
Setting up Neo4j/PostgreSQL Database

N. Greiner

August 18, 2022

1 Introduction

1.1 Neo4j

In progress.

1.2 PostgreSQL

In progress.

2 Prerequisites

2.1 Conda environment

1. Install Conda
2. Clone GitHub repository:

```
git clone https://github.com/BackofenLab/protein-graph-database.git
cd protein-graph-database/
```

3. Create a conda environment:

```
make env
```

4. Activate the conda environment:

```
conda activate pgdb
```

2.2 Install neo4j & apoc library

- Mac OS Installation:

```
brew install neo4j
```

- Install Apoc
- Move Apoc.jar into the neo4j plugins folder
- Add permissions to the neo4j.conf:

```
dbms.security.procedures.whitelist=apoc.export.*
apoc.import.file.use_neo4j_config=false
apoc.export.file.enabled=true
```

2.3 Install **postgreSQL**

- Mac OS Installation:

```
brew install postgresql
```

- Start the postgresql database:

```
brew services start postgresql
```

- Create superuser postgres if not exists:

```
createuser postgres -s
```

3 Load **STRING** data onto **PostgreSQL**

1. First we need to download our STRING Database subsets:

- STRING_Mouse.zip
- STRING_Human.zip

2. Creating an empty PostgreSQL database "string":

```
psql -U postgres -c "DROP DATABASE string;"  
psql -U postgres -c "CREATE DATABASE string;"
```

3. Loading dump & schema(mouse) onto the "string" database:

```
psql -U postgres string < dump.schema.psql  
psql -U postgres string < dump.mouse.psql
```

4 Build a Graph Database (**PostgreSQL -> Neo4j**)

To build now our database in neo4j we have to setup both databases

1. Starting the neo4j database:

```
neo4j start
```

2. Build the graph database with build_graph_db.py:

```
python build_graph_db.py --credentials tests/credentials.test.yml  
--species_name "mus_musculus"  
--combined_score_threshold 750
```

Arguments for build_graph_db.py:

```
--credentials CREDENTIALS
    Path to the credentials YAML file that will be used
    (default: credentials.yml)
--species_name SPECIES_NAME
    Species name (default: Homo sapiens)
--protein_list PROTEIN_LIST
    Path to the file containing protein Ensembl IDs
    (default: None)
--combined_score_threshold COMBINED_SCORE_THRESHOLD
    Threshold above which the associations between
    proteins will be considered (default: None)
--skip_actions SKIP_ACTIONS
    Do not add protein - protein actions to the resulting
    graph database (default: True)
--skip_drugs SKIP_DRUGS
    Do not add drugs to the resulting graph database
    (default: False)
--skip_compounds SKIP_COMPOUNDS
    Do not add compounds to the resulting graph database
    (default: False)
--skip_diseases SKIP_DISEASES
    Do not add diseases to the resulting graph database
    (default: False)
--keep_old_database KEEP_OLD_DATABASE
    Do not overwrite the existing Neo4j graph database
    (default: False)
```

Thank You ...