# Documentation for **AliStat**

## Quantifying the Completeness of Multiple Sequence Alignments used in Phylogenetic and Phylogenomic Research

Copyright (C) 2014, CSIRO

April, 2020

## 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 a collaboration between Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, Australia, the University of Freiburg, Freiburg, Germany, and the Zoologica Research Museum A. Koenig ZFMK, Zentrum für Molekulare Biodiversitätsforschung (zmb), Bonn, Germany.

## Citation

Wong TKF, