SDA v 0.11.x

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

September, 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 to Ramon Flores, `ramon.flores.r@outlook.com`.

3 Requirements

SDA requires:

- 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 or v3.1b2.
- For the GUI, you'll need Java 7 or later.

SDA makes some searches inside the PFAM database to annotate the input sequences. Here, we use the PFAM version 29 (latest, 30). After installing the HMMER suite please format your database for hmmscan searches.[3]

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

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

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

4 File format and headers

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 file) 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, a tab separated file list with PFAM annotations.

It's very important that your annotation file has the following requirements: the header for the Pfam column must begin with **pfam** (capitals or small letter, doesn't matter). If the file is a genome file, the header of the genes must be or contain **gene_id**, and for a transcriptome file must be or contain **transcript**; absence of some of these headers will result in a suddenly termination of this tool.

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

5 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|--acc` <*VALUE*>	Defines the accuracy for each PFAM in the hmmscan summary table. It must be value between 0 and 1 (default, 0.85).

`-b|--bitscore` <*VALUE*>	Defines the minimum bit score of the hmm table for each domain. Must be a positive value (default, 50).
`-c|--cpu` <*CPUs*>	Defines the number of CPUs used to create the HMM profile with `hmmscan`.
`-e|--eval` <*VALUE*>	E-value threshold for hmmscan search and result filtering (default, 1e-10). Must be positive with format input as an integer or in exponential notation.
`-f|--fasta` <*FILE*>	The path for the FASTA file with amino acid sequence(s).
`-g`	Use it if your annotation file is a gene file. The expected annotation file input is a transcriptome.
`-h|--help`	Displays this help menu.
`-l|--list` <*FILE*>	Use it if your input is one or more PFAMs architecture(s) in a file.
`-o|--output` <*PATH*>	Changes the output path directory (default, parent working directoty).
`-p|--pfam` <*FILE*>	*Compulsory.* Indicate the location of your Pfam-A.hmm file.
`-t|--tabular` <*FILE*>	Tab-delimited file with PFAM annotation for the genome/transcriptome file.
`-v|--version`	Shows the SDA version.

6 Output Files

SDA generates up to three different files in directory called *SDA.Results* in the output directory (∼/ as default).

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.

.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 with all the PFAM architectures found, similarity score and it's relative frequency.

If you feed SDA with a PFAM architectures file, 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:

7 Similarity score

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, the similarity scores is defined as

$$S_{SDA} = \frac{n}{\max\{l_1, l_2\}} \tag{1}$$

where n is the amount of domains shared by the both of sequences, and l_1, l_2 are the total amount of domains found for each sequence.

8 Working with SDA

There are many ways to use SDA, we shown some ways to use it.

8.1 Identifying domains architectures with SDA

The first use of SDA is to identify domains in a sequence of amino acids. The simplest way to do this is typing

`$ ./SDA -p` *<Pfam-A path>* `-f` *<fastaFile path>*

SDA shown the results of processing the file and at the end, it will indicate you where you will find the results.

SDA will create the files `.table` and `.pdf` for results (figure 1).

