SDA v 0.10.0

Flores Rodriguez Ramon
ramon.flores.r@outlook.com

February, 2016

1 Before you start

Scan Domain Architecture (SDA) is an script intended to help you with the analysis of protein functional domain architecture and evolution.

2 Feedback

Any comments, suggestions, bugs or feedback please contact us at: *ramon.flores.r@outlook.com.*

3 Requirements

Before running SDA you will need to download or install:

- Perl v5.18.2 or later version.
- Perl SVG module for graphics.
- The Pfam-A.hmm database[1].
- The HMMER[2] suite for hmm profiles. SDA was developed and tested under HMMER v3.1b1, but current version v3.1b2 should work as well.

SDA makes some searches inside the PFAM database to annotate the input sequences. Here, we use the PFAM version 29 (latest). After installing the HMMER suite please format your database for hmmscan searches.[3]

To search a list of proteins with PFAM annotation in order to study their functional domain architecture you will need: (1) An input file (or PFAM list) of the sequence(s) you want to find and must be amino acid sequences. This program doesn't work with nucleic acid sequences; (2) if you want to compare your sequence PFAM annotation to other proteins, an annotation list in the **trinotate** format can be used.

[1] `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam29.0/`

[2] `http://hmmer.org/`

[3] To format the Pfam-A.hmm file, please execute `hmmpress` *Pfam-A.hmm* in the terminal. This will create four binary files .h3{fimp}

It's very important that your annotation file has the following requirements: the header for the swissprot field must be *sprot_blastp_hit* (capitals or small letter, doesn't matter). Also, the file must have at least columns with the following headers: *gene_id*, *pfam*, and *transcript* (if the annotation file comes from a transcriptome); absence of some of these headers will result in a suddenly termination of this tool.

SDA allows you to search in an annotation file like Trinotate transcriptome results, for proteins with similar architectures or new combinations of the domains of PFAM domains of your interest.

SDA works with annotation files from genomes or transcriptomes as long as they have the required format.

4 Usage

SDA creates up to three files depending on the results found. Please, verify next section for details.

The way you have to use SDA is as follows:

```
./SDA -OPTION <PARAMETER> ...
```

The user's interest sequence might be a FASTA transcriptome sequence in a file, or a PFAM architecture in a file or as a list.

The PFAM list may have two forms, (1) as a file with one architecture per line and each PFAM joined by a comma without spaces; each PFAM can only have the letters *PF* and five digits. (2) As a parameter list, in which case this input is interpreted just as one architecture; you can list up to three PFAMS this way, each one joined by a comma and without spaces.

The all possible options and their description are the next:

-a *accuracy* Defines the accuracy for each PFAM in the hmmscan summary table. It's a value between 0 and 1 (default, 0.85).

-e *VALUE* E-value threshold for the hmmscan search and result filtering (default, 1e-10). Must be positive. The format input might be as an integer or `_[._]e[-]__`, where the values inside square brackets are optional.

-f *FILE* FASTA file with amino acid sequence(s).

-g *Optional*. Use it if your annotation file is a gene file. The expected annotation file input is a transcriptome.

-h –**help** Display this help menu.

-l *PFAMS* Use it if your input is one or more PFAMs architecture(s).

-p *FILE* *Compulsory.* Indicate the location of your Pfam-A.hmm file.

-t *FILE* Tab-delimited file with PFAM annotation.

-v Shows the SDA version.

5 Output Files

SDA generates up to three different files in directory called *SDA.Results* in your parent working directory (PWD) (∼/).

If you feed SDA with a FASTA file, inside the results directory, SDA will create another directory with the name of your FASTA file and inside, you'll find the results. That means you can work with different FASTA files and you'll find your results in each different directory.

If you feed SDA with a PFAM file or list, the result will be saved inside the *SDA.Results* directory if there is at least one coincidence or similarity between the PFAMs input and the annotation file.

The output files and their content is described above:

.table Contains the hmmscan results for the `hmmscan` fasta file. Please, check the `hmmscan` documentation for further details.

.pdf This file contains a graphical representation of your result for each sequence in the FASTA file. It draws the domains found for the sequence(s) and it's PFAM name.

.out It's created if SDA finds at least one coincidence or similarity between each FASTA sequence (or PFAM input) and the annotation file. Contains the gene_id or transcript, location, e-value, and description of the match. At the end of this file you'll find a summary table with all the PFAM architectures found and it's relative frequency.

6 Criteria

The similarity of the PFAMs for the `.out` file depends on the PFAM cardinality of each sequence in the FASTA file or PFAM list, and the number of matches. This way, a similarity will be considered as a result if at least a third of its domains match with an element of the annotation list. In the .table file SDA keeps just the hmmscan results that has at least a score greater than 52 and an accuracy greater or equal than .90.

