PathwayForte Documentation

Release 0.0.1-dev

Sarah Mubeen and Daniel Domingo-Fernández

May 29, 2019

COMMAND LINE INTERFACE :

A Python package for benchmarking pathway databases with functional enrichment and prediction methods tasks.

CHAPTER

ONE

COMMAND LINE INTERFACE

PathwayForte commands.

1.1 pathway_forte

Run PathwayForte.

```
pathway_forte [OPTIONS] COMMAND [ARGS]...
```

1.1.1 datasets

List the available cancer datasets.

```
pathway_forte datasets [OPTIONS]
```

1.1.2 export

Generate gene set files using ComPath.

```
pathway_forte export [OPTIONS]
```

1.1.3 fcs

List of FCS Analyses.

```
pathway_forte fcs [OPTIONS] COMMAND [ARGS]...
```

gsea

Run GSEA on TCGA data.

```
pathway_forte fcs gsea [OPTIONS]
```

Options

-d, --data <data>
Name of the cancer dataset from TCGA [required]

-p, --permutations <permutations>
Number of permutations [default: 100]

gsea-msig

Run GSEA on TCGA data using MSigDB gene sets.

```
pathway_forte fcs gsea-msig [OPTIONS]
```

Options

-d, --data <data>
Name of the cancer dataset from TCGA [required]

ssgsea

Run ssGSEA on TCGA data.

```
pathway_forte fcs ssgsea [OPTIONS]
```

Options

-d, --data <data>
Name of the cancer dataset from TCGA [required]

1.1.4 ora

Perform ORA analysis.

```
pathway_forte ora [OPTIONS] COMMAND [ARGS]...
```

hypergeometric

Performs one-tailed hypergeometric test enrichment.

```
pathway_forte ora hypergeometric [OPTIONS]
```

Options

-d, --genesets <genesets>
Path to GMT file [required]

-s, --fold-changes <fold_changes>
Path to fold changes file [required]

--no-threshold
Do not apply threshold

-o, --output <output>
Optional path for output JSON file

1.1.5 prediction

List of Prediction Methods.

```
pathway_forte prediction [OPTIONS] COMMAND [ARGS]...
```

binary

Train elastic net for binary prediction.

```
pathway_forte prediction binary [OPTIONS]
```

Options

-d, --data <data>
Name of the cancer dataset from TCGA [required]

--outer-cv <outer_cv>
Number of splits in outer cross-validation [default: 10]

--inner-cv <inner_cv>
Number of splits in inner cross-validation [default: 10]

-i, --max_iterations <max_iterations>
Number of max iterations to converge [default: 1000]

--turn-off-warnings
Turns off warnings

subtype

Train subtype analysis.

```
pathway_forte prediction subtype [OPTIONS]
```

Options

-d, --ssgsea <ssgsea>
Path to ssGSEA file [required]

-s, --subtypes <subtypes>
Path to the subtypes file [required]

--outer-cv <outer_cv>
Number of splits in outer cross-validation [default: 10]

--inner-cv <inner_cv>
Number of splits in inner cross-validation [default: 10]

--chain-pca

--explained-variance <explained_variance>
Explained variance [default: 0.95]

--turn-off-warnings
Turns off warnings

survival

Train survival model.

```
pathway_forte prediction survival [OPTIONS]
```

Options

-d, --data <data>
Name of dataset [required]

--outer-cv <outer_cv>
Number of splits in outer cross-validation [default: 10]

--inner-cv <inner_cv>
Number of splits in inner cross-validation [default: 10]

--turn-off-warnings
Turns off warnings

test-stability-prediction

```
pathway_forte prediction test-stability-prediction [OPTIONS]
```

Options

-s, --ssgsea-scores-path <ssgsea_scores_path>
ssGSEA scores file [required]

-p, --phenotypes-path <phenotypes_path>
Path to the phenotypes file [required]

--outer-cv <outer_cv>
Number of splits in outer cross-validation [default: 10]

--inner-cv <inner_cv>
Number of splits in inner cross-validation [default: 10]

-i, --max_iterations <max_iterations>
Number of max iterations to converge [default: 1000]

--turn-off-warnings
Turns off warnings

CHAPTER

TWO

PIPELINE

Pipelines from Pathway Forte.

CHAPTER

THREE

CONSTANTS

This module contains all the constants used in the PathwayForte repo.

pathway_forte.constants.**BIO2BEL_DATA_DIR = '/home/docs/.bio2bel/pathwayforte'**
Cancer Data Sets

pathway_forte.constants.**make_classifier_results_directory**()
Ensure that the result folder exists.

pathway_forte.constants.**MSIG_GSEA = '/home/docs/checkouts/readthedocs.org/user_builds/pathw**
Output files with results for GSEA

pathway_forte.constants.**make_gsea_export_directories**()
Ensure that gsea export directories exist.

pathway_forte.constants.**MSIG_SSGSEA = '/home/docs/checkouts/readthedocs.org/user_builds/pat**
Pickles with results for ssGSEA

pathway_forte.constants.**make_ssgsea_export_directories**()
Ensure that gsea export directories exist.

pathway_forte.constants.**check_gmt_files**()
Check if GMT files exist and returns GMT files as constant variables.

pathway_forte.constants.**GENESET_COLUMN_NAMES = {'kegg': 'KEGG Geneset', 'reactome': 'Reac**
Columns to read to perform ORA analysis.

CHAPTER

FOUR

OVER REPRESENTATION METHODS

CHAPTER

FIVE

FUNCTIONAL CLASS SCORE

Functional Class Scoring Methods such as GSEA.

CHAPTER

SIX

PATHWAY TOPOLOGY METHODS

This module contain the topology-based topology methods implemented in PathwayForte used R wrappers and are located outside the main Python package in its corresponding R folder https://github.com/pathwayforte/results/tree/master/R.

CHAPTER

SEVEN

UTILS

Complementary methods for prediction analysis.

pathway_forte.prediction.**utils**
alias of *pathway_forte.prediction.utils*

CHAPTER

EIGHT

BINARY PREDICTION

Prediction of binary classes such as tumor vs. normal patients.

pathway_forte.prediction.**binary**
alias of *pathway_forte.prediction.binary*

CHAPTER

NINE

MULTI-CLASS PREDICTION

Prediction of multi-class labels such as tumor subtypes.

pathway_forte.prediction.**multiclass**
alias of *pathway_forte.prediction.multiclass*

CHAPTER

TEN

SURVIVAL PREDICTION

Prediction of survival based on clinical and pathway patient data.

pathway_forte.prediction.**survival**
alias of *pathway_forte.prediction.survival*

CHAPTER

ELEVEN

UTILS

Complementary methods for prediction analysis.

pathway_forte.prediction.**utils**
alias of *pathway_forte.prediction.utils*

CHAPTER

TWELVE

MAPPINGS METHODS

Methods related to ComPath mappings.

pathway_forte.**mappings**
alias of *pathway_forte.mappings*

CHAPTER

THIRTEEN

INSTALLATION |PYPI_VERSION| |PYTHON_VERSIONS| |PYPI_LICENSE|

`pathway_forte` can be installed from PyPI with the following command in your terminal:

```
$ python3 -m pip install pathway_forte
```

The latest code can be installed from GitHub with:

```
$ python3 -m pip install git+https://github.com/pathwayforte/pathway-forte.git
```

For developers, the code can be installed with:

```
$ git clone https://github.com/pathwayforte/pathway-forte.git
$ cd pathway-forte
$ python3 -m pip install -e .
```

CHAPTER

FOURTEEN

MAIN COMMANDS

The table below lists the main commands of PathwayForte.

Command	Action
datasets	Lists of Cancer Datasets
export	Export Gene Sets using ComPath
ora	List of ORA Analyses
fcs	List of FCS Analyses
prediction	List of Prediction Methods

CHAPTER

FIFTEEN

FUNCTIONAL ENRICHMENT METHODS

- **ora**. Lists Over-Representation Analyses (e.g., one-tailed hyper-geometric test).
- **fcs**. Lists Functional Class Score Analyses such as GSEA and ssGSEA using GSEAPy.

CHAPTER

SIXTEEN

PREDICTION METHODS

`pathway_forte` enables three classification methods (i.e., binary classification, training SVMs for multi-classification tasks, or survival analysis) using individualized pathway activity scores. The scores can be calculated from any pathway with a variety of tools (see[1]) using any pathway database that enables to export its gene sets.

- **binary**. Trains an elastic net model for a binary classification task (e.g., tumor vs. normal patients). The training is conducted using a nested cross validation approach (the number of cross validation in both loops can be selected). The model used can be easily changed since most of the models in scikit-learn (the machine learning library used by this package) required the same input.
- **subtype**. Trains a SVM model for a multi-class classification task (e.g., predict tumor subtypes). The training is conducted using a nested cross validation approach (the number of cross validation in both loops can be selected). Similarly as the previous classification task, other models can quickly be implemented.
- **survival**. Trains a Cox's proportional hazard's model with elastic net penalty. The training is conducted using a nested cross validation approach with a grid search in the inner loop. This analysis requires pathway activity scores, patient classes and lifetime patient information.

[1] Lim, S., *et al.* (2018). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings in bioinformatics*, bby125.

CHAPTER

SEVENTEEN

OTHER

- **export**. Export GMT files with current gene sets for the pathway databases included in ComPath[2].
- **datasets**. Lists the TCGA data sets[3] that are ready to run in `pathway_forte`.

[2] Domingo-Fernández, D., *et al.* (2018). ComPath: An ecosystem for exploring, analyzing, and curating mappings across pathway databases. *npj Syst Biol Appl.*, 4(1):43.

[3] Weinstein, J. N., *et al.* (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10), 1113.

CHAPTER

EIGHTEEN

REFERENCES

CHAPTER

NINETEEN

INDICES AND TABLES

- genindex
- modindex
- search

PYTHON MODULE INDEX

p

INDEX

Symbols

B

C

G

M

P

S

U