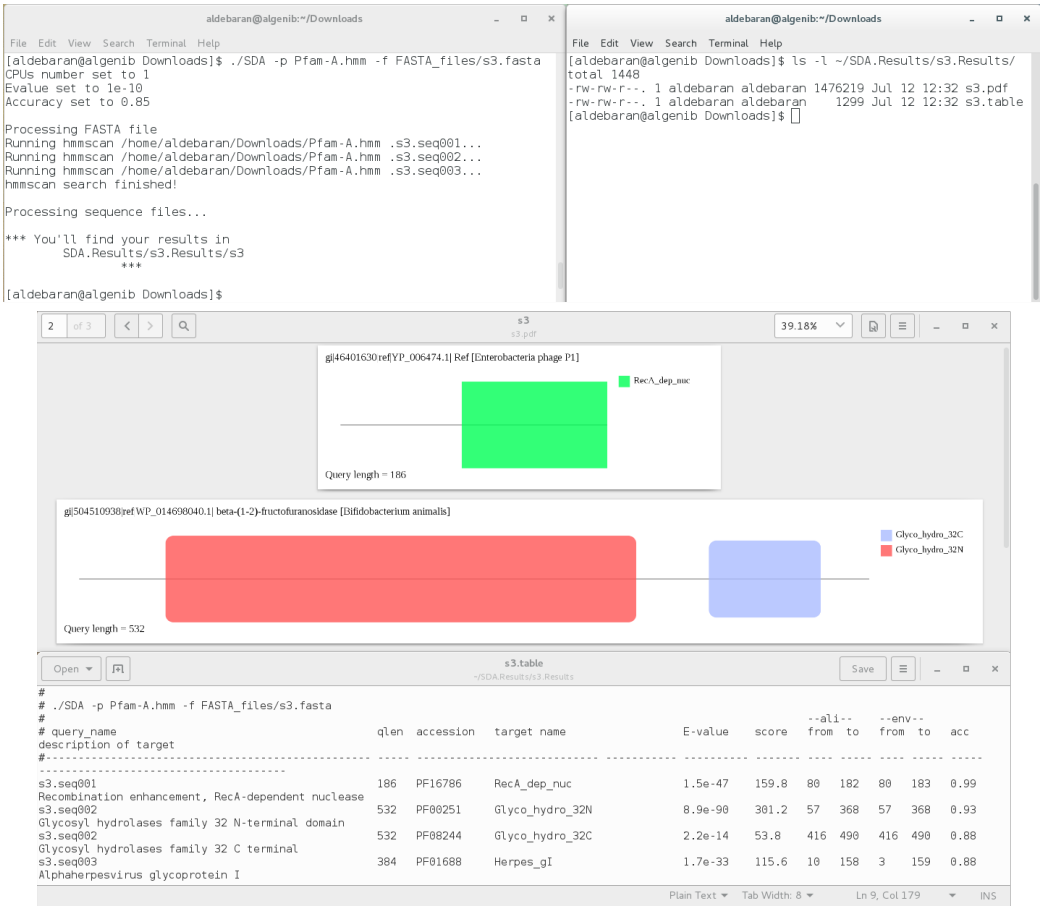

Figure 1: Results from SDA. Up, example of run. Down, .table and .pdf files results.

Additionaly, you can indicate the number of CPUs used to create the `hmmscan` profiles, the e-value and bitscore treshold, and the minimum accuracy of the results adding

`-c` *<CPUs>* `-a` *<accuracy>* `-e` *<e-value>* `-b` *<bitscore>*

8.2 SDA with a FASTA file and an annotation file

For searching domain architectures in a trinotate annotation file, you must add the option `-t` with the path to the annotation file, and the option `-g` if the annotation is a genome file (default is a transcriptome). Remember that the column names in the annotation file must have a format (section 4).

This way, the command will be

`$ ./SDA -p` *<Pfam-A path>* `-f` *<fastaFilePath>* `-t` *<annotFilePath>* `-g`

And again, you can specify the number of processors and accuracy of results. The example run and results are shown in figure 2.

