MARS Documentation

Daniel J. Parente
dparente@kumc.edu

November 24, 2014

Contents

1 Quick Start

MARS accepts the path to two FASTA-formatted files, a target alignment (target) and a reference alignment (refer), and the path to one similarity matrix (simmat) and writes results to standard output (stdout). Call the program, under Windows, by opening a command prompt, changing the directory to one that contains MARS.exe and all input files, and executing:

```
MARS.exe [target] [refer] [simmat] > [Output Path]
```

Under Linux, execute the command:

```
./MARS [target] [refer] [simmat] > [Output Path]
```

If this does not work, you may need to set permissions, see Permissions: MARS will not execute under linux due to a Permission Denied Error.

2 Overview

MARS solve the problem of merging two multiple sequence alignments (MSAs) that share some subset of sequences ("guide sequences") into a single alignment. This is achieved by prioritizing alignment of the guide sequences with their counterpart in each of the two input alignments, over minimization of some similarity criteria (e.g. as might be done by CLUSTAL or MUSCLE). In columns that contain gaps in all guide sequences, MARS uses columwise profile similarity as a fallback. Thus, MARS prioritizes guide sequence alignment constraints over similarity information, without completely ignoring similarity information.

The Maintainer of Alignments using Reference Sequences (**MARS**) program accepts three arguments arguments, which must be paths to files: a Target Alignment, a Reference Alignment, and a Similarity Matrix. The alignments must be in FASTA format (see FASTA format) and the similarity matrix must be similar to matrices accepted by BLAST (see Similarity Matrix). Output will be written to standard output (stdout). Execute the program as shown in Quick Start.

3 Dependencies

MARS requires installation of the .NET 4.0 (or greater) framework.

3.1 Windows

Under Windows, this program can be freely obtained from Microsoft at https://www.microsoft.com/en-us/download/details.aspx?id=17851. One newer Windows machines, this framework may already be installed by default.

3.2 Linux

Under Linux, the Mono .NET framework can be used instead. Consult your package manager (apt, yum, etc) for specific installation instructions. For example, Mono can be installed under Debian using:

```
sudo apt-get install mono-complete
```

Under some Linux distributions, mono may be installed by default.

4 Input files

4.1 FASTA format

MARS accepts two FASTA-formatted files. The FASTA format, as it is used by BLAST, is described by NCBI here: http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml

For the purposes of MARS, FASTA format is defined as follows:

FASTA files stores a series of sequence records. Each records begins with a description line (defline), followed by one or more lines of sequence information. Description lines must begin with the greater-than (>) symbol and contain a set of alphanumeric characters (possibly including space and some special characters, e.g. !@#$_, but not >), which typically describe the gene or protein in words (e.g. "E.Coli LacI"). Sequence information is recorded as a string of characters representing amino acid residues/nucleotides or a gap (represented only by a dash, -). Characters are case-sensative (e.g. 'A' is different from 'a') and are limited to the set of characters present in the supplied similarity matrix, see Similarity Matrix, plus the gap character (-). The sequence information in every record **must** be of the same length, including gaps, as should be true of any multiple sequence alignment. End of lines (EOLs) may use either a LF (Linux-style) character or CRLF (Windows-style) character pair, but should be chosen consistently. The final sequence must end with a EOL character.

An example FASTA file is shown below:

```
>SomeProtein
MKPVTLYDVAEYAGVSYQ----VVNQASHVSAKTREKVEAAMAELNYIPNRVAQQLAGKQ
>AnotherHomolog
--MATIKDVAKRANVSTTTVSHVINKTRFVAEETRNAVWAAIKELHYSPSAVARSLKVNH
```

4.2 Similarity Matrix

Similarity matrices are accepted in the same format as those accepted by BLAST, except that (a) no comment lines are permitted, and (b) gaps are denoted with a dash (-), not an asterisk (*); comments lines are those which are preceeded by a hash (#). Example matrices are available from NCBI via FTP at ftp://ftp.ncbi.nih.gov/blast/matrices/

In brief, the similarity matrices are space-delimited text files, with multiple spaces allowed between fields. The first (header) line contains a list of amino acids/nucleotide one one letter codes, starting from the second field. Each subsequent line, starts with a one letter code in the first field, with integers in each subsequent field indicating the score derived from the alignment of the amino acid/nucleotide in the first field of the current line with the corrosponding amino acid/nucleotide in this field of the header line. This format has the benefit of being easily editable from text-editors when using a monospaced font.

For example, (part of the) BLOSUM62 matrix is shown below:

	A	R	N	D	-
A	4	-1	-2	-2	-4
R	-1	5	0	-2	-4
N	-2	0	6	1	-4
D	-2	-2	1	6	-4
-	-4	-4	-4	-4	1

5 Troubleshooting

5.1 Permissions: MARS will not execute under linux due to a Permission Denied Error

Possibly, the binary's permissions are not correctly set. Try making the binary executable by executing this command, in the same directory as the MARS binary:

```
chmod u+x MARS
```

6 Contact

If have have specific questions, bug reports or feature requests, please contact dparente@kumc.edu. Though I am quite interested in the general usefulness of the software, I can not explicitly guarantee that I will be able to provide full support for every user and environment.

7 Disclaimer

This document is intended to provide technical information that many be useful to operators of MARS. Any hardware/software configurations discussed are for illustrative purposes only and do not imply an endorsement of any company, product or service, by the author of MARS or the University of Kansas Medical Center. All advice is advisory only and provided without any warranty. You are responsible for the integrity of the machines on which this software is run. Neither the author, nor the University of Kansas Medical Center, will be held liable for any damage that is a result of following the guidance provided in this document or due to the use of MARS.