hmdb Documentation

Release 0.1.1-dev

Colin Birkenbihl, Charles Tapley Hoyt

Apr 29, 2019

Contents:

Bio2BEL HMDB is a package which allows the user to work with a local sqlite version of the Human Metabolome Database (HMDB).

Next to creating the local database there are also functions provided, which will enrich given Biological Expression Language (BEL) graphs with information about metabolites, proteins and diseases, that is present in HMDB.

HMDB BEL namespaces for these BEL graphs can be written.

CHAPTER 1

Installation

1.1 Get the Latest

Download the most recent code from GitHub with:

```
$ python3 -m pip install git+https://github.com/bio2bel/hmdb.git
```

1.2 For Developers

Clone the repository from GitHub and install in editable mode with:

```
$ git clone https://github.com/bio2bel/hmdb.git
$ cd hmdb
$ python3 -m pip install -e .
```

CHAPTER 2

Setup

2.1 1. Create a `bio2bel_hmdb.Manager` object

```
>>> from bio2bel_hmdb import Manager
>>> manager = Manager()
```

2.2 2. Create the tables in the database

```
>>> manager.create_all()
```

2.3 3. Populate the database

This step will take sometime since the HMDB XML data needs to be downloaded, parsed, and fed into the database line by line.

```
>>> manager.populate()
```

CHAPTER 3

Enrichment

3.1 Enrich BEL graphs

In the current build it is possible to enrich BEL graphs containing metabolites with associated disease or protein information and to enrich BEL graphs containing disease or protein information with associated metabolites. This can be done with the functions further explained in BEL Serialization

3.1.1 2. Enriching BEL graphs

Using an BEL graph with metabolites (represented using the HMDB namespace) it can be enriched with disease and protein information from HMDB.

2.1 Metabolites-Proteins

For a graph containing metabolites:

```
>>> enrich_metabolites_proteins(bel_graph, manager)
```

The result of this will be a BEL graph which now includes relations between the metabolites and proteins.

For a graph containing proteins (named using uniprot identifiers):

```
>>> enrich_proteins_metabolites(bel_graph, manager)
```

This will result in a BEL graph where the proteins are linked to associated metabolites.

2.2 Metabolites-Diseases

For a graph containing metabolites:

```
>>> enrich_metabolites_diseases(bel_graph, manager)
```

The result of this will be a BEL graph which now includes relations between the metabolites and diseases.

For a graph containing diseases (named using HMDB identifiers):

```
>>> enrich_diseases_metabolites(bel_graph, manager)
```

This will result in a BEL graph where the diseases are linked to associated metabolites.

bio2bel_hmdb.enrich.**enrich_diseases_metabolites**(*graph: pybel.struct.graph.BELGraph*, *manager: Optional[bio2bel_hmdb.manager.Manager] = None*)

Enrich a given BEL graph, which includes HMDB diseases with HMDB metabolites, which are associated to the diseases.

bio2bel_hmdb.enrich.**enrich_metabolites_diseases**(*graph: pybel.struct.graph.BELGraph*, *manager: Optional[bio2bel_hmdb.manager.Manager] = None*)

Enrich a given BEL graph, which includes metabolites with diseases, to which the metabolites are associated.

bio2bel_hmdb.enrich.**enrich_metabolites_proteins**(*graph: pybel.struct.graph.BELGraph*, *manager: Optional[bio2bel_hmdb.manager.Manager] = None*)

Enrich a given BEL graph, which includes metabolites with proteins, that are associated to the metabolites.

bio2bel_hmdb.enrich.**enrich_proteins_metabolites**(*graph: pybel.struct.graph.BELGraph*, *manager: Optional[bio2bel_hmdb.manager.Manager] = None*)

Enrich a given BEL graph, which includes uniprot proteins with HMDB metabolites, that are associated to the proteins.

CHAPTER 4

Manager

The Manager is a key component of HMDB. This class is used to create, populate and query the local HMDB version.

class bio2bel_hmdb.manager.**Manager**(**args*, ***kwargs*)

Metabolite-proteins and metabolite-disease associations.

count_biofunctions() → int

Count the number of biofunctions in the database.

count_cellular_locations()

Count the number of cellular locations in the database.

count_diseases() → int

Count the number of diseases in the database.

count_metabolites() → int

Count the number of metabolites in the database.

count_pathways() → int

Count the number of pathways in the database.

count_proteins() → int

Count the number of proteins in the database.

count_references()

Count the number of literature references in the database.

count_tissues() → int

Count the number of tissues in the database.

get_hmdb_accession()

Create a list of all HMDB metabolite identifiers present in the database.

Return type list

get_hmdb_diseases()

Create a list of all disease names present in the database.