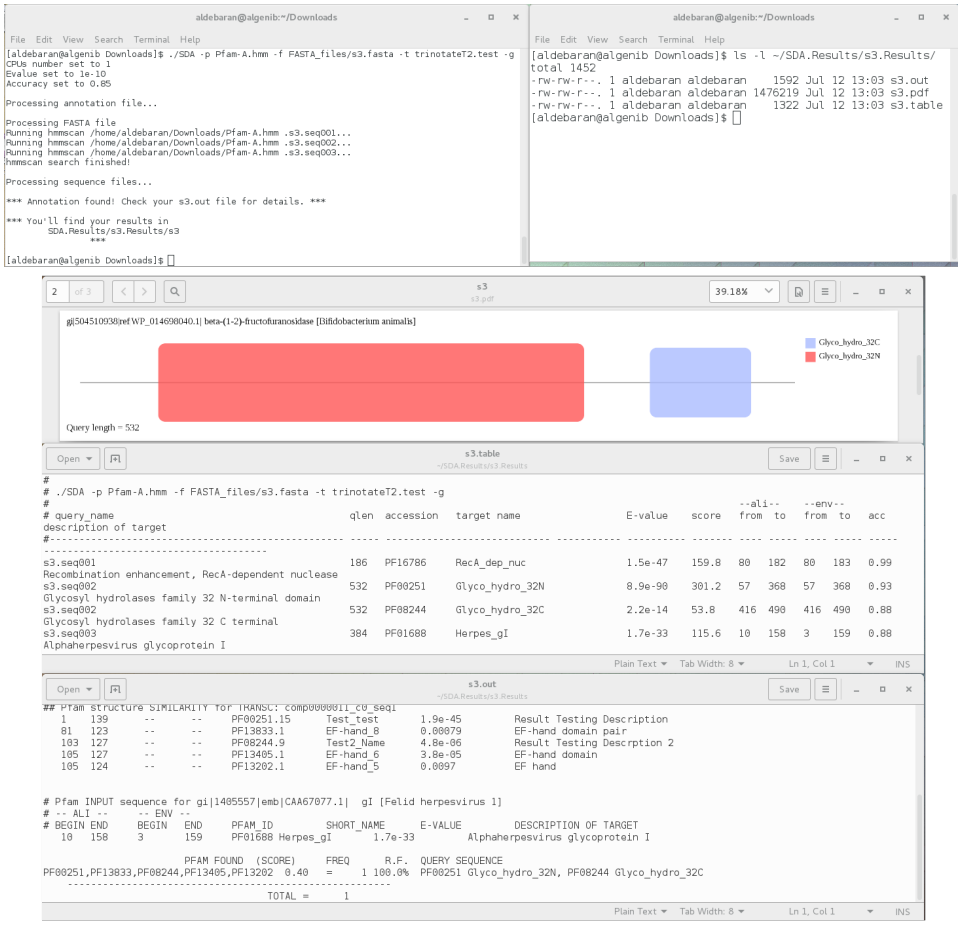

Figure 2: Results from SDA with annotation file. Up, example of run. Down, .table, .pdf and .out files results.

The `.out` file contains a table with information of the query domain and the

domain(s) found. At the end, there is a resume with the domain architectures found.

8.3 SDA with a PFAM architectures file and an annotation file

SDA can also be used as a PFAM architectures searcher. If you have an annotation file and you are looking for some domains architectures, you can put the architectures in a file `.txt` and look for them in the annotation file. To do this, type

`$ ./SDA -l` *<Pfam Architectures path>* `-t` *<annotationFile path>* `-g`

For this usage, the values of accuray, e-value and CPUs are nonsense because SDA will just look for architectures and will not build the HMM profile.

You will get just a file in the home directory called as the input file with extension `.out`. This file will contain the similarities found (if there are) and the distribution resume. If there are non results, no one file will be created.

Figure 3 shows an ejecution example and the results.

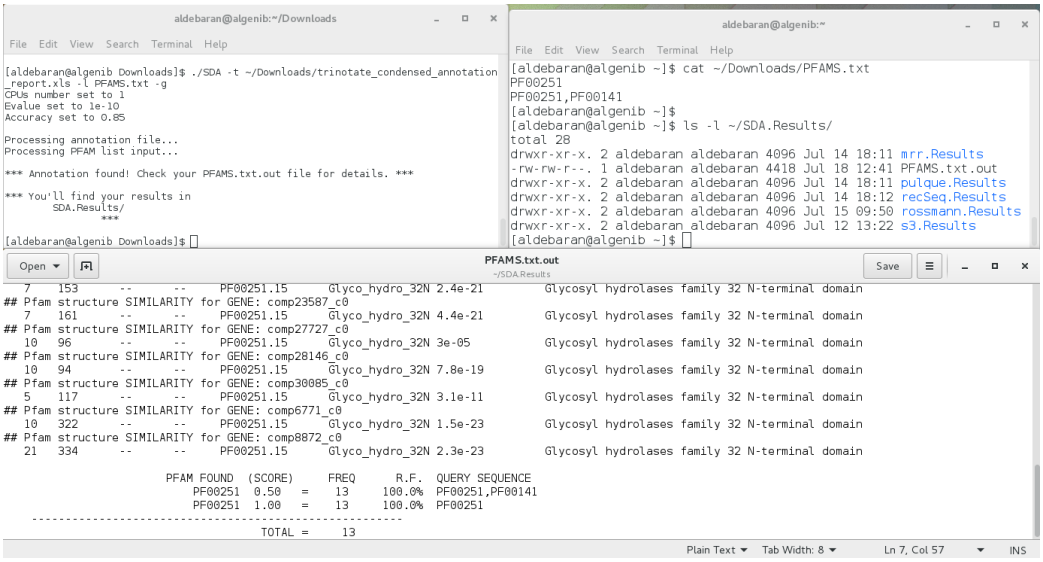

Figure 3: Results from SDA with PFAM architecture file and annotation file. Up left, example of run. Up right, PFAM file content and result file. Down, .out file content.

