RDTextractor

Installation

```
pip install -r requirements.txt
python setup.py install
```

Once it's installed, you can run the extractor by typing:

```
% extract -h
```

Introduction

This tool is designed to extract data from the *in vivo* repeat-dose toxicity (RDT) studies' database generated within the context of the eTOX project. These data are expanded using an histopathological observation and an anatomical entity ontology. The histopathological ontology is obtained from Novartis and can be used under the Apache License 2.0. The anatomical entities ontology is extracted from the following paper:

- Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. Genome Biol. 2005; 6(3): R29

It can work with version 2016.1 or with later versions. For the former, you need to request access to the data files from us. For the latter, you need to have the Oracle database provided by Lhasa installed and run the script from the Oracle server. Additionally, you'll need to set up the ORACLE_HOME and LD_LIBRARY_PATH environment variables. This project is an extension of the work published in the following paper:

- López-Massaguer O, Pinto-Gil K, Sanz F, Amberg A, Anger LT, Stolte M, Ravagli C, Marc P, Pastor M. Toxicol Sci. 2018 Mar; 162(1): 287–300.

Manual

Exract studies' findings based on the given filtering and the organs' and morphological changes' ontologies-based expansions of these findings.

- Required arguments:
    - -a / --organ ORGAN Anatomical entity that the finding refers to (case insensitive). You can filter for more than one organ by passing a blank space-separated list.
- Optional arguments:
    - Version-related arguments:
        - -v / --version *{local, oracle}* Vitic database version (default: oracle).
        - -d / --sid SID If working with the Oracle database, provide the Oracle SID's.
        - -u / --user USER If working with the Oracle database, provide the Oracle database user name.
        - -p / --passw PASSW If working with the Oracle database, provide the Oracle database password.
    - Study design-related arguments:

        - -i / --min_exposure MIN_EXPOSURE Minimum exposure period (days).
        - -e / --max_exposure MAX_EXPOSURE Maximum exposure period (days).
        - -r / --route *{Cutaneous, Diertary, Oral, Oral gavage, Intragastric, Nasogastric, Oropharyngeal, Endotracheal, Intra-articular, Intradermal, Intraesophageal, Intraileal, Intramuscular, Subcutaneous, Intraocular, Intraperitoneal, Intrathecal, Intrauterine, Intravenous, Intravenous bolus, Intravenous drip, Parenteral, Nasal, Respiratory (inhalation), Percutaneous, Rectal, Vaginal, Subarachnoid}* Administration route (case insensitive). You can filter for more than one administration route by passing a blank space-separated list.
        - -s / --species *{Mouse, Rat, Hamster, Guinea pig, Rabbit, Dog, Pig, Marmoset, Monkey, Baboon}* Species (case insensitive). You can filter for more than one species by passing a blank space-separated list.
        - -x / --sex *{F,M,Both}* Study design sex.
    - Finding-related arguments:
        - -m / --observation OBSERVATION Morphological change type that the finding refers to (case insensitive). You can filter for more than one morphological change by passing a blank space-separated list.
        - -t / --treatment_related Keep only treatment-related findings.
    - Output-related arguments:
        - -o / --output_basename OUTPUT_BASENAME Output file base name. Two output files will be generated: basename_quant.tsv and basename_qual.tsv, with quantitative and qualitative results respectively. (default: output).

Use examples

1. Extract all studies with liver-related findings

- vitic 2016.1:
  ```
  extract -v local -a liver
  ```
- latest vitic:
  ```
  extract -v oracle -d ORACLE_SID -u ORACLE_USER -p ORACLE_PASSWORD -a liver
  ```

2. Extract all studies with liver- and kidney-related findings

Note that you can filter for more than one organ by passing a blank space-separated list.

- vitic 2016.1:
  ```
  extract -v local -a liver kidney
  ```
- latest vitic:
  ```
  extract -v oracle -d ORACLE_SID -u ORACLE_USER -p ORACLE_PASSWORD -a liver
  kidney
  ```

3. Extract only studies of interest

Filter the studies of interest based on exposure time (days), administration route, and species. Note that for route and species you can filter for more than one value by passing a blank space-separated list.

