. \(2010\)](#)). It's convenient because ready to use but not flexible as BridgeDB or BED which allow an empowered user to focus on required information. We are going to modify this statement to make it less ambiguous in the next version.

- **In the introduction, three main challenges are mentioned which are addressed by BED.**
- **(1) Integration of mappings from different resources - very relevant but the difficult question is if transitive mappings are always biological meaningful. They can also lead to conflicting statements when resources show inconsistent relationships (we have experienced this when comparing Ensembl → UniProt and UniProt → Ensembl mappings) - how are you dealing with that? We want to state that mygene is also integrating mappings from multiple resources.**

The transitivity mechanism is managed by the 2 following relationships: "corresponds_to" and "is_associated_to". On one hand the "corresponds_to" relationships make the mapping transitive since 2 BEIDs connected through this kind of relationship are considered to identify the same BE. On the other hand a BEID which "is_associated_to" to another one does not automatically "identify" the same BE making this kind of relationship not available for indirect mappings. When the BED database is fed, the user chooses which relationship should be of type "corresponds_to" or of type "is_associated_to" for each resource taken into account. For example, in the instance we provide, cross references provided by Ensembl from Ensembl gene ID to Entrez, HGNC and Vega gene ID are considered as "corresponds_to" relationships whereas cross references to miRbase, Unigene and OMIM are considered as "is_associated_to" relationship. In Ensembl the Hs.745351 Unigene ID is mapped to ENSG00000184033 and to ENSG00000268651 Ensembl gene IDs which correspond to 2 different genes in Ensembl but also in Entrez and in HGNC and these genes are located on the same chromosome but at different positions. This Unigene identifier will be mapped to both Ensembl gene IDs but another external identifier mapped to only one of these 2 Ensembl gene ID won't be mapped to the other (the association to Hs.745351 won't be used indirectly).

The cross references provided by Ensembl and Uniprot between Ensembl peptide IDs and Uniprot IDs are considered as "corresponds_to" relationship in the BED instance we provide. If mygene integrates mapping from multiple resources it does not apply transitive mapping between Ensembl, Entrez and HGNC gene IDs (as shown in figure 4) and it does not allow

the user to do it.

- **(2) Mapping of deprecated identifiers - this is indeed an interesting problem when analysing older datasets and the visualization in Figure 3 can be very useful when running into such issues. While you mention that BED contains all deprecated identifiers, it is not discussed why g:Profiler has five deprecated identifiers that are not in BED (Figure 4).**

These deprecated identifiers are not associated to any up-to-date identifier in Ensembl and as such they are not considered anymore for mapping in BED. We will develop this point in the next version of the article in order to make it clearer.

- **(3) Mapping scope - It is not clear why the automation of mapping between different scopes needs to be done differently and how BED is solving this. Importantly, BioMart and mygene also provide easy ways to map between the different scopes (gene - gene / gene - protein / gene - homolog).**

BED use the biological relationship between genes, transcript and peptides to convert identifiers. For example, when converting peptides identifiers from the same species it will use only mapping done at the peptide level and won't use mapping to transcript and gene mapping. This strategy seems to be applied by biomaRt but not by mygene nor by gProfilerR which map for example one Uniprot ID to all the Ensembl peptide ID coded by the same gene. For example the A6NI28 Uniprot identifier is unambiguously mapped to the ENSP00000298815 Ensembl peptide identifier by BED and biomaRt but is mapped to three additional Ensembl peptide identifiers (ENSP00000431776, ENSP00000434304 and ENSP00000435961 which are encoded by the same gene: ENSG00000165895) by mygene and gProfilerR. Mapping biological entities identifier which are not genes from two different organisms using ortholog information requires at least two steps in biomaRt, mygene and gProfilerR: one for find the ortholog gene and the other to find the relevant biological entity identifier. These two steps are integrated and transparent in BED. We will add clarifying sentences in the next version of the article to address this.

- **Figure 3 - we believe that it would make sense to use two different edge styles for is_replaced_by and is_associated_to since they have very different meaning. Also check the layout (in this example, it looks like the blue node is placed over the edge from the purple to the light-purple node).**

The visNetwork library only provides 2 types of edges: solid or dash. And we would prefer not using too many colors for different types of relationships. The "is_replaced_by" and "is_associated_to" relationship can easily be differentiated using the colors of the nodes: if the nodes have the same color it is an "is_replaced_by" relationship"; if the nodes have different colors it is an "is_associated_to" relationship. In this kind of graph "is_known_as", "identifies" (optional) or "targets" (optional) relationships can also be differentiated according to the shapes of the nodes. We will clarify this point in the figure legend.

