

# Documentation for **AliStat**

## **Quantifying the Sparseness of Multiple Sequence Alignments used in Phylogenetic and Phylogenomic Studies**

Copyright (C) 2014, CSIRO

Dec, 2019

# Disclaimer

CSIRO Open Source Software License Agreement (variation of the BSD / MIT License)

Copyright (c) 2014, Commonwealth Scientific and Industrial Research Organisation (CSIRO) ABN 41 687 119 230.

All rights reserved. CSIRO is willing to grant you a license to this AliStat on the following terms, except where otherwise indicated for third party material.

Redistribution and use of this software in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of CSIRO nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission of CSIRO.

EXCEPT AS EXPRESSLY STATED IN THIS AGREEMENT AND TO THE FULL EXTENT PERMITTED BY APPLICABLE LAW, THE SOFTWARE IS PROVIDED "AS-IS". CSIRO MAKES NO REPRESENTATIONS, WARRANTIES OR CONDITIONS OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY REPRESENTATIONS, WARRANTIES OR CONDITIONS REGARDING THE CONTENTS OR ACCURACY OF THE SOFTWARE, OR OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT, THE ABSENCE OF LATENT OR OTHER DEFECTS, OR THE PRESENCE OR ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE.

TO THE FULL EXTENT PERMITTED BY APPLICABLE LAW, IN NO EVENT SHALL CSIRO BE LIABLE ON ANY LEGAL THEORY (INCLUDING, WITHOUT LIMITATION, IN AN ACTION FOR BREACH OF CONTRACT, NEGLIGENCE OR OTHERWISE) FOR ANY CLAIM, LOSS, DAMAGES OR OTHER LIABILITY HOWSOEVER INCURRED. WITHOUT LIMITING THE SCOPE OF THE PREVIOUS SENTENCE THE EXCLUSION OF LIABILITY SHALL INCLUDE: LOSS OF PRODUCTION OR OPERATION TIME, LOSS, DAMAGE OR CORRUPTION OF DATA OR RECORDS; OR LOSS OF ANTICIPATED SAVINGS, OPPORTUNITY, REVENUE, PROFIT OR GOODWILL, OR OTHER ECONOMIC LOSS; OR ANY SPECIAL, INCIDENTAL, INDIRECT, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES, ARISING OUT OF OR IN CONNECTION WITH THIS AGREEMENT, ACCESS OF THE SOFTWARE OR ANY OTHER DEALINGS WITH THE SOFTWARE, EVEN IF CSIRO HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH CLAIM, LOSS, DAMAGES OR OTHER LIABILITY.

APPLICABLE LEGISLATION SUCH AS THE AUSTRALIAN CONSUMER LAW MAY APPLY REPRESENTATIONS, WARRANTIES, OR CONDITIONS, OR IMPOSES OBLIGATIONS OR LIABILITY ON CSIRO THAT CANNOT BE EXCLUDED, RESTRICTED OR MODIFIED TO THE FULL EXTENT SET OUT IN THE EXPRESS TERMS OF THIS CLAUSE ABOVE "CONSUMER GUARANTEES". TO THE EXTENT THAT SUCH CONSUMER GUARANTEES CONTINUE TO APPLY, THEN TO THE FULL EXTENT PERMITTED BY THE APPLICABLE LEGISLATION, THE LIABILITY OF CSIRO UNDER THE RELEVANT CONSUMER GUARANTEE IS LIMITED (WHERE PERMITTED AT CSIRO'S OPTION) TO ONE OF FOLLOWING REMEDIES OR SUBSTANTIALLY EQUIVALENT REMEDIES:

- (a) THE REPLACEMENT OF THE SOFTWARE, THE SUPPLY OF EQUIVALENT SOFTWARE, OR SUPPLYING RELEVANT SERVICES AGAIN;
- (b) THE REPAIR OF THE SOFTWARE;
- (c) THE PAYMENT OF THE COST OF REPLACING THE SOFTWARE, OF ACQUIRING EQUIVALENT SOFTWARE, HAVING THE RELEVANT SERVICES SUPPLIED AGAIN, OR HAVING THE SOFTWARE REPAIRED.

IN THIS CLAUSE, CSIRO INCLUDES ANY THIRD PARTY AUTHOR OR OWNER OF ANY PART OF THE SOFTWARE OR MATERIAL DISTRIBUTED WITH IT. CSIRO MAY ENFORCE ANY RIGHTS ON BEHALF OF THE RELEVANT THIRD PARTY.

## Credits

The software *AliStat* was developed by the bioinformatics and phylogenomics team in Ecosystem Sciences, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, Australia.

This project is collaboration between Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, Australia, and ZFMK, Zentrum für Molekulare Biodiversitätsforschung (zmb), Bonn, Germany.