- Using long arguments:
  ```
  extract -v local --organ liver --min_exposure 1 --max_exposure 10 --route ORAL
  --species MOUSE RAT
  ```
- Using short arguments:
  ```
  extract -v local -a liver -i 1 -e 10 -r ORAL -s MOUSE RAT
  ```

4. Extract treatment-related findings only

```
extract -v local -a liver -i 1 -e 10 -r ORAL -s MOUSE RAT -t
```

5. Output example

After extracting data using this tool, two output files are generated, one with quantitative and the other with qualitative data. Both have five common columns, namely:

- subst_id: Substance ID.
- study_count: Number of relevant studies (according to the current filtering scheme) in which the substance appears.
- dose_max: Maximum dose at which the substance has been tested among the relevant studies.
- dose_min: Minimum dose at which the substance has been tested among the relevant studies.
- is_active: Boolean indicating whether the substance has been found to have any toxicity according to the current finding-related filtering criteria.

After these, there is a column for each relevant finding. In these columns a value is provided if the finding is reported for the given substance, and it is empty otherwise. The value will be 1 in the qualitative file and the minimum dose at which the finding is reported in the quantitative file.

This is an example of the qualitative output:

subst_id	study_count	dose_max	dose_min	is_active	liver_basophilic focus	liver_benign tumor	liver_bile duct hyperplasia	liver_cell adaptation/injury/death	liver_cell injury/death	liver_cellular adaptation	liver_cellular adaptation of growth	liver_chronic inflammatory/proliferative/metaplastic change	liver_clear cell focus	liver_compartmental cell accumulation	liver_congestion	liver_cyst, with squamous/ciliated epithelium	liver_cytoplasmic alteration	liver_decreased
X	9	650	1	True			1	1	1	1	1							1
X	1	504	97.9	False														
X	1	25	8	False														
X	2	400	400	False														
X	1	10	1	False														
X	1	1000	1000	False														
X	8	2000	10	True				1			1							
X	1	110	64	False														
X	1	140	120	True				1	1	1								1
X	8	2000	1	True	1	1	1	1	1	1	1							1
X	4	150	18.9	False														
X	1	2000	2000	False														
X	1	30	30	False														
X	1	3	0.2	False														
X	1	2	1	False														
X	4	120.5	0.5	True				1	1		1							
X	2	650	60	True				1		1	1							1
X	5	500	5	True				1	1	1	1				1			
X	1	50	13	False														
X	5	12	0.04	True				1	1		1							
X	1	100	100	False														
X	2	1000	50	True				1										
X	1	350	71	False														
X	6	500	50	True				1	1	1	1							1
X	1	150	10	True				1	1									
X	5	450	45	True				1		1				1				

This is an example of the quantitative output:

_subst_id	study_count	dose_max	dose_min	is_active	erythropoiesis	liver_increased mitoses	liver_increased, cellular content	liver_increased, histiocyte number	liver_increased, lipid content	liver_infarct	liver_infectious diseases	liver_inflammation	liver_inflammation, granulomatous	liver_inflammatory cell infiltration	liver_inflammatory processes	liver_intracellular accumulation	liver_intracellular
X	9	650	1	True		5	1		1			650			650	650	
X	1	504	97.9	False													
X	1	25	8	False													
X	2	400	400	False													
X	1	10	1	False													
X	1	1000	1000	False													
X	8	2000	10	True													
X	1	110	64	False													
X	1	140	120	True												120	
X	8	2000	1	True			1		1			10		5	5	5	
X	4	150	18.9	False													
X	1	2000	2000	False													
X	1	30	30	False													
X	1	3	0.2	False													
X	1	2	1	False													
X	4	120.5	0.5	True		0.5											
X	2	650	60	True													
X	5	500	5	True			5		5			5		5	5	5	
X	1	50	13	False													
X	5	12	0.04	True		1.2										0.12	
X	1	100	100	False													
X	2	1000	50	True												50	