DomRates-Seq User manual

Abdulbaki Coban

September 5, 2024

Contents

1 Introduction

DomRates-Seq uses protein sequences to infer overall domain rearrangement events and track domain rearrangement events in a given phylogeny.

This manual is created to ease the usage of DomRates-Seq and tries to answer possible questions. Even though we tried to keep the software easy to use and bug free, if you encounter any issues or have any questions, please contact us at `domainworld@uni-muenster.de`

2 Installation

We try to keep the dependencies as little as possible. Current dependencies are:

- cmake
- compiler with c++20 and OpenMP support
- BioSeqDataLib (`https://zivgitlab.uni-muenster.de/domain-world/BioSeqDataLib`) (will be downloaded with cmake)
- boost (http://www.boost.org)

2.1 Installation with Git

```
git clone https://zivgitlab.uni-muenster.de/domain-world/domratesseq.git
cd domratesseq
cmake -S . -B build
cmake --build build
```

2.2 Manual installation without Git

The repository of DomRates-Seq can be downloaded manually from `https://zivgitlab.uni-muenster.de/domain-world/domratesseq`

Inside the source folder you can run:

```
cd domratesseq
cmake -S . -B build
cmake --build build
```

3 Input file formats

DomRates-Seq requires a pyhlogeny, protome files and domain annotation files.

- Phylogeny should be in newick format and bifurcating.

  ```
  (((A,(B,C)),(D,E)),Outgroup);
  ```

- Proteomes can be in FASTA
- Domain annotations can be in Pfam or InterProScan

3.1 Data preparation

It is important to note that file prefixes for proteome and domain files should be the same with the names on the phylogeny including the outgroup. By default domRates-Seq searches for .dom and .fasta files with the leaf names in the given folder.

To reduce the effect of multiple isoforms on the domain rearrangement analysis, we recommend to filter isoforms and keep only one of the isoforms. To do this you can use *isoformCleaner* (`https://zivgitlab.uni-muenster.de/domain-world/dw-helper`).

After the isoforms are filtered, domains should be annotated. Domain annotation can be easily performed using PfamScan (`https://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/`). The resulted annotation files can directly be used in domRates-Seq.

It is highly recommended that the quality of the proteome should be checked before starting the analysis. The DOGMA software (`https://zivgitlab.uni-muenster.de/domain-world/DOGMA`) offers a quick way to check the quality of the proteomes.

4 Running domRates-Seq

As usual, to run the installed software, you have to call the software from the build directory or add it to the PATH variable. To see the possible run options you can run:

```
./domRatesSeq ––help
```

or

```
./domRatesSeq –h
```

This will print out the help message with possible options:

```
DomRatesSeq version 1.0.0 (C) 2024 Abdulbaki Coban
This program comes with ABSOLUTELY NO WARRANTY;

Allowed options are displayed below.:

General options:
  -h [ --help ]                       Produces this help message
  -v [ --version ]                    Shows program version
  -t [ --tree ] arg                   The phylogenetic tree in newick format.
  -a [ --annotationDomains ] arg      A directory with all domain annotation
                                      files for species in the tree. Note that
                                      species names in tree and the regarding
                                      domain annotation file names have to be
                                      the same.
  -r [ --proteomesDirectory ] arg     A directory with all proteome files for
                                      species in the tree. Note that species
                                      names in tree and the regarding proteome
                                      file names have to be the same.
  -g [ --outgroup ] arg               The name of the outgroup as it is labeled
                                      in the tree.
  -e [ --ending ] arg (=.dom)         The filename extension of your domain
                                      annotation files.
  -f [ --prot_ending ] arg (=.fasta)  The filename extension of your domain
                                      annotation files.
  -m [ --amat ] arg                   Path to alignment matrix
  -o [ --out ] arg                    The output file. If no output file is
                                      chosen, results will be printed to
                                      console.
  -s [ --statistics ] arg             File to store additional information
                                      (such as number of events per node in the
                                      tree). Additional information is just
                                      stored in file, if specified.
  -n [ --node ] arg                   If two species names devided by ':' are
                                      provided, all arrangements involved in
                                      rearrangement events at the node
```

```
                                        representing the last common ancestor of
                                        both species will be listed in the
                                        statistics file.
                                        Just usable if statistics file (-s
                                        parameter) is set.Example for use: '-n
                                        Drosophila_melanogaster:Caenorhabditis_el
                                        egans'
  -d [ --detailed ]                     If this parameter is set, the output
                                        files also contain statistics about
                                        identical arrangements that have not
                                        changed. i.e. the arrangement stays
                                        conserved, and complex solutions, i.e.
                                        the rearrangement event leading to the
                                        new arrangement cannot be determined.
                                        (This can heavily increase file size.)
  -p [ --threads ] arg (=1)             Number of parallel threads to use for
                                        computation.
  -c [ --track ] arg (=0)               Option to track domain arrangements of a
                                        set of proteins
  -i [ --protIds ] arg                  Protein names to track
```

