alignESS

A program for Enzymatic Step Sequence (ESS) alignment using the Dynamic Programming (DP) Needleman–Wunsch algorithm.

The program can perform the following alignments:

```
pair                ESS command line pairwise comparisson using DP
dbalign             ESSs database(s) alignment using DP
multi               ESSs multiple alignment using GA
```

The pairwise alignments are generated with the DP algorithm, and the multiple ESS alignement is generated using a Genetic Algorithm (GA).

Dependencies

The program runs in Python 3. It runs properly in an Anaconda base installation! (that includes numpy, sqlite, matplotlib, cython and pytest).

- Pair-wise alignment (simple pairwise or database alignment):
  - cython
  - numpy
  - sqlite3
  - matplotlib [optional] only for ecs_entropy.py script
  - pytest [optional] for running test suite
- Multiple alignment:
  - C boost library: only for compiling the alignment algorithm*

(*) This repo contains a compiled (linux 64 bit) copy of the multiple alignment algorithm that may work fine in the majority of linux systems. The source code of this part of the program is not yet included.

For convenience a conda environment can be build to fullfill all dependencies with the conda .yml file in this repo

```
$ conda create env -f conda_env.yml
```

Installing

For now, clone this git repository and use the alginESS.py script.

Usage.

Enzimatic Step Sequences (ESS)

The ESSs represent lineal consecutive steps of enzymatic activities. These steps are represented using the Enzyme Commission (EC) numbers that describe catalytic function. Thus, the ESS are a form of functional representation of a metabolic processes. In this case, only the first 3 numbers are considered,

because in general seems to be more informative and tends to be less prone to annotation issues.

The ESS can be obtained in any way, but must have the following form:

1. Each enzime is represented by a 3 digit EC number: 3.1.4.
   Invalid numbers (i.e. inexistent in KEGG database) will rise error.
2. The sequence (ESS) is constructed joining the enzimes using colons (:).
   In this form (a 3 step ESS): 3.2.4:1.6.12:4.4.1
3. An undetermined enzymatic setp must be specified in this form: 9.9.9

The program accepts ESS written on the terminal, in a text file or in a Sqlite database. Examples of each type of file can be found in the test folder in this repo.

Pair-wise alignment.

```
usage: alignESS.py pair [-h] [-l] ess1 ess2
positional arguments:
  ess1            ESS (3 levels EC numbers). Colon separated.
                  (1.2.3:3.5.-:...:9.9.9)
  ess2            ESS (3 levels EC numbers). Colon separated.
                  (1.2.3:3.5.-:...:9.9.9)

optional arguments:
  -h, --help      show this help message and exit
  -l, --localize  The alignment is trimmed to the coverage of the shortest ESS
                  and the score is then calculated to the trimmed alignment.
                  This method allows to find 'local-like' alignments between
                  ESS of different size
```

ess1 and ess2 must be ESS written in the command line.

Example

```
(ess-env) $ python3 alignESS.py pair 2.7.1:5.3.1:5.3.1:2.7.1:4.1.2:1.2.1 5.3.1:5.3.1:4.2.1
<r 2.7.1:5.3.1:5.3.1:2.7.1:4.1.2:1.2.1 5.3.1:5.3.1:4.2.1
ess1:   2.7.1:5.3.1:5.3.1:2.7.1:4.1.2:1.2.1
ess2:   -.-.-:5.3.1:5.3.1:-.-.-:4.2.1:-.-.-
score = 0.566987156867981
>>> Done!!! <<<
:D, see you soon.
```

Pair-wise database alginment.

```
usage: alignESS.py dbalign [-h] [-db2 ESSDB2] [-o OUTFILE] [-t THRESHOLD]
                           [-nproc NPROC] [-l] [-align]
                           essdb1
```

```
positional arguments:
  essdb1                ESSs database 1. Sqlite3 file with nrseqs table or
                        text file with one ESS in each line. If the essdb2
                        argument is not specified, the program performs the
                        all-vs-all alignment in essdb1. This argument also can
                        be a single ESS, in this case the -db2 argument is
                        necessary

optional arguments:
  -h, --help            show this help message and exit
  -db2 ESSDB2, --essdb2 ESSDB2
                        ESSs databse 2. Sqlite3 file with nrseqs table or text
                        file with one ESS in each line
  -o OUTFILE, --outfile OUTFILE
                        Outfile name to report scores. By default the file
                        only contains the id of the ESSs and the score of the
                        alignment. If argument '-align' is set, then the file
                        contains the alignmed ESSs
  -t THRESHOLD, --threshold THRESHOLD
                        Threshold score to filter results in the range 0-1
                        [0.3]. If the threshold is high (>0.6) and the
                        databases are large, results may saturate the RAM
                        memmory, beware!
  -nproc NPROC          Number of processes to execute analysis [2]. It can be
                        created more processes than cores in the the
                        processor, so the speedup of the analysis depends on
                        the number of cores available
  -l, --localize        The alignment is trimmed to the coverage of the
                        shortest ESS and the score is then calculated to the
                        trimmed alignment. This method allows to find 'local-
                        like' alignments between ESS of different size
  -align                If set, the outputfile contains the alignment of each
                        ESS pair bellow the threshold. Beware, if the
                        databases are large and the threshold high, the file
                        may be huge or the RAM memmory colapse.
```

essdb1, essdb2 can be text files or sqlite databases (with specific format --coming soon--). Examples can be found in test folder

Multiple alignment

```
usage: alignESS.py multi [-h] [-o OUTPUTFILE] [-p FILENAME] FILENAME

positional arguments:
  FILENAME              ESSs file. Each line must contain an ESS name and the
                        ESS separeated by a TAB. Accepts commentaries with '#'
```

```
optional arguments:
  -h, --help            show this help message and exit
  -o OUTPUTFILE, --multiout OUTPUTFILE
                        Multiple alignment outputfile
  -p FILENAME, --pcomp FILENAME
                        If spificied, stores the pairwise comparisson of the
                        ESSs in the ESSs file
```

File name must be a text file. An example can be found in test folder.

More info coming soon...!!!

Papers.

The programs presented here were used all or in parts in the following papers.

1. Comparison of Metabolic Pathways in Escherichia coli by Using Genetic Algorithms P Ortegon, AC Poot-Hernández, E Perez-Rueda, K Rodriguez-Vazquez Computational and structural biotechnology journal 13, 277-285. 2015.
2. The alignment of enzymatic steps reveals similar metabolic pathways and probable recruitment events in Gammaproteobacteria AC Poot-Hernandez, K Rodriguez-Vazquez, E Perez-Rueda BMC genomics 16 (1), 957. 2015.
3. Identification of functional signatures in the metabolism of the three cellular domains of life P Escobar-Turriza, R Hernandez-Guerrero, AC Poot-Hernández, ... PloS one 14 (5). 2019.