DAMPSA: Domain-aware Multiple Protein Sequence Aligner

Roll number: 820

DAMPSA is a multiple sequence analysis (MSA) tool that utilises protein domain annotations as biological constraints.

DAMPSA generates MSA outputs where protein domains are anchored. This means:

1. Domain and linker segments of the protein will not be mixed by random similarity.
2. Potentially, *distant* sequences with *similar* structures can be aligned.

Installation

```
conda create -n dampsa python=3.9
conda activate dampsa
```

In project folder (`cd DAMPSA`),

1. Install python libraries through `pip`.

```
pip install -r py_requirements.txt
```

2. Install command line programs through `bioconda`.

```
conda install --file=cmd_requirements.txt -c bioconda
```

3. Prepare the **local Pfam database** for domain annotation.
    - The following code download Pfam database files and then use `hmmpress` to index them for efficient processing.
    - *Note*: The ready database will takes ~**3.4 GB** space.
    - *Note*: The links are updated on 30/05/2022. Check Pfam if any link fails.

```
mkdir data/Pfam_scan_db
cd data/Pfam_scan_db

wget http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hm
wget http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hm
wget http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/active_si
gunzip *.gz

hmmpress Pfam-A.hmm
```

:) You are now ready to run DAMPSA main pipeline.

Further notes:

- To run visualisation scripts in R (not the main pipeline), you need the following packages.

  ▼ click here

  ```
          tidyverse
          msa
          ggmsa
          RColorBrewer
          Biostrings
          stringr
          getopt
  ```

- DAMPSA is developed and tested on `MacOS 12.1`, Python `3.9.12`, and R `4.1.2`.

Getting started (or see a walkthrough tutorial here)

In project folder (`cd DAMPSA`),

```
python bin/main.py -h
```

▼ parameter descriptions

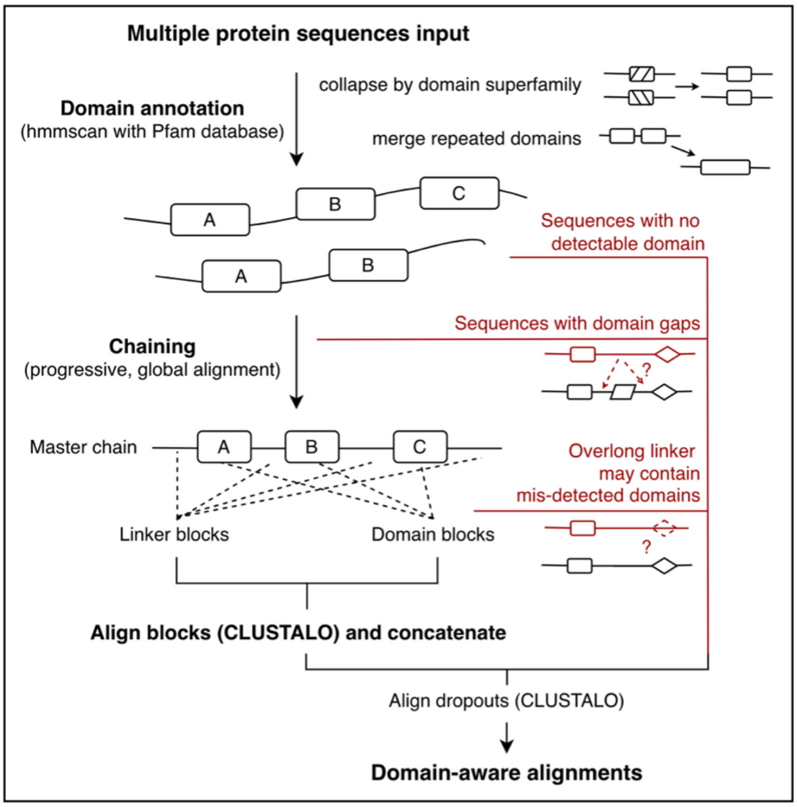
```
 usage: main.py [-h] [--input INPUT] [--output OUTPUT] [--domain-out D
       [--focus-clan FOCUS_CLAN] [--cache-dom CACHE_DOM] [--domain-app
       [--n-thread N_THREAD]
DAMPSA input arguments.


optional arguments:
-h, --help            show this help message and exit
--input INPUT         Path to the input .fasta file.
--output OUTPUT       Path to the alignment .fasta file output.
--domain-out DOMAIN_OUT
Path to output domain annotation results.
--refine-edge         Refine alignments at the edge between domain and
--no-check-linker     Not to check if the linker is too long - likely
--focus-clan FOCUS_CLAN
Only consider the specified Clan IDs (domain superfamily) - comma sepa
--cache-dom CACHE_DOM
Skip hmmscan, use supplied filepath to cached domain table (TSV-like).
--domain-app DOMAIN_APP
Aligner for domain segments (clustalo or mafft), default clustalo.
--linker-app LINKER_APP
Aligner for linker segments (clustalo or mafft), default clustalo.

--log                 Store log file in the same folder as the alignme
```

```
  --n-thread N_THREAD    Thread number for running hmmscan (domain annot
```

Run an example command (align the RASSF family)

```
python bin/main.py --input tutorial/RASSF/raw.fasta \
--output tutorial/RASSF/aligned.fasta \
--domain-out tutorial/RASSF/domain.txt \
--no-check-linker --log
```

Architecture


Multiple protein sequences input
Domain annotation
(hmmscan with Pfam database)
collapse by domain superfamily
merge repeated domains
A
B
C
A
B
Sequences with no
detectable domain
Sequences with domain gaps
?
Chaining
(progressive, global alignment)
Master chain
A
B
C
Overlong linker
may contain
mis-detected domains
?
Linker blocks
Domain blocks
Align blocks (CLUSTALO) and concatenate
Align dropouts (CLUSTALO)
Domain-aware alignments

The DAMPSA pipeline involves three stages:

1. Annotating domains with `hmmscan` and Pfam database.
2. Chaining domain sequence using progressive global alignment implemented here.
3. Align blocks defined by the chain, using Clustal-Omega (CLUSTALO). These blocks are concantenated to generate full alignment.

In three cases, sequences cannot be considered by DAMPSA. They are dropped out and aligned finally using *sequence-to-profile* method in CLUSTALO.

- Sequences with no domain detected.
- Domain sequences that are gapped (e.g. -A-C- vs. -A-B-C-, the first sequence will be dropped out)
- Sequences with overlong linker which may indicate domain misdetection.

All dropout cases are logged. Please check `log.txt` in DAMPSA output.

API documentation

See here.