# Contents

1. Terminologies.....	4
2. Installation of AliStat .....	5
3. Usage of AliStat .....	5
4. Output of AliStat.....	9
4.1 Table 1 - completeness scores for individual sequences ( $C_r$ ) .....	9
4.2 Table 2 - completeness scores for individual sites ( $C_c$ ).....	9
4.3 Table 3 - Distribution of C scores for individual sites ( $C_c$ ).....	10
4.4 Table 4 - Matrix with C scores for pairs of sequences ( $C_{ij}$ ) .....	10
4.5 Table 5 - Matrix with incompleteness scores for pairs of sequences ( $I_{ij}$ ) .....	11
4.6 Table 6 - Table with completeness scores ( $C_{ij}$ ) and incompleteness scores ( $I_{ij}$ ) .....	11
4.7 Summary statistics of the alignment.....	12
4.8 P-distance table – p-distance for pairs of sequences ( $P_{ij}$ ) .....	12
4.9 Masked and discarded (reordered) alignment.....	13
4.10 Heatmap.....	14
5. More examples.....	15
6. Contact person .....	17
7. References.....	17

# 1. Terminologies

## Ambiguous characters

For DNA nucleotides, all the characters which are not considered in the coding types selected by the users are regarded as ambiguous characters. For example, if the selected coding type is "C, T, R (i.e. C, T, AG)", then the characters other than A, C, G, T, R will be regarded as ambiguous characters, like '-' (i.e. gap), Y, S, W, K, M, B, D, H, V, etc.

For amino acids, we only consider the standard amino acids. Thus, all the characters except A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y are considered as ambiguous characters, like '-' (i.e. gap), B, J, O, U, X, Z, etc.

## $C_a$ - Completeness score for the alignment

$C_a$  = Total number of unambiguous characters / (number of sequences \* length of alignment)

## $C_r$ - Completeness scores for individual sequences

$C_r$  = Number of unambiguous characters in the sequence / alignment length

## $C_c$ - Completeness scores for individual sites

$C_c$  = Number of unambiguous characters in the column / number of sequences

## $C_{ij}$ - Completeness scores for pairs of sequences

$C_{ij}$  = Number of columns in which the corresponding characters of both  $i$ - and  $j$ -th sequence are unambiguous / length of alignment. By definition,  $C_{ii}$  is always 1.

## $I_{ij}$ - Incompleteness scores for pairs of sequences

$$I_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 - C_{ij} & \text{otherwise} \end{cases}$$

## $P_{ij}$ - P-distance for pairs of sequences

Let  $W_{ij}$  be the set of columns in which the corresponding characters of both  $i$ - and  $j$ -th sequence in the alignment are unambiguous.

$$P_{ij} = (\text{Number of substitutions between } i\text{- and } j\text{-th sequence along the columns in } W_{ij}) / |W_{ij}|$$

$P_{ij}$  is undefined if  $|W_{ij}| = 0$ , and  $P_{ii}$  is always 0.

## 2. Installation of AliStat

The software was written in C++, and it has been tested under linux and MacOS platform. You need to have C++ compiler installed in the machine in order to compile the source codes. The compilation steps are shown as follows:

```
$ tar -zxvf AliStat_1.11.tar.gz
$ cd AliStat_1.11
$ make
```

Then an executable file named *AliStat* will appear.

## 3. Usage of AliStat

```
Syntax: ./alistat <alignment file> <data type> [other options]
        ./alistat -h
```

<alignment file>	<p>Multiple alignment file in FASTA format</p> <p>The gap is represented by hyphen (i.e. "-")</p> <p>An example of a nucleotide alignment file:</p> <pre>&gt;Seq_1 -----CGTACCTCTATTGTATGTTTGATGCCTCTATTGTATGTTTGATGC &gt;Seq_2 -----CGTACATCTGTTGCATGATTAATGGATCTGTTGCATGATTAATGG &gt;Seq_3 -----CGTAC-----CTCTATTGTATGTTTGATGC &gt;Seq_4 -----CGTACATCTGTTGCATGATTAATGG----- &gt;Seq_5 CGTAC-----ATCTGTTGCATGATTAATGGCTCTATTGTATGTTTGATGC</pre>
<data type>	<p>The data type of the alignment file.</p> <p>1 - Nucleotide; 2 - Amino acid; 3 - Mixture of nucleotide and amino acid</p> <p>For 3, the user has to specify the data type for each partition inside the <i>partition file</i>. For details, please check the option '-p'.</p>

## Other options:

-b	<p>Report the brief summary of figures to the screen</p> <p>Output format: File, #seqs, #sites, Ca, Cr_max, Cr_min, Cc_max, Cc_min, Cij_max, Cij_min (this option cannot work with the options: -o, -t, -r, -m, -i, -d)</p>
-c <coding type>	<p>The coding type in the alignment.</p> <p>This option is only valid when &lt;data type&gt; = 1.</p> <p>The nucleotides in the alignment will be converted into the corresponding characters according to the coding type the user selected.</p> <p>The possible coding types are:</p> <pre> 0 - A, C, G, T [default]; 1 - C, T, R (i.e. C, T, AG); 2 - A, G, Y (i.e. A, G, CT); 3 - A, T, S (i.e. A, T, CG); 4 - C, G, W (i.e. C, G, AT); 5 - A, C, K (i.e. A, C, GT); 6 - G, T, M (i.e. G, T, AC); 7 - K, M      (i.e. GT,  AC); 8 - S, W      (i.e. GC,  AT); 9 - R, Y      (i.e. AG,  CT); 10 - A, B      (i.e. A,   CGT); 11 - C, D      (i.e. C,   AGT); 12 - G, H      (i.e. G,   ACT); 13 - T, V      (i.e. T,   ACG); </pre>
-o <output prefix>	<p>Prefix for output files (default: &lt;alignment file&gt; w/o .ext)</p>
-n <FILE>	<p>Only consider sequences with names listed in FILE</p> <p>An example of the file (with a sequence name on each line):</p> <pre> Seq_2 Seq_4 Seq_5 </pre>
-t <n <sub>1</sub> ,n <sub>2</sub> ,...>	<p>Only output the tables n<sub>1</sub>, n<sub>2</sub>, ...</p> <ol style="list-style-type: none"> <li>1 - Completeness scores for individual sequences (C<sub>r</sub>)</li> <li>2 - Completeness scores for individual sites (C<sub>c</sub>)</li> <li>3 - Distribution of completeness scores for individual sites (C<sub>c</sub>)</li> <li>4 - Matrix with completeness scores for pairs of sequences (C<sub>ij</sub>) (Together with the heatmaps for C<sub>ij</sub> scores)</li> <li>5 - Matrix with incompleteness scores for pairs of sequences (I<sub>ij</sub>)</li> <li>6 - Table with completeness scores (C<sub>ij</sub>) and incompleteness scores (I<sub>ij</sub>)</li> </ol> <p>By default, the program does not output any table or the heat map image. For details about the tables, please refer to the Section "Output of AliStat".</p>

<p>-p &lt;FILE&gt;</p>	<p>Specify the partitions in the <i>partition file</i></p> <p>For &lt;data type&gt; = 1 or 2, partition file format:</p> <p>"&lt;partition name 1&gt;=&lt;start pos&gt;-&lt;end pos&gt;, ..."  "&lt;partition name 2&gt;= ... "</p> <p>An example:</p> <pre>part1=1-50, 60-100 part2=101-200</pre> <p>For &lt;data type&gt; = 3, partition file format:</p> <p>"&lt;NU/AA&gt;, &lt;partition name 1&gt;=&lt;start pos&gt;-&lt;end pos&gt;, ..."  "&lt;NU/AA&gt;, &lt;partition name 2&gt;= ..."</p> <p>An example:</p> <pre>NU,part1=1-50, 60-100 AA,part2=101-200</pre> <p>For nucleotide data, in order to consider specific codon position:</p> <p>"&lt;partition name&gt;=&lt;start pos&gt;-&lt;end pos&gt;, ... \3", or  "&lt;NU&gt;, &lt;partition name&gt;=&lt;start pos&gt;-&lt;end pos&gt;, ... \3"</p> <p>For example:</p> <pre>NU,codon1=1-7\3 NU,codon2=2-8\3 NU,codon3=3-9\3</pre>
<p>-s &lt;n<sub>3</sub>,n<sub>4</sub>&gt;</p>	<p>Sliding window analysis: window size = n<sub>3</sub>; step size = n<sub>4</sub></p> <p>The first sliding window: [1, n<sub>3</sub>]  The second sliding window: [1+n<sub>4</sub>, n<sub>3</sub>+n<sub>4</sub>]  .....</p> <p>Each sliding window will be regarded as a partition</p> <p>This option cannot be used with '-p' at the same time</p>
<p>-r &lt;row col both&gt;</p>	<p>Reorder the rows/columns (or both) of the alignment according to the C<sub>r</sub>/C<sub>c</sub> scores</p> <p>All the tables will be displayed according to the reordered alignment</p> <p>To output the reordered alignment, please also use the option '-m'</p>

-m <n <sub>5</sub> >	<p>Mask the alignment; <math>0 \leq n_5 \leq 1</math></p> <p>Output the (reordered) alignment with columns <math>C_c \geq n_5</math> in the file 'Mask.fst', the (reordered) alignment with columns <math>C_c &lt; n_5</math> in the file 'Disc.fst', and the (reordered) alignment with an extra row in the first line to indicate whether the column is masked in the file 'Stat.fst'.</p> <p>Special case: if no &lt;n<sub>5</sub>&gt; is supplied, whole alignment is outputted in the file 'Mask.fst'; if &lt;n<sub>5</sub>&gt; is 0, the alignment with columns <math>C_c \geq 0</math> is outputted in the file 'Mask.fst'.</p>
-i <n <sub>6</sub> >	<p>Generate heat map image for <math>C_{ij}</math> scores of sequence pairs</p> <p>1 - Triangular heat map 2 - Rectangular heat map 3 - Both</p> <p>(if no &lt;n<sub>6</sub>&gt;, then both triangular and rectangular heat map files are outputted)</p>
-d	<p>Report the p distances between sequences in the file with extension '.p-dist.csv'</p> <p>By default, this option is disabled</p> <p>Computation of p distances may take long time for large number of sequences</p> <p>For more information, please refer to the Section “Output of AliStat”.</p>
-h	The help page



## 4. Output of AliStat

### 4.1 Table 1 - completeness scores for individual sequences ( $C_r$ )

Name of the file: *<output prefix>.table1.csv*, or *<output prefix>.<partition name>.table1.csv* if partition information is provided.

