PEMT Documentation

Release 0.0.3-dev

Yojana Gadiya and Andrea Zaliani

Aug 08, 2022

CONTENTS:

Release notes : https://github.com/Fraunhofer-ITMP/PEMT/releases

CHAPTER

ONE

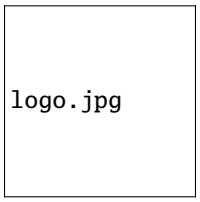
PEMT: A PATENT ENRICHMENT TOOL FOR DRUG DISCOVERY.

PEMT takes a two-step approach to collect patent documents relevant for drug discovery.

1. The `chemical_extractor` module extraction of chemicals that directly regulate (i.e. activation or inhibition) genes of interest based on functional or biochemical assays found within ChEMBL.
2. The `patent_extractor` module interlinking these chemicals to patent documents by systematically querying SureChEMBL, a patent database.

1.1 Welcome to PEMT's documentation!

Release notes : https://github.com/Fraunhofer-ITMP/PEMT/releases



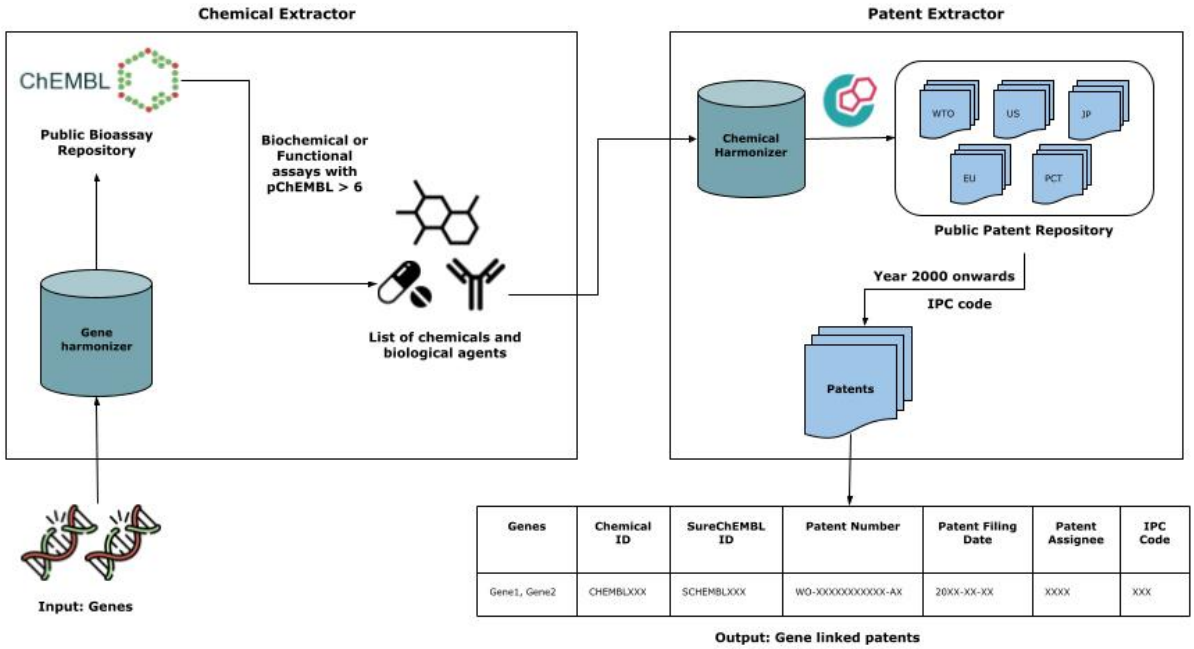
1.1.1 PEMT: A patent enrichment tool for drug discovery.

PEMT takes a two-step approach to collect patent documents relevant for drug discovery.

1. The `chemical_extractor` module extraction of chemicals that directly regulate (i.e. activation or inhibition) genes of interest based on functional or biochemical assays found within ChEMBL.
2. The `patent_extractor` module interlinking these chemicals to patent documents by systematically querying SureChEMBL, a patent database.

1.1.2 General info

PEMT is a patent extractor tool that enables users to retrieve patents relevant to drug discovery. The framework is depicted in the graphic below



1.1.3 Installation

You can install PEMT package from pypi.

```
# Use pip to install the latest release
$ python3 -m pip install pemt
```

You may instead want to use the development version from Github, by running

```
$ python3 -m pip install git+https://github.com/Fraunhofer-ITMP/PEMT.git
```

For contributors, the repository can be cloned from GitHub and installed in editable mode using:

```
$ git clone https://github.com/Fraunhofer-ITMP/PEMT.git
$ cd PEMT
$ python3 -m pip install -e .
```

1.1.4 Dependency

- Python 3.8+
- Installation of chromedriver

Mandatory

- Pandas
- CheMBL Webresource
- PubChemPy
- Chromedriver

For API information to use this library, see the *Developmental Guide*.

1.1.5 Issues

If you have difficulties using PEMT, please open an issue at our GitHub repository.

1.1.6 Disclaimer

PEMT is a scientific tool that has been developed in an academic capacity, and thus comes with no warranty or guarantee of maintenance, support, or back-up of data.

1.2 Command Line Interface

PEMT commands.

1.2.1 pemt

Run PEMT.

```
pemt [OPTIONS] COMMAND [ARGS]...
```

run-chemical-extractor

Extract chemicals for genes of interest

```
pemt run-chemical-extractor [OPTIONS]
```

Options

--name <name>

Required Name of the analysis that is to be run

--data <data>

Required Path to tab-separated gene data file

--input-type <input_type>

Type of data file i.e. 'tab' for tsv or 'comma' for csv files

--uniprot, --no-uniprot

Boolean value indicating whether the gene data file has uniprot ids or not.

run-patent-extractor

Extract patent for filtered chemicals

```
pemt run-patent-extractor [OPTIONS]
```

Options

--name <name>

Required Name of the analysis that is to be run

--os <os>

The OS system on which is the script is running

Options

linux | mac | windows

--chromedriver-path <chromedriver_path>

Required The path where the chromedriver can be found on the users computer

--year <year>

The year from which you want to retrive patents from

--chemical, --no-chemical

Boolean value indicating whether the chemical data is provided by the user or not

--chemical-data <chemical_data>

Path to tab-separated chemical data file with single column of chembl_ids

run-pemt

Run the PEMT tool with gene data

```
pemt run-pemt [OPTIONS]
```

Options

--name <name>

Required Name of the analysis that is to be run

--data <data>

Required Path to tab-separated gene data file

--input-type <input_type>

Type of data file i.e. 'tab' for tsv or 'comma' for csv files

--uniprot, --no-uniprot

Boolean value indicating whether the gene data file has uniprot ids or not.

--chromedriver-path <chromedriver_path>

Required The path where the chromedriver can be found on the users computer

--os <os>

The OS system on which is the script is running

Options

linux | mac | windows

--year <year>

The year from which you want to retrive patents from

1.3 Developmental Guide

1.3.1 Core Module APIs

Gene Harmonizer

pemt.utils.**get_hgnc_id**() → Dict[str, str]

Mapping dictionary for HGNC symbol to HGNC identifiers

pemt.utils.**hgnc_to_chembl**(*chemical_mapper: Dict[str, str]*, *uniprot_mapper: Dict[str, str]*, *hgnc_symbol: str*) → Optional[str]

Mapping HGNC symbol to ChEMBL identifiers.

Parameters

- **chemical_mapper** – A dictionary mapping the UNIPROT identifiers to ChEMBL
- **uniprot_mapper** – A dictionary mapping the HGNC identifiers to UNIPROT
- **hgnc_symbol** – A HGNC symbol

pemt.utils.**uniprot_to_chembl**(*chemical_mapper: dict*, *uniprot_id: str*) → Optional[str]

Mapping UniProt identifiers to ChEMBL identifiers.

Parameters

- **chemical_mapper** – A dictionary mapping the UNIPROT identifiers to ChEMBL
- **uniprot_id** – UNIPROT identifier of a protein

pemt.utils.**get_chemical_names**(*chembl_id: str*) → str

Method to get chemical name from ChEMBL id.

Parameters

chembl_id – ChEMBL identifier of a compound

pemt.utils.**uniprot_to_chembl**()

Mapping UniProt identifiers to ChEMBL identifiers.

Parameters

- **chemical_mapper** – A dictionary mapping the UNIPROT identifiers to ChEMBL
- **uniprot_id** – UNIPROT identifier of a protein

pemt.utils.**hgnc_to_chembl**()

Mapping HGNC symbol to ChEMBL identifiers.

Parameters

- **chemical_mapper** – A dictionary mapping the UNIPROT identifiers to ChEMBL

- **uniprot_mapper** – A dictionary mapping the HGNC identifiers to UNIPROT
- **hgnc_symbol** – A HGNC symbol

Chemical Extractor

pemt.chemical_extractor.experimental_data_extraction.**extract_chemicals**()

Enrich genes with chemical data from CheMBL bioassays.

Parameters

- **analysis_name** – The name of the analysis you want to run. This name would be used to save the resultant file
- **gene_list** – The list of gene you want to extract chemicals for.
- **gene_file_path** – The path of the gene file
- **file_separator** – The separator used within the file. This can be 'comma', 'tab', or 'semi-colon'. By default,

the file separator is set to csv. :param is_uniprot: A boolean value indicating whether the given gene list or file containing uniprot ids or HGNC symbols. By default, the value is set to False indicating that a "symbol" column is present with the respective HGNC symbols. If set to True, the file with "uniprot" column is expected.

Chemical Harmonizer

pemt.patent_extractor.patent_chemical_harmonizer.**harmonize_chemicals**()

Method that allows mapping from ChEMBL to SureChEMBL identifiers.

Parameters

- **analysis_name** – The name of the analysis you want to run. This name would be used to save the resultant file.
- **from_genes** – Boolean indicating where the process needs to get chemicals based on genes or not.

Patent Extractor

pemt.patent_extractor.patent_enrichment.**extract_patent**()

Extract and store all valid patent document metadata.

Parameters

- **analysis_name** – Name of the analysis.
- **os_system** – The OS on which the code is running. It can be either of these: linux, mac, window.
- **chrome_driver_path** – The path of the chrome driver is located.
- **patent_year** – The cutt-off year for searching the patent documents

CHAPTER

TWO

INDICES AND TABLES

- genindex
- modindex
- search

PYTHON MODULE INDEX

p

INDEX

Symbols

E

G

H

M

P

U