Return type list

get_metabolite_by_accession(*hmdb_metabolite_accession: str*) → Optional[bio2bel_hmdb.models.Metabolite]

Query the constructed HMDB database and extract a metabolite object.

Parameters **hmdb_metabolite_accession** – HMDB metabolite identifier

Example:

```
>>> import bio2bel_hmdb
>>> manager = bio2bel_hmdb.Manager()
>>> manager.get_metabolite_by_accession("HMDB00072")
```

get_reference_by_pubmed_id(*pubmed_id: str*) → Optional[bio2bel_hmdb.models.Reference]

Get a reference by its PubMed identifier if it exists.

Parameters **pubmed_id** – The PubMed identifier to search

is_populated() → bool

Check if the database is already populated.

populate(*source: Optional[str] = None*, *map_dis: bool = True*, *group_size: int = 500000*)

Populate the database with the HMDB data.

Parameters

- **source** – Path to an .xml file. If None the whole HMDB will be downloaded and used for population.
- **map_dis** – Should diseases be mapped?

query_disease_associated_metabolites(*disease_name: str*) → List[bio2bel_hmdb.models.Metabolite]

Query function that returns a list of metabolite-disease interactions, which are associated to a disease.

Parameters **disease_name** – HMDB disease name

query_metabolite_associated_diseases(*hmdb_metabolite_id: str*) → List[bio2bel_hmdb.models.Disease]

Query the constructed HMDB database to get the metabolite associated disease relations for BEL enrichment

Parameters **hmdb_metabolite_id** – HMDB metabolite identifier

query_metabolite_associated_proteins(*hmdb_metabolite_id: str*) → Optional[List[bio2bel_hmdb.models.Protein]]

Query the constructed HMDB database to get the metabolite associated protein relations for BEL enrichment

Parameters **hmdb_metabolite_id** – HMDB metabolite identifier

query_protein_associated_metabolites(*uniprot_id*)

Query function that returns a list of metabolite-disease interactions, which are associated to a disease.

Parameters **uniprot_id** (*str*) – uniprot identifier of a protein for which the associated metabolite relations should be outputted

Return type list

summarize() → Mapping[str, int]

Summarize the contents of the database in a dictionary.

CHAPTER 5

Models

The data model for the local HMDB version consists of 22 different tables that represent the relations found in the original HMDB data.

class bio2bel_hmdb.models.**Biofluid**(***kwargs*)
Table storing the different biofluids.

biofluid
Name of the biofluid

class bio2bel_hmdb.models.**Biofunction**(***kwargs*)
Table for storing the 'biofunctions' annotations

class bio2bel_hmdb.models.**CellularLocation**(***kwargs*)
Table for storing the cellular location GO annotations

class bio2bel_hmdb.models.**Disease**(***kwargs*)
Table storing the diseases and their ids.

dion
Disease Ontology name for this disease. Found using string matching

hpo
Human Phenotype Ontology name for this disease. Found using string matching

mesh_diseases
MeSH Disease name for this disease. Found using string matching

name
Name of the disease

omim_id
OMIM identifier associated with the disease

serialize_to_bel() → pybel.dsl.node_classes.Pathology
Function to serialize a disease object to a PyBEL node data dictionary.

class bio2bel_hmdb.models.**Metabolite**(***kwargs*)
Table which stores the metabolites and all the information provided about them in HMDB.

accession
Accession ID for the metabolite

average_molecular_weight
Average molecular weight of the metabolite

bigg_id
Bigg ID of the metabolite

biocyc_id
BioCyc ID of the metabolite

cas_registry_number
Cas registry number of the metabolite

chebi_id
ChEBI identifier of the metabolite

chemical_formula
Chemical formula of the metabolite

chemspider_id
Chemspider ID of the metabolite

creation_date
Date when the metabolite was included into HMDB

description
Description including some information about the metabolite

drugbank_id
DrugBank identifier of the metabolite

drugbank_metabolite_id
Drugbank metabolite ID of the metabolite

foodb_id
FooDB ID of the metabolite

het_id
Het ID of the metabolite

inchi
InChi of the metabolite

inchikey
InCHI key of the metabolite

iupac_name
IUPAC name of the metabolite

kegg_id
KEGG ID of the metabolite

knapsack_id
Knapsack ID of the metabolite

metagene
Metagene ID of the metabolite

metlin_id
Metlin ID of the metabolite

monisotopic_molecular_weight
Monisotopic weight of the molecule

name
Name of the metabolite

nugowiki
NukoWiki ID of the metabolite

phenol_explorer_compound_id
Phenol explorer compound ID of the metabolite

phenol_explorer_metabolite_id
Phenol explorer metabolite ID of the metabolite

pubchem_compound_id
PubChem compound ID of the metabolite

serialize_to_bel() → pybel.dsl.node_classes.Abundance
Function to serialize a metabolite object to a PyBEL node data dictionary.