$C_r$  provides the completeness score for individual sequences in the alignment (considering only the positions specified for the partition if the partition exists).

$C_r$  = Number of unambiguous characters in the sequence / alignment length (or partition length)

Referring to the alignment example before, the corresponding table 1 is as follows:

Seq ID	Sequence	Sites	$C_r$
1	Seq_1	45	0.9
2	Seq_2	45	0.9
3	Seq_3	25	0.5
4	Seq_4	25	0.5
5	Seq_5	45	0.9

Note: Comma is used as the delimiter inside the file.

### 4.2 Table 2 - completeness scores for individual sites ( $C_c$ )

Name of the file: *<output prefix>.table2.csv*, or *<output prefix>.<partition name>.table2.csv* if partition information is provided.

$C_c$  provides the completeness score for individual sites in an alignment (considering only the positions specified for the partition if the partition exists).

$C_c$  = Number of unambiguous characters in the column / number of sequences

Referring to the alignment example before, the corresponding table 2 is as follows:

Site ID	$C_c$
1	0.2
2	0.2
3	0.2
4	0.2
5	0.2
6	0.8
7	0.8
8	0.8
9	0.8
10	0.8
.....	

Note: Comma is used as the delimiter inside the file.

### 4.3 Table 3 - Distribution of C scores for individual sites (C<sub>c</sub>)

Name of the file: `<output prefix>.table3.csv`, or `<output prefix>.<partition name>.table3.csv` if partition information is provided

This table shows the distribution of C<sub>c</sub> scores.

Referring to the alignment example before, the corresponding table 3 is as follows:

Here X represents the number of ambiguous characters at a given site

X	Cc	Sites	Proportion	Cumulative
0	1	0	0	0
1	0.8	45	0.9	0.9
2	0.6	0	0	0.9
3	0.4	0	0	0.9
4	0.2	5	0.1	1
5	0	0	0	1

Note: Comma is used as the delimiter inside the file.

### 4.4 Table 4 - Matrix with C scores for pairs of sequences (C<sub>ij</sub>)

Name of the file: `<output prefix>.table4.dis`, or `<output prefix>.<partition name>.table4.dis` if partition information is provided

C<sub>ij</sub> provides the completeness score for each pairs of sequences in an alignment (considering only the positions specified for the partition if the partition exists).

C<sub>ij</sub> = Number of columns in which the corresponding characters of both *i*- and *j*-th sequence are unambiguous / length of alignment (or length of partition)

By definition, C<sub>ii</sub> is always 1.

Referring to the alignment example before, the corresponding table 4 is as follows:

5					
Seq_1	1	0.9	0.5	0.5	0.8
Seq_2	0.9	1	0.5	0.5	0.8
Seq_3	0.5	0.5	1	0.1	0.4
Seq_4	0.5	0.5	0.1	1	0.4
Seq_5	0.8	0.8	0.4	0.4	1

Note: Tab is used as the delimiter inside the file.

The number in the first line is the total number of sequences.

Starting from the second line, it shows a symmetric square matrix.

The first column states the name of the sequence for each row of the matrix, and the order of the sequences for the columns is always the same as that for the rows.

## 4.5 Table 5 - Matrix with incompleteness scores for pairs of sequences ( $I_{ij}$ )

Name of the file: `<output prefix>.table5.dis`, or `<output prefix>.<partition name>.table5.dis` if partition information is provided

$I_{ij}$  provides the incompleteness score for each pairs of sequences in an alignment (considering only the positions specified for the partition if the partition exists).

$$I_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 - C_{ij} & \text{otherwise} \end{cases}$$

Referring to the alignment example before, the corresponding table 5 is as follows:

5					
Seq_1	0	0.1	0.5	0.5	0.2
Seq_2	0.1	0	0.5	0.5	0.2
Seq_3	0.5	0.5	0	0.9	0.6
Seq_4	0.5	0.5	0.9	0	0.6
Seq_5	0.2	0.2	0.6	0.6	0

Note: Tab is used as the delimiter inside the file.

The number in the first line is the total number of sequences.

Starting from the second line, it shows a symmetric square matrix.

The first column states the name of the sequence for each row of the matrix, and the order of the sequences for the columns is always the same as that for the rows.

The format of this file is suitable as a distance matrix for some phylogenetic software tools, like PHYLIP.