There are two possible usage of domRates-Seq:

- Inferring overall domain rearrangemet events.
- Tracking domain rearrangement events.

4.1 Inferring domain rearrangement events

To infer domain rearrangement events, you must run:

```
./domrRatesSeq −t path/to/tree.nwk −a path/to/domain/annotations −r
     path/to/proteome/annotations −m path/to/alignment/matrix −g
    Outgroup  −o path/to/outfile
```

This command creates a file containing the overall domain rearrangement events in the given phylogeny. To have a better understanding on the domain rearrangement events and to identify the events in each node you must include *-s / –statistics* and *-d / –detailed* options.

```
./domRatesSeq −t path/to/tree.nwk −a path/to/domain/annotations −r
    path/to/proteome/annotations −m path/to/alignment/matrix −g
    Outgroup  −o path/to/outfile −s path/to/statfile −d
```

When these options are selected, together with a main output file, two more files are generated. The first file created has the name as provided with the -s parameter while the second file has the same output name with the appendix ”epd”. Detailed description of the content of these files can be found at section 5.

4.2 Tracking domain rearrangement events

Using domRates-Seq, one can easily track domain rearrangement events in the evolutionary history of a protein or a protein set. To track the domain rearrangement events, one needs to add protein ids as they are saved in the proteome files. Then domRates-Seq should be run with *-c* option:

```
./domRatesSeq −t path/to/tree.nwk −a path/to/domain/annotations −r
    path/to/proteome/annotations −m path/to/alignment/matrix −g
    Outgroup  −o path/to/outfile −c 1 −i protId1,protId2,protId3
```

4.3 Options to use domRates-Seq

As overviewed in the previous sections, there are two main usage of domRates-Seq. We also provide some parameters that may be helpful to optimize domRates-Seq for user specific cases.

4.3.1 Parameter for file endings: -e and -f

By default, domRates-Seq searches for .fasta and .dom files in the provided folders for proteome and domain files. When the domain and proteome files are structured differently, one should define the endings accordingly using -e and -f parameters.

4.3.2 Parameter for alignment matrix: -m

domRates-Seq requires an alignment matrix to calculate alignment scores.

4.3.3 Parameter for node selection: -n

If you are interested in statistics about one specific node in your tree, you can use the -n parameter in combination with the -s parameter. If two species names devided by ':' are provided, all arrangements involved in rearrangement events at the node representing the last common ancestor of both species will be listed in the statistics file in a separate section. This parameter is just usable if -s parameter is set.

5 Results

domRates-Seq defines six event and four solution types and infer domain rearrangements within these sets. 1 and 2 give an overview on the defined event and solution types, respectively.