smiles
Smiles representation of the metabolite

state
Aggregate state of the metabolite

synthesis_reference
Synthesis reference citation of the metabolite

trivial
Trivial name of the metabolite

update_date
Date when the entry was last updated

version
Current version listing that metabolite

wikipedia
Wikipedia name of the metabolite

class bio2bel_hmdb.models.**MetaboliteBiofluid**(***kwargs*)
Table representing the Metabolite and Biofluid relations.

class bio2bel_hmdb.models.**MetaboliteBiofunction**(***kwargs*)
Table storing the many to many relations between metabolites and cellular location GO annotations

class bio2bel_hmdb.models.**MetaboliteCellularLocation**(***kwargs*)
Table storing the many to many relations between metabolites and cellular location GO annotations

class bio2bel_hmdb.models.**MetaboliteDiseaseReference**(***kwargs*)
Table storing the relations between disease and metabolite

class bio2bel_hmdb.models.**MetabolitePathway**(***kwargs*)
Table storing the different relations between pathways and metabolites.

class bio2bel_hmdb.models.**MetaboliteProtein**(***kwargs*)
Table representing the many to many relationship between metabolites and proteins.

class bio2bel_hmdb.models.**MetaboliteReference**(***kwargs*)
Table representing the many to many relationship between metabolites and references.

class bio2bel_hmdb.models.**MetaboliteSynonym**(**kwargs)
Table storing the synonyms of metabolites.

synonym
Synonym for the metabolite

class bio2bel_hmdb.models.**MetaboliteTissue**(**kwargs)
Table storing the different relations between tissues and metabolites

class bio2bel_hmdb.models.**Pathway**(**kwargs)
Table storing the different tissues.

kegg_map_id
KEGG Map identifier of the pathway.

name
Name of the pathway.

smpdb_id
SMPDB identifier of the pathway.

class bio2bel_hmdb.models.**PropertyKinds**(**kwargs)
Table storing the 'kind' of chemical properties e.g. logP.

Not used for BEL enrichment

kind
the 'kind' of chemical properties e.g. logP, melting point etc

class bio2bel_hmdb.models.**PropertySource**(**kwargs)
Table storing the sources of properties e.g. software like 'ALOGPS'.

Not used for BEL enrichment

class bio2bel_hmdb.models.**PropertyValues**(**kwargs)
Table storing the values of chemical properties.

Not used for BEL enrichment

value
value of a chemical property (e.g. logp) that will be linked to the properts and metabolites

class bio2bel_hmdb.models.**Protein**(**kwargs)
Table to store the protein information.

gene_name
Gene name of the protein coding gene

protein_accession
HMDB accession number for the protein

protein_type
Protein type like 'enzyme' etc.

serialize_to_bel() → pybel.dsl.node_classes.Protein
Function to serialize a protein object to a PyBEL node data dictionary.

uniprot_id
UniProt identifier of the protein

class bio2bel_hmdb.models.**Reference**(**kwargs)
Table storing literature references.

pubmed_id
PubMed identifier of the article

reference_text
Citation of the reference article

class bio2bel_hmdb.models.**SecondaryAccession**(***kwargs*)
Table storing the different synonyms of metabolites.

secondary_accession
Other accession numbers for the metabolite

class bio2bel_hmdb.models.**Tissue**(***kwargs*)
Table storing the different tissues.

tissue
Tissue type

CHAPTER 6

Creating BEL Namespaces

CHAPTER 7

Current Status

7.1 What is still missing?

Not all of the information found in HMDB is yet integrated.

Bio2BEL HMDB does not yet include: - Taxonomy information - Spectra information - Experimental properties (datamodel is implemented but tables will not get populated) - Predicted properties (datamodel is implemented but tables will not get populated) - Normal concentration - Abnormal concentration

Bio2BEL HMDB still lacks functions to: - convert metabolite namespaces from and to HMDB identifiers - query functions (only querying with metabolite identifiers for diseases and proteins and vice versa is supported right now)

7.2 Roadmap

The next steps in the development of Bio2BEL HMDB are:

1. add namespace mappings from metabolite HMDB identifiers to different databases/namespaces
2. add query functions for several tables and entries
3. change BEL enrichment functions to automatically work even when pathology nodes are not in HMDB disease namespace
4. include missing HMDB tables and relations listed above
5. maybe add parallelization to the database population to improve run time

CHAPTER 8

Indices and tables

- genindex
- modindex
- search

Python Module Index

b

Index

G

H

I

K

M

N

O

P

Q

R

S

T

U

V

W