## 4.6 Table 6 - Table with completeness scores ( $C_{ij}$ ) and incompleteness scores ( $I_{ij}$ )

Name of the file: `<output prefix>.table6.csv`, or `<output prefix>.<partition name>.table6.csv` if partition information is provided

$C_{ij}$  provides the completeness score for each pairs of sequences in an alignment (considering only the positions specified for the partition if the partition exists).

$I_{ij}$  provides the incompleteness score for each pairs of sequences in an alignment (considering only the positions specified for the partition if the partition exists).

Referring to the alignment example before, the corresponding table 6 is as follows:

taxon 1	taxon 2	$C_{ij}$	$I_{ij}$
Seq_1	Seq_1	1	0
Seq_1	Seq_2	0.9	0.1
Seq_1	Seq_3	0.5	0.5
Seq_1	Seq_4	0.5	0.5
Seq_1	Seq_5	0.8	0.2
Seq_2	Seq_1	0.9	0.1
Seq_2	Seq_2	1	0
Seq_2	Seq_3	0.5	0.5
Seq_2	Seq_4	0.5	0.5
.....			

Note: Comma is used as the delimiter inside the file.

## 4.7 Summary statistics of the alignment

Name of the file: *<output prefix>.summary.txt*, or *<output prefix>.<partition name>.summary.txt* if partition information is provided

This file shows the summary statistics of the alignment (considering only the regions specified for the partition if the partition exists), which includes the following statistics:

- $C_a$  provides the completeness score for the alignment.  
 $C_a = \text{Total number of unambiguous characters} / (\text{number of sequences} * \text{length of alignment})$
- Maximum and minimum values of  $C_r$  (i.e. completeness score for individual sequences in the alignment)
- Maximum and minimum values of  $C_c$  (i.e. completeness score for individual sites in the alignment)
- Maximum and minimum values of  $C_{ij}$ , where  $i \neq j$  (i.e. completeness score for each pairs of sequences in the alignment)

As an example, referring to the alignment before (i.e. the example file name: test.fa), the corresponding summary statistics is as follows:

```
Name of input file ..... test.fa
Type of data ..... nucleotides
Characters used ..... ACGT
Number of sequences in the alignment ..... 5
Number of sites in the alignment ..... 50
Pairs of sequences ..... 10
Completeness (C) score for the alignment (Ca) ..... 0.74
Maximum C-score for individual sequences (Cr_max) ..... 0.9
Minimum C-score for individual sequences (Cr_min) ..... 0.5
Maximum C-score for individual sites (Cc_max) ..... 0.8
Minimum C-score for individual sites (Cc_min) ..... 0.2
Maximum C-score for pairs of sequences (Cij_max, i!=j) ..... 0.9
Minimum C-score for pairs of sequences (Cij_min, i!=j) ..... 0.1

C scores for individual sequences (Cr) stored in ..... test.table1.csv
C scores for individual sites (Cc) stored in ..... test.table2.csv
Distribution of C scores for individual sites (Cc) in ..... test.table3.csv
Matrix with C scores for individual pairs of seqs (Cij) in .... test.table4.dis
Matrix with incompleteness scores (Iij = 1.0 - Cij) in ..... test.table5.dis
Table with C scores (Cij) and incompleteness scores (Iij) in .. test.table6.csv
```

## 4.8 P-distance table – p-distance for pairs of sequences ( $P_{ij}$ )

This table will be outputted only when the p-distance option (i.e. “-d”) is used.

Name of the file: *<output prefix>.p-dist.csv*, or *<output prefix>.<partition name>.p-dist.csv* if partition information is provided

$P_{ij}$  provides the p-distance for each pairs of sequences in the alignment (considering only the positions specified for the partition if the partition exists).

Let  $W_{ij}$  be the set of columns in which the corresponding characters of both  $i$ - and  $j$ -th sequence in the alignment are unambiguous.

$$P_{ij} = (\text{Number of substitutions between } i\text{- and } j\text{-th sequence along the columns in } W_{ij}) / |W_{ij}|$$

$P_{ij}$  is undefined if  $|W_{ij}| = 0$ , and  $P_{ii}$  is always 0.

Referring to the alignment example before, the corresponding p-distance table is as follows:

5					
Seq_1	0	0.27	0	0.24	0.15
Seq_2	0.27	0	0.24	0	0.15
Seq_3	0	0.24	0	0	0
Seq_4	0.24	0	0	0	0
Seq_5	0.15	0.15	0	0	0

Note: Comma is used as the delimiter inside the file.

The number in the first line is the total number of sequences.

Starting from the second line, it shows a symmetric square matrix.

The first column states the name of the sequence for each row of the matrix, and the order of the sequences for the columns is always the same as that for the rows.

## 4.9 Masked and discarded (reordered) alignment

The masked and the discarded alignment will be outputted only when the option “-m” is used.

The masked alignment with columns  $C_c \geq n7$  will be outputted in the file ‘Mask.fst’ (or ‘[partition name].Mask.fst’ if a partition file is used).

