ADOMA: Alternative Display Of Multiple Alignment

User manual

Dionne Zaal & Benjamin Nota
Department of Blood Cell Research, Sanquin Research
Plesmanlaan 125, 1066 CX Amsterdam
The Netherlands

Contact: dionnezaal@gmail.com

Contents

Quick start

Run ADOMA:

```
$ ./adoma.sh -p [prefix_output] -i [fasta_file(s)]
```

For more parameters check chapter *Run ADOMA* or the help file:

```
$ ./adoma.sh --help
$ ./adoma.sh -h
```

Introduction

About

ADOMA stands for: Alternative Display Of Multiple Alignment. ADOMA can create four different displays of a multiple sequence alignment: a ClustalW alignment in HTML format, a simplified ClustalW alignment in HTML and/or txt format and a colored ClustalW alignment in HTML format. A brief explanation of the output files can be found in the next sections. ADOMA uses ClustalW[1,2] to create the multiple alignment from DNA or protein sequences and displays them slightly different than the normal output of ClustalW.

ClustalW alignment

This output is the default ClustalW alignment, but then in HTML format. It can be easily viewed in an internet browser and with the HTML format no shift of the alignment can occur.

Colored ClustalW alignment

The colored ClustalW alignment gives nucleotides or amino acids a color depending on which nucleotide/amino acid and the alignment consensus. For ADOMA the same color scheme is used as in ClustalX [3]. This colored ClustalW alignment is in HTML format.

Simplified ClustalW alignment

When aligning homologous sequences it can be difficult to identify the mismatches between sequences, for example when there is only one of the 30 sequences that has an mismatch on a certain position. Identifying this mismatch might be needed for allele determination for example. The simplified alignment creates a simpler alignment so that identifying of mismatches between sequences is easier. It displays the parent sequence and for all the other sequences it displays "-" when there is a match between the sequence and the parent sequence, the nucleotide/amino acid when there is a mismatch and "*" when there is a deletion. The simplified ClustalW alignment has a default output in HTML format but ADOMA can also output a txt file of the simplified ClustalW alignment.

! For a full explanation of the output files: see examples of the
! output files in the folder Examples in adoma-[version].

Requirements

ADOMA must be executed in a unix command line.

ADOMA needs ClustalW (version 2.1) to create a multiple sequence alignment from the input sequences. ClustalW must be installed in the unix commandline. For installing and downloading ClustalW: http://www.clustal.org/clustal2/

ADOMA also needs python (version 2.7 or higher) to create the different displays of the alignment. For installing and downloading python: https://www.python.org/downloads/source/

For the display of the alignment files in HTML an internet browser is needed (Google Chrome, Mozilla Firefox, Internet Explorer, etc.). For the TXT file a text editor is needed to view the alignment.

Run ADOMA

ADOMA can be executed with a set of different parameters. All parameters can be found in the help file that can be assessed by typing:

```
$ ./adoma.sh -h
$ ./adoma.sh --help
```

All parameters can be assigned in the order the user prefers. All parameters will be discussed in the next sections. The different parameters can create different output files, these output files are discussed in section: Output ADOMA.

Parameters

! Every parameter that will be discussed in this chapter can only be assigned once in
! the command line. When a parameter is assigned twice the first definition of the
! parameter will be overwritten with the last assignment of that parameter.

Mandatory parameters

There are two mandatory parameters that always must be assigned for ADOMA:

```
-p    Define the prefix for the output files here
-i    The input files in fasta format
      This can be one file or multiple files
```

Prefix parameter

After the prefix parameter the user can define what name the output files must include. If desired the user can also define the path of destination for the output files. But when ADOMA is run from the folder where the output files are requested this is not necessary, two examples:

Running ADOMA with the prefix defined output folder:

```
$ ./adoma.sh -p /PATH/TO/OUTPUT/FOLDER/[prefix] -i [sequences.fa]
```

Running ADOMA from the output folder:

```
$ /PATH/TO/ADOMA/adoma.sh -p [prefix] -i [sequences.fa]
```

! Behind the prefix parameter only one word (or string) can be assigned, for example
! 'dna_sequences'. When 'dna sequences' is defined with a space in between
! only the first word will be assigned to the prefix.

Input file(s) parameter

After the input parameter one or multiple fasta files can be defined for the multiple sequence alignment. ADOMA will always create a new fasta file with all the sequences together (with the defined prefix), even when only one fasta file is defined.

! The input files must contain at least two sequences, otherwise no multiple
! sequence alignment can be created.

Running ADOMA with multiple fasta files:

```
$ ./adoma.sh -p [prefix] -i [sequence1.fa] [sequence2.fa]
```

Type parameters

There are two parameters that can be assigned for defining the type of sequences in the input fasta file(s):

```
-dna        Define that the type of input sequences is DNA [default]
-prot       Define that the type of input sequences is protein
```

These two parameters are important when a colored ClustalW alignment is requested, because the coloring scheme is different for DNA and protein. When none of these two parameters are assigned, ADOMA will assign DNA on default. When inputting DNA sequences therefore it is not necessary to specify this at the parameters.

! When the user wants to input RNA sequences, define DNA as type.

An example for input with protein sequences:

```
$ ./adoma.sh -p [prefix] -i [sequences.fa] -prot
```

! The type parameters cannot be used both at the same time. The sequences are
! either DNA/RNA or protein sequences. When the user defines both parameters,
! ADOMA will print an error message and quit.

Output parameters

The next two parameters can be assigned when extra output files are needed:

```
-txt        Create also a TXT output for the simplified alignment
-color      Create a colored ClustalW alignment in HTML
```

The parameters for the extra output files can both be assigned at the same time.