9 Using the GUI

There is a GUI for the users that dislike using shell, it is call `SDA.run` and was created with Java.

To execute it just open terminal in the folder where you have the files `SDA.run` and `SDA` (both of them must be in the same location) and type

```
$  ./SDA.run
```

The GUI is easy to use, you just have to select the files in your computer and click the button **start** to get results (figure 4).

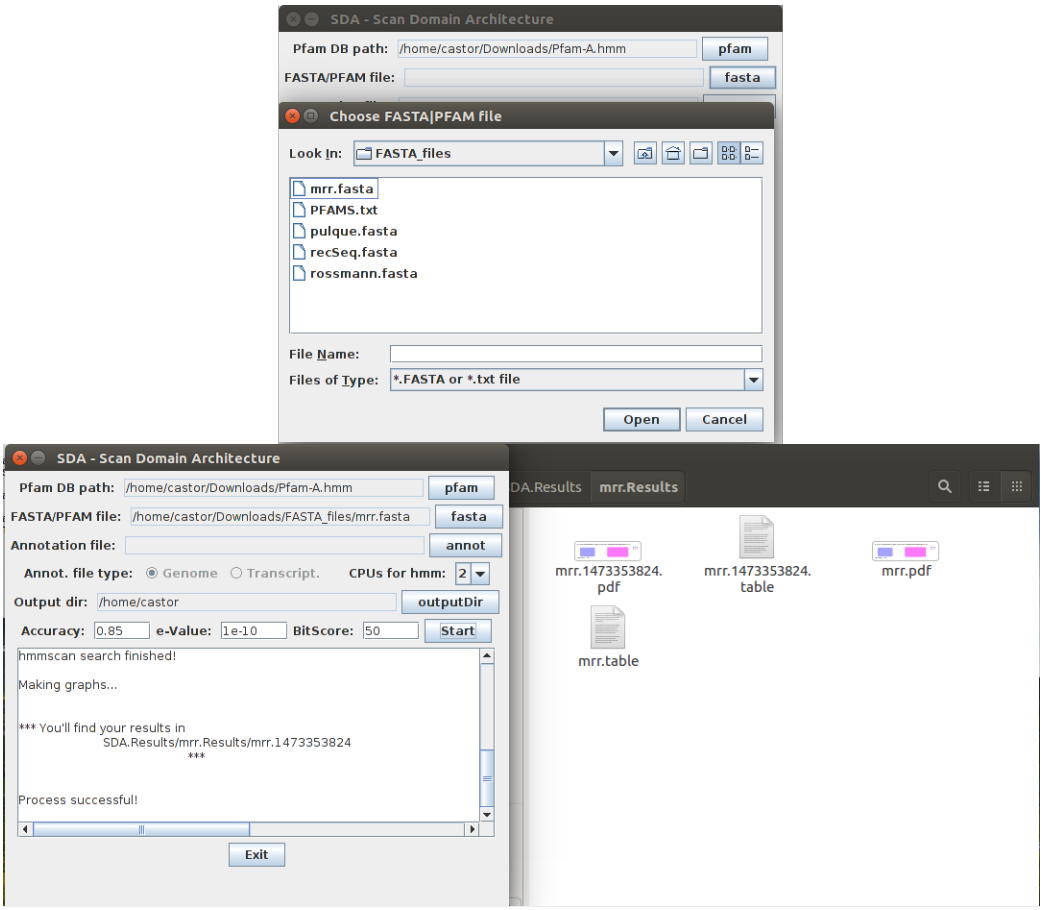

Figure 4: Example of use of SDA GUI.

If you want to use SDA GUI as a Pfam architecture searcher, select the PFAM file with the FASTA button, and the annotation file with the annot button. When the program finishes, SDA automatically will open the file with the results.