The discarded alignment with columns  $C_c < n7$  will be outputted in the file ‘Disk.fst’ (or ‘[partition name].Disc.fst’ if a partition file is used).

If the reorder option “-r” is also used, then the rows/columns (or both) of alignment will be sorted according  $C_r/C_c$  (or both) scores.

The original alignment with an extra line in the beginning to show which columns are included in the masked alignment, will be outputted in the file ‘Stat.fst’ (or ‘[partition name].Stat.fst’ if a partition file is used).

Referring to the alignment example before, when using the option “-m 0.7”, the file ‘Stat.fst’ will be as follows:

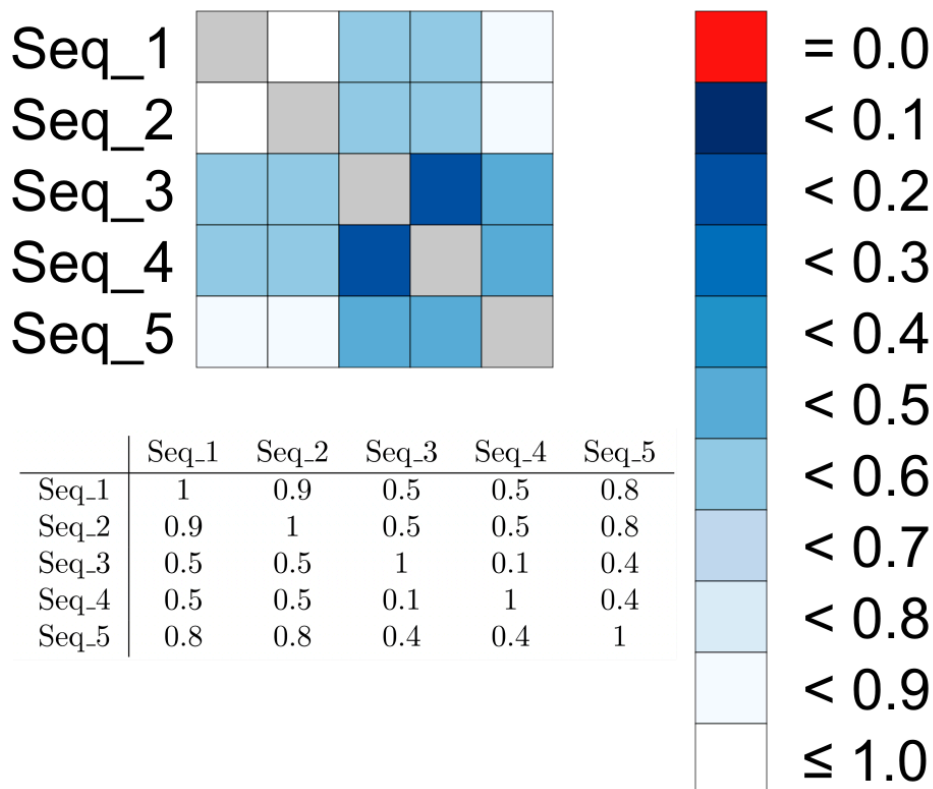
```
>Mask
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
>Seq_1
-----CGTACCTCTATTGTATGTTTGATGCCTCTATTGTATGTTTGATGC
>Seq_2
-----CGTACATCTGTTGCATGATTAATGGATCTGTTGCATGATTAATGG
>Seq_3
-----CGTAC-----CTCTATTGTATGTTTGATGC
>Seq_4
-----CGTACATCTGTTGCATGATTAATGG-----
>Seq_5
CGTAC-----ATCTGTTGCATGATTAATGGCTCTATTGTATGTTTGATGC
```

Note: ‘E’ means the column is excluded and ‘I’ means the column is included in the masked alignment.

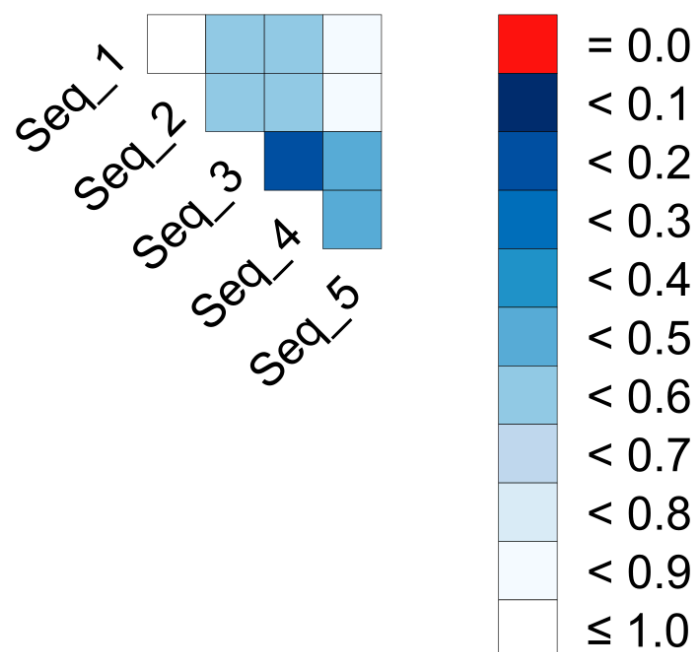
## 4.10 Heatmap

The heatmap files can visually display the  $C_{ij}$  values and they are in SVG image format. The rectangular heatmap will be outputted in the file with extension *'full.svg'*, while the triangular one will be outputted in the file with extension *'triangular.svg'*.