7 Working with SDA

7.1 SDA with a FASTA file and a genome annotation file

This is an example working with a FASTA sequence file and a genome annotation file. We run SDA as:

```
./SDA -p ~/Downloads/Pfam-A.hmm -f new.fasta
      -a trinotate_annotation_report.xls -g
```

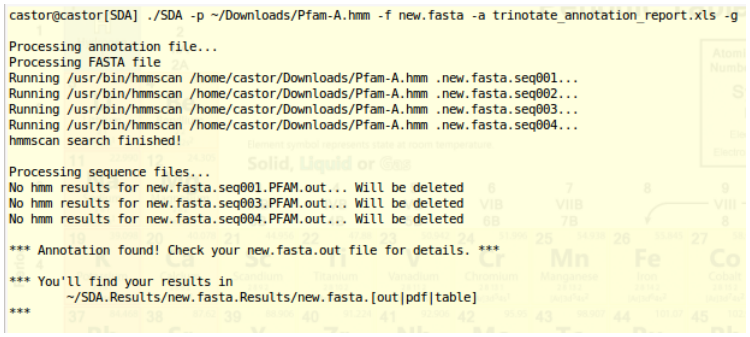

Figure 1: SDA show us information while it's running.

Because SDA found coincidences or similarities, there will be three files in the ~/SDA.Results/new.fasta.Results directory. The file .out keeps the coincidences or similarities (figure 2).

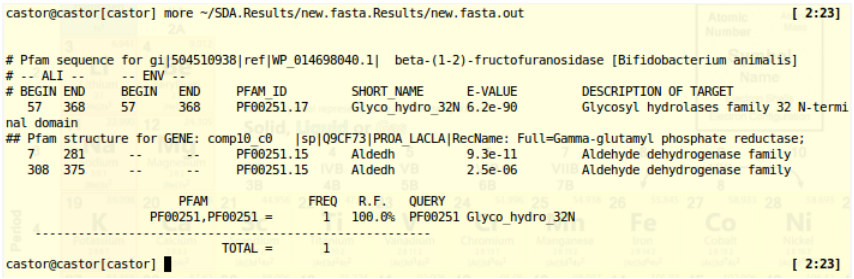

Figure 2: Coincidences between FASTA input and annotation file.

The .table file keeps the PFAMs for the FASTA file, this is created with `hmmscan` (figure 3).

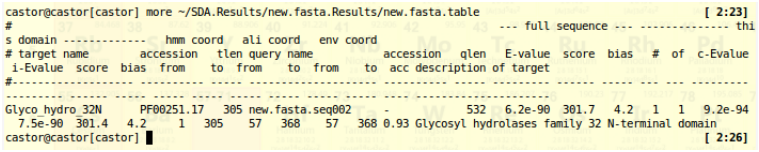

Figure 3: Content of the .table file.

Figure 4: Graphics for each PFAM in the .table file.

The .pdf file contents a graphic for each FASTA sequence in the FASTA file (figure 4).

7.2 SDA with a PFAM list and a transcriptome file

With a PFAM list[4], we run SDA as:

```
./SDA -p ~/Downloads/Pfam-A.hmm -l PF00899,PF00410 -a trinotateT.xls
```

```
castor@castor[SDA] ./SDA -p ~/Downloads/Pfam-A.hmm -l PF00899,PF00410 -a trinotateT2.test

Processing annotation file...
Processing PFAM input...

*** Annotation found! Check your results.out file for details. ***

*** You'll find your results in
        ~/SDA.Results/results.out
***
```

Figure 5: SDA running with a PFAM list in a transcriptome file.

In this way, you only obtain a single file, the .out file:

```
castor@castor[castor] more SDA.Results/results.out                                                    [ 2:37]

## Pfam arquitecture input: PF00899,PF00410
## Pfam structure for comp0000015_c0_seq1|sp|Q03799|RT08_YEAST|RecName: Full=37S ribosomal protein S8, mitochondrial;
   6    157      --       --      PF00410.14      Ribosomal_S8    6.7e-18         Ribosomal protein S8
## Pfam structure for comp0000963_c0_seq1|sp|Q96W54|RS22_CANAL|RecName: Full=40S ribosomal protein S22;
   6    130      --       --      PF00410.14      Ribosomal_S8    4.9e-25         Ribosomal protein S8
## Pfam structure for comp0001283_c0_seq1|sp|Q96W54|RS22_CANAL|RecName: Full=40S ribosomal protein S22;
   6    130      --       --      PF00410.14      Ribosomal_S8    4.1e-26         Ribosomal protein S8

                    PFAM          FREQ   R.F.   QUERY
                        PF00410 =       3   100.0%  PF00899,PF00410
    ------------------------------------------------------
                              TOTAL =       3
castor@castor[castor] ▉                                                                               [ 2:37]
```

Figure 6: SDA running with a PFAM list in a transcriptome file.

[4] If you want, you could try with the PFAMS.data file instead of the PFAM list.