We will also fix the layout issue of figure 3 in the next version of the article.

- **Figure 5 - what do the bold borders of nodes mean in the network? Preferred identifiers? How are those selected? Additionally, when talking about the shortest relevant path, the arrows on the edges might be misleading and confusing (since there is no path from ILMN_1220595 to Q16552 taking the directionality into account).**

Bold borders in this figure indeed meant preferred identifiers. In the first version of the BED instance we provided (bed-ucb-human:2018.01.03), the preferred status of RefSeq transcripts and peptides is determined according to the status field provided in the

gene2refseq file provided by the NCBI. The ID is preferred if the status is "MODEL". The way to define the preferred status of Entrez gene, RefSeq transcripts and peptides will change in the next version of this instance where we will consider the assembly information also provided in the gene2refseq file: identifiers associated to non-alternative assembly will be "preferred".

We will remove the arrows from the edges in figure 5 to avoid the confusion about the use of directionality to find a path between two identifiers.

- **The authors shortly mention the neo2R package to build the database. The functionality is not discussed in detail and it is unclear why the existing R package provided by Neo4j (<https://neo4j.com/developer/r/>) was not used. Neo4j can also be easily queried from other programming languages. Are you planning to provide APIs in other languages that would allow the integration in tools other than R?**

Two reasons motivated our choice to develop neo2R : (i) The development of Rneo4j package was on hold for a long time period (according to github commits) and (ii) it used legacy cypher HTTP endpoint. We wanted to use the transactional HTTP endpoint as recommended in the neo4j documentation (<https://neo4j.com/docs/rest-docs/current/#rest-api-cypher>). As the scope of the article is on the biological entity mapping and not on Neo4j as such, we don't want to put emphasis on this point because we don't think we provide strong additional value at this level.

We do not plan to provide API in other languages but we would be happy if it is done by other developers and we would be ready to help in this frame. Indeed one of the reason to make the BED package publicly available under a GPL-3 license is to allow the community to build on it and to improve it.

- **While the conversion rate from Ensembl to Entrez Gene is very interesting, we are missing a comparison between the tools for real research examples, e.g. selection of several datasets and mapping from probe to Ensembl identifier / Entrez Gene identifier (one of the most common use cases in R workflows). This is also mentioned under the criteria for a software tool article in F1000: "The article should provide examples of suitable input data sets and include an example of the output that can be expected from the tool and how this output should be interpreted."**

We will provide such an example in the next version. The example will be focused on the comparison of results from different experiments with different designs: different microarray platforms and organisms.

- **Is it possible to only include edges from certain resources when performing the identifier conversion? Or do the users need to build their own database with only those selected resources?**

As mentioned here-above, the conversion strategy is defined when feeding the BED database and the use of the relationships: "corresponds_to" and "is_associated_to". At the end-user level, refinements of mapping can be achieved by using the "restricted" (which focus the mapping to non-deprecated identifiers) and the "preFilter" (which focus the mapping to preferred identifiers) parameters. Also the "getDirectProduct" and "getDirectOrigin" functions allow the user to find direct products or direct origins of molecular biology processes. For example the direct products of an Ensembl gene ID will be Ensembl transcript IDs. This is particularly useful when the user wants to focus on canonical transcription or translation events when this information is available (this is the case for

Ensembl transcripts and peptides).

- **As a final comment, we think that structure of the article is sometimes hard to follow and paragraphs are often not linked to each other. In the section “Converting identifiers” you state the following: “The aim of BED is to improve the efficiency of identifier conversion in a well defined context (organism, information resources of interest. . .) and not to replace biomaRt, mygene, gProfileR or other tools which provide many more features for many organisms and which should not be narrowed to this task for a complete comparison.” We believe that this efficiency, especially in the context of run time, is the key advantage of this tool and this should be made more clear in the article (abstract/intro/conclusion).**

We adopted the structure recommended by F1000Research for a “software tool article”.

Nevertheless we take note of this comment and we will try to improve the flow of the text in the next version of the article.

We will put higher emphasis on the efficiency statement in the next version of the article.

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

For pre-submission enquiries, contact research@f1000.com

F1000Research