Example for both txt and colored ClustalW alignment output with protein sequences:

```
$ ./adoma.sh -p [prefix] -i [sequences.fa] -color -txt -prot
```

ClustalW parameter

The last parameter that can be assigned for ADOMA is the clustalw parameter:

```
-clustalw   ClustalW options can be defined here
```

When the user would like to adjust parameters for the ClustalW alignment this can be done behind the clustalw parameter. Only when the parameters for ClustalW are defined behind the clustalw parameter they will be given to ClustalW, otherwise they are ignored. When for example the user would like to output the sequence numbers in the output file and change the gap open penalty score to 30, the user can define:

```
$ ./adoma.sh -p [prefix] -i [sequences.fa] -clustalw -seqnos=ON -gapopen=30
```

When the user does not define parameters a default ClustalW multiple sequence alignment will be created. Some parameters for ClustalW are automatically assigned by ADOMA, so the user does not have to define those parameters, these parameters can be found in table 1. There are also parameters that can be assigned to ClustalW, but cannot be assigned to ADOMA, these parameters can be found in table 2 with their restrictions. Only the general settings and setting for multiple alignments are discussed, because the other parameters are irrelevant for ADOMA.

Table 1: ClustalW parameter already assigned by ADOMA

ClustalW parameter	Explanation Parameter	Parameter from ADOMA
-infile	For input sequences	ADOMA uses the merged fasta file as input for ClustalW.
-outfile	Define name output file	ADOMA already uses the prefix that is assigned by the user for this parameter.
-align	Do full multiple sequence alignment	Set parameter.
-type	Define if sequences are DNA or protein	ADOMA uses type that is put in by the user. If the user did not define the type, default type is DNA.

Table 2: ClustalW parameters with restrictions for ADOMA

ClustalW parameter	Explanation Parameter	Restriction
-output	Define output format	This parameter cannot be changed. ADOMA can only convert a clustal format output. This format is the default output from ClustalW.
-interactive	Open an interactive menu for ClustalW	This parameter cannot be used. ClustalW will open an interactive program, but ADOMA does not respond to this and will not continue.
-seqno_range=ON	Displays the length of the sequence numbers behind the sequence name.	It is possible that the sequence range is not displayed. The sequence name has a limit of 12 characters with the txt format and 30 characters for html format.

Output ADOMA

When ADOMA is executed different output files are created. The output files that ADOMA creates can be found in table 3.

Table 3: Output files that ADOMA can create

Output file	Content output file	Output created
logfile.txt	If errors occur during the program, the errors will be written to this file	Default
[prefix].fa	Merged fasta file from the input	Default
[prefix].aln	ClustalW alignment file	Default
[prefix].dnd	ClustalW guide tree	Default
ClustalW_*[prefix]*_aln.html	ClustalW alignment in HTML format	Default
simplified_*[prefix]*_aln.html	Simplified alignment in HTML format	Default
simplified_*[prefix]*_aln.txt	Simplified alignment in txt format	When txt parameter is assigned
color_*[prefix]*_aln.html	Colored ClustalW alignment in HTML	When color parameter is assigned

Examples with test data

When ADOMA is downloaded and unpacked a folder Test_data is also created in the folder adoma-[version]. The test data was derived from the IPD KIR database [4]. The DNA data are the CDS from alleles of KIR2DL2, the protein data is from KIR2DL4. With these test data the user can test if ADOMA works properly and to see the different files that are created by ADOMA. In the folder Examples there are subfolders with different output files from the test data. In each subfolder a file *command.txt* can be found with the command that was used.

Troubleshouting

Errors can occur during ADOMA. All the errors that will occur are not displayed in the stdout but are written to a file: logfile.txt. When the output is not as expected please check this file to see what errors might have occurred. In this logfile all the parameters that were given by the user are also written to this file. Checking if the parameters were given correctly can therefore also be done by checking the logfile.

The output files are not created or empty

When the output files are empty or not created many things could have gone wrong. The most likely explanation is that ClustalW could not make an alignment. The logfile can contain an error when the files are empty, always check this file. Otherwise:

- Check if the files where defined correctly on the command line
- Check if the files are correct fasta files
- Check if the correct parameters were assigned for ADOMA
- Check if the input files at least contain two sequences
- Check if the ClustalW parameters are defined correctly
- Check if ClustalW and python are installed properly

A part of the alignment that is exactly the same still displays all its nucleotides or amino acids in the simplified ClustalW alignment

This indeed can occur in the simplified ClustalW alignment. ADOMA compares every sequence in the alignment with the parent (or first) sequence in the alignment. When this sequence contains a deletion, indicated with a "*", the sequences will be compared with this deletion. When the other sequences do not contain this deletion their nucleotides or amino acids will be shown in the simplified alignment. Therefore it can occur that a part of the alignment where the parent sequence has a deletion is fully shown with all the nucleotides or amino acids.

Not the whole sequence name is displayed in the alignments

This problem can occur. When writing the sequence name to the txt format of the simplified alignment there is a limitation to the amount of characters (12 characters) that can be written to the alignment file. This was built in the code to ensure that the alignment will align properly so it can be easily viewed. In the HTML format of the simplified, colored and regular alignment there is also a restriction of the amount of characters but this is larger (30 characters) than the txt format.

Could not find a solution for the error? Or do you want to report a bug? Contact the developers at: http://sourceforge.net/projects/adoma or email: dionnezaal@gmail.com.

References

1. Thompson JD, Higging DG, Gibson TJ. (1994). CLUSTALW: improving the sensitivity of progressive multiples sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., 22, 4673-4680.
2. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007). ClustalW and Clustal X version 2.0. *Bioinformatics*, 23, 2947-2948.
3. https://www.cgl.ucsf.edu/chimera/docs/ContributedSoftware/multalignviewer/colprot.par
4. http://www.ebi.ac.uk/ipd/kir/