Referring to the alignment example before, the following is rectangular heatmap for table 4:



And the corresponding triangular heatmap table is:



## 5. More examples

Sequence alignment file: M\_Ento.fas (Source: <http://doi.org/10.5061/dryad.mk8p7> [1])

Data type: Amino acid

Partition file: M\_Ento.partition

In this example, we consider three partitions (as stated inside the file *M\_Ento.partition*) :

Partition name	Positions
gene409969g	1 - 120
gene409979g	121-290
gene409986g	291-460

Command to run *AliStat* with the option to reorder both the rows and the columns of the alignment according to the Cr and Cc scores, respectively:

```
$. /AliStat M_Ento.fas 2 -p M_Ento.partition -r both -i 1
```

Then for each partition, the tables 1-4 and the summary file are outputted.

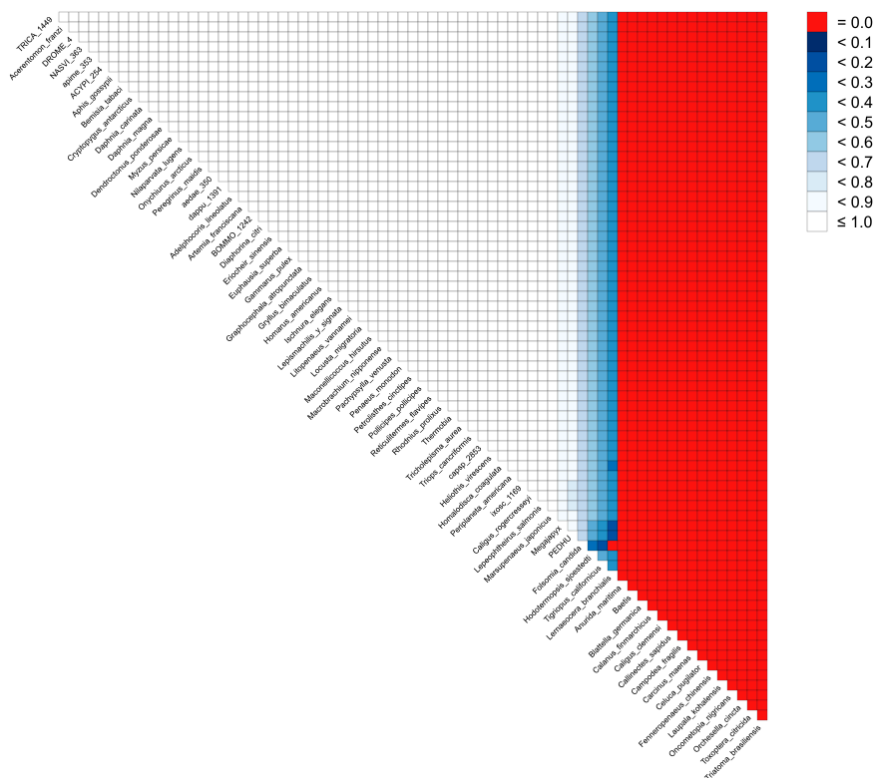
According to the summary files, the completeness scores of the partitions are:

Partition	Completeness score for the partition ( $C_a$ )
gene409969g	0.725231
gene409979g	0.790033
gene409986g	0.362173