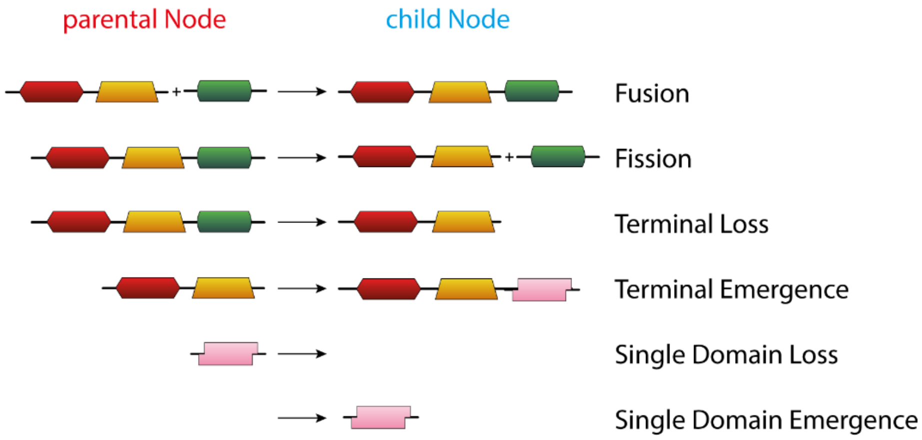

Figure 1: Defined event types in domRates-Seq.

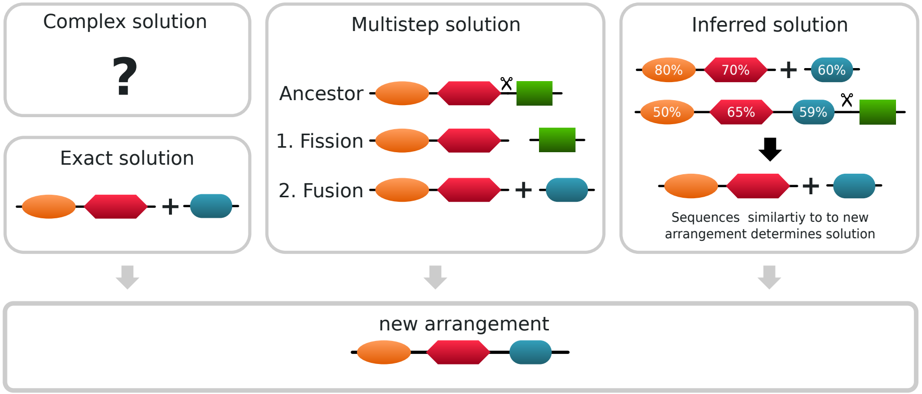

Figure 2: Defined solution types in domRates-Seq.

5.1 Sample output files

The following sections provide example output files on both usage of domRatesSeq.

5.1.1 Inferring domain rearrangement events

As presented in the previous chapters, to infer domain rearrangement events in a phylogeny, one should run a similar command as below:

```
./domRatesSeq -t species.tree -a domain_annotations/ -g Outgroup -o
     results.out -s results_stat.out -d -m BLOSUM62.txt -r
    protein_annotations
```

This command will generate three files: *results.out*, *results_stat.out*, and *results_stat_epd.out*.

A sample main results file (*results.out*) can be found below:

```
# DomRatesSeq version 1.0.0 at Thu May  9 09:02:13 2024
# domRatesSeq -t species.tree -a domain_annotations/ -g Outgroup
-e .dom -f.fasta -o results.out -s results_stat.out -n
-d 1 -p 1 -m BLOSUM62.txt -r protein_annotations -c 0 -i
# Solution types
Exact solutions: 3195
Complex solutions: 308
Multiple steps solutions: 947
Inferred solutions: 902
Loss: 2522
Maintained arrangements total: 121056

Exact and inferred solutions: 4097

# Solution rates
Exact and inferred rate: 3.18%
Complex rate: 0.24%
Multistep rate: 0.73%
Loss rate: 1.96%
Maintained rate: 93.89%

# Event types
Fusions: 4243
Fissions: 1078
Terminal Loss: 286
Terminal Emergences: 20
Single Domain Losses: 856
Single Domain Emergences: 69
Domain arrangement loss: 2522
```

```
# Event rates
Fusion rate: 64.76%
Fission rate: 16.45%
Terminal Loss rate: 4.37%
Terminal Emergence rate: 0.31%
Single Domain Loss rate: 13.06%
Single Domain Emergence rate: 1.05%
```

First two lines contain run-related information like software version, run date and used parameters. Then following four result groups contain the total count of solutions, events and their respective frequencies.