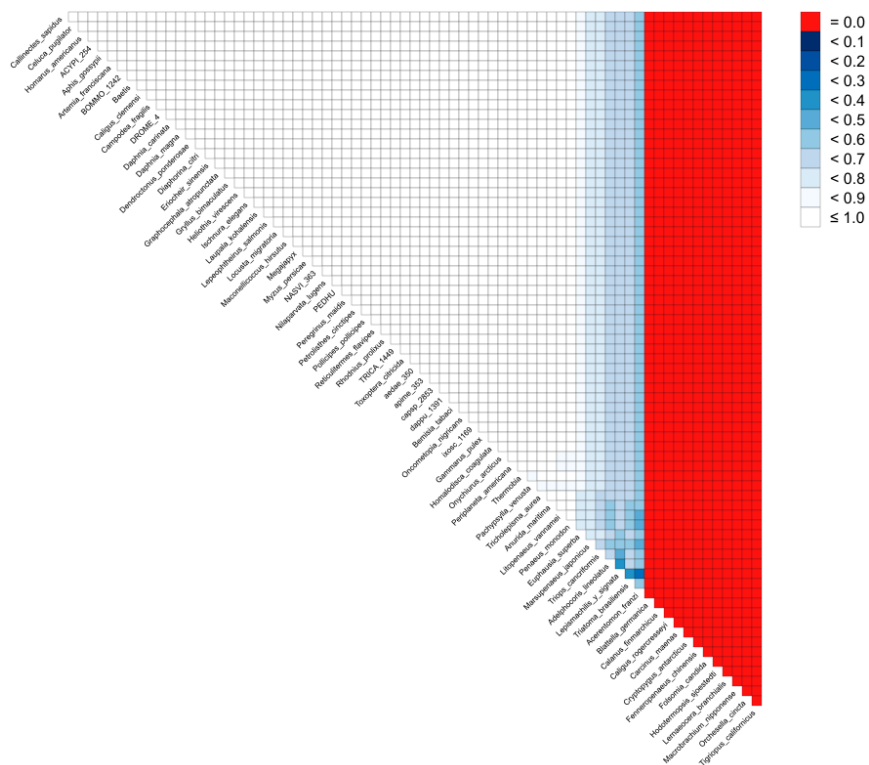
It indicates that the partition *gene409979g* has highest level of completeness, while the partition *gene409986g* has lowest level of completeness.

Since *AliStat* was run with the option to reorder both the rows and the columns of the alignment according to the  $C_r$  and  $C_c$  scores respectively, we can see that the resulting sets of heatmap are displayed nicely in a sorted manner.

The corresponding heatmap for partition *gene409969g*:

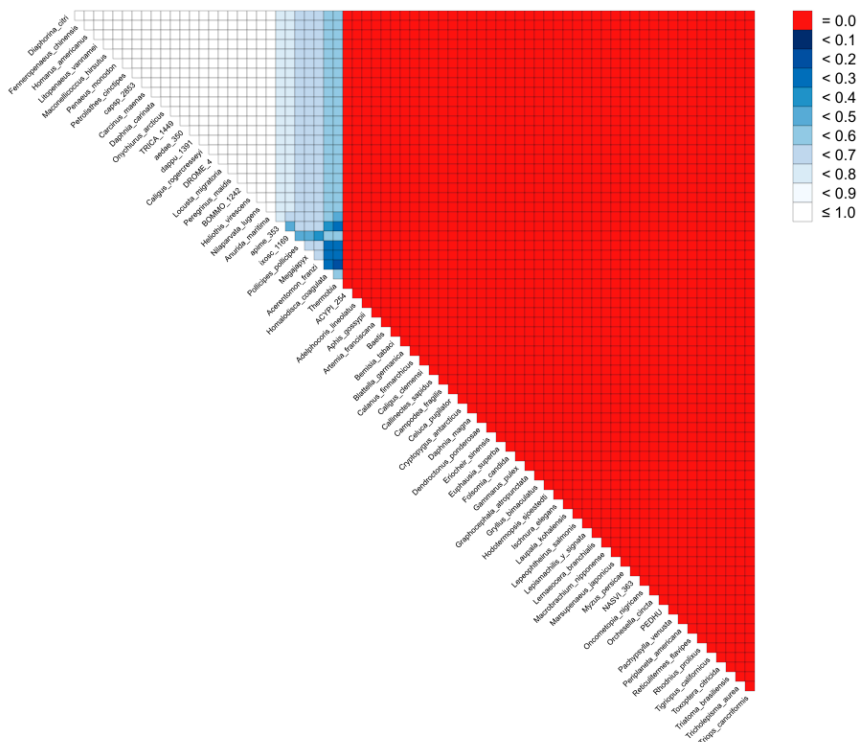


The corresponding heatmap for partition *gene409979g*:





The corresponding heatmap for partition *gene409986g*:



According to these sets of heatmap, there is a large proportion of black colour for the partitions *gene409969g* and *gene409979g*, which indicates that many pairs of sequences overlap completely with each other within these two partitions. On the other hand, for the partition *gene409986g*, there is a large proportion of red colour, which indicates that many pairs of sequences however do not have any overlap with each other within that partition.

## 6. Contact person

Dr Lars Jermiin

Email: [lars.jermiin@anu.edu.au](mailto:lars.jermiin@anu.edu.au)

Dr Thomas Wong

Email: [thomas.wong@anu.edu.au](mailto:thomas.wong@anu.edu.au)

Research School of Biology, Australian National University, ACT 2601, Australia

## 7. References

1. Dell'Ampio E., Meusemann K., U. Szucsich N., S. Peters R., Meyer B., Borner J., Petersen M., J. Aberer A., Stamatakis A., G. Walz M., Quang Minh B., von Haeseler A., Ebersberger I., Pass G., Misof B. (2013) Decisive Datasets in Phylogenomics: Lessons from Studies on the Phylogenetic Relationships of Primarily Wingless Insects. *Molecular Biology and Evolution* doi: 10.1093/molbev/mst196