When domRatesSeq is used with -s option as in the sample case above, domRatesSeq also generates statistics output containing events per node:

```
# Number of events per node.
# Node ID #Fusions #Fissions #TerminalLosses #TerminalEmergences #SingleDomainLosses
#SingleDomainEmergences #DomainArrangementLoss #Maintained
0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0
2 153 33 6 5 147 6 5210
3 103 18 4 5 21 26 5295
4 30 7 0 0 17 8 5445
5 82 11 1 0 36 1 5416
6 86 20 3 0 14 1 5467
7 100 19 1 0 49 3 5546
8 101 24 4 1 12 5 5543
9 0 0 0 0 0 0 5734
10 0 0 0 0 0 0 5734
11 125 18 2 0 78 1 5378
12 59 10 3 1 23 2 5431
13 108 27 5 0 13 0 5458
14 148 31 9 0 46 0 5497
15 147 32 7 0 6 3 5570
16 107 18 8 1 39 4 5702
17 70 12 1 1 17 0 5780
18 114 33 20 0 102 0 5488
19 407 70 14 1 17 4 5844
20 146 24 6 1 79 0 5359
21 54 7 2 0 46 1 5419
22 215 29 10 3 58 3 5357
23 216 25 7 1 36 1 5383
24 0 0 0 0 0 0 0 0
```

The file contains tab separated values of Node ID, Fusions, Fissions, Terminal Losses, Terminal Emergences, Single Domain Losses, Single Domain Emergences, Domain Arrangement Loss, Maintained solutions, respectively.

The last file contains tab separated values for all the events and the solutions for each node:

```
# Node-ID solution type event type new arrangement at node arrangement at parental node
5 exact solution single domain loss PF21711
5 inferred solution fission PF00063 | PF00612 PF01843 PF00063 PF00612 PF01843
5 inferred solution fusion PF13833 PF13499 PF00153 PF13833 + PF13499 PF00153
5 inferred solution terminal loss PF00071 PF08356 PF00071 PF08356 PF08355
5 maintained maintained PF00004 PF00004
5 maintained maintained PF00004 PF08542 PF00004 PF08542
```

5.1.2 Tracking domain rearrangement events

To track domain rearrangement events, one should run:

```
domRates —t tree.txt —a domain_annotations/ —g outgroup —o results.
    out —s results_stat.out —d —m BLOSUM62.txt —r
    proteome_annotation/ —c 1 —i NP_003254.2,NP_001305716.1,
    NP_003256.1,NP_612564.1,NP_003259.2,NP_006059.2,NP_057646.1,
    NP_059138.1,NP_057694.2,NP_001017388.1
```

This command will generate five output files. Three of the files have the same structure as we described in 5.1.1. However, instead of presenting the domain rearrangement event of the whole phylogeny, the results show only the domain rearrangement events of the selected proteins. In this example we used human TLR proteins with protein IDs *NP_003254.2, NP_001305716.1, NP_003256.1, NP_612564.1, NP_003259.2, NP_006059.2, NP_057646.1, NP_059138.1, NP_057694.2, NP_001017388.1*

```
# DomRatesSeq version 1.0.0 at Sun Dec 17 18:51:09 2023
# domRatesSeq -t tree.txt -a domain_annotations/ -g outgroup -e .dom -f.fasta
-o results.out -s results_stat.out -n  -d 1 -p 1 -m BLOSUM62.txt -r proteome_annotation/
-c 1 -i NP_003254.2,NP_001305716.1,NP_003256.1,NP_612564.1,NP_003259.2,
NP_006059.2,NP_057646.1,NP_059138.1,NP_057694.2,NP_001017388.1

# Solution types
Exact solutions: 10
Complex solutions: 2
Multiple steps solutions: 3
Inferred solutions: 1
Loss: 0
Maintained arrangements total: 144

Exact and inferred solutions: 11

# Solution rates
Exact and inferred rate: 6.88%
Complex rate: 1.25%
Multistep rate: 1.88%
```

```
Loss rate: 0.00%
Maintained rate: 90.00%

# Event types
Fusions: 12
Fissions: 2
Terminal Loss: 0
Terminal Emergences: 0
Single Domain Losses: 0
Single Domain Emergences: 4

# Event rates
Fusion rate: 66.67%
Fission rate: 11.11%
Terminal Loss rate: 0.00%
Terminal Emergence rate: 0.00%
Single Domain Loss rate: 0.00%
Single Domain Emergence rate: 22.22%
```

The fourth and fifth files can directly be uploaded to ITOL to visualize domain rearrangement events in the evolutionary history of the selected proteins. One of the file is just partially manipulated version of the same newick tree that user provided and the other one specifies the domain rearrangement events. A sample file can be found below:

```
DATASET_PIECHART
SEPARATOR TAB
DATASET_LABEL Domain Rearrangement Events
COLOR #ff0000
FIELD_COLORS #7B68EE #D2691E #556B2F #FF6347 #4682B4 #DB7093
FIELD_LABELS Fusion Fission Terminal domain loss Terminal domain emergence
Single domain loss Single domain emergence
LEGEND_TITLE Legend
LEGEND_POSITION_X 100
LEGEND_POSITION_Y 100
LEGEND_HORIZONTAL 0
LEGEND_SHAPES 2 2 2 2 2 2
LEGEND_COLORS #7B68EE #D2691E #556B2F #FF6347 #4682B4 #DB7093
LEGEND_LABELS Fusion Fission Terminal domain loss Terminal domain emergence
Single domain loss Single domain emergence
LEGEND_SHAPE_SCALES 1 1 1 1 1 1
MAXIMUM_SIZE 10
DATA
intNode1 0 0.5 0 0 0 0 0 0
bacteria 0 0.5 0 0 0 0 0 0
intNode2 0 0.5 0 0 0 0 0 0
```

```
archaea 0 0.5 0 0 0 0 0 0
intNode3 0 0.5 0 0 0 0 0 4
intNode4 0 0.5 0 0 0 0 0 0
44689 0 0.5 0 0 0 0 0 0
intNode5 0 0.5 0 0 0 0 0 0
intNode6 0 0.5 0 0 0 0 0 0
4787 0 0.5 0 0 0 0 0 0
intNode7 0 0.5 0 0 0 0 0 0
67593 0 0.5 0 0 0 0 0 0
164328 0 0.5 0 0 0 0 0 0
intNode8 0 0.5 0 0 0 0 0 0
intNode9 0 0.5 0 0 0 0 0 0
39946 0 0.5 0 0 0 0 0 0
39947 0 0.5 0 0 0 0 0 0
intNode10 0 0.5 0 0 0 0 0 0
29760 0 0.5 0 0 0 0 0 0
intNode11 0 0.5 0 0 0 0 0 0
3702 0 0.5 0 0 0 0 0 0
3694 0 0.5 0 0 0 0 0 0
```

When the tree file and the itol output file are uploaded to ITOL, the user can visualize the domain rearrangement events.

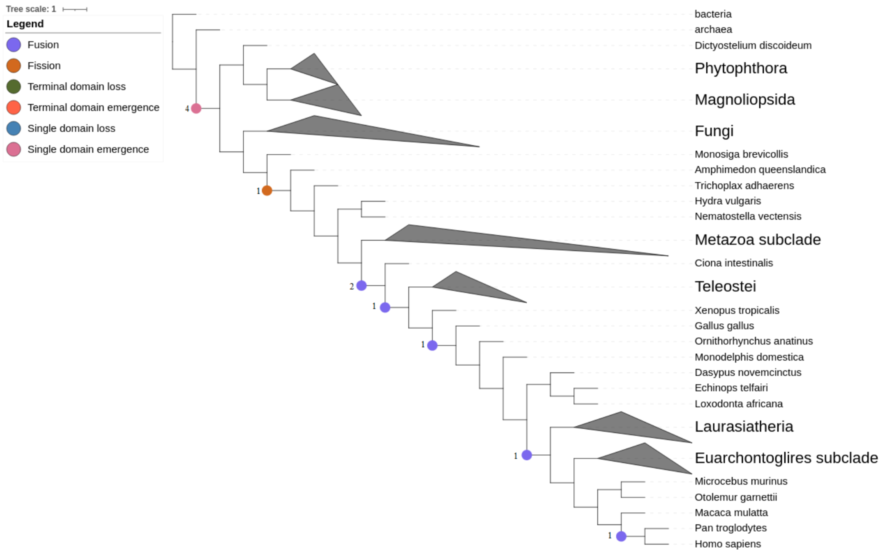

Figure 3: Domain rearrangement events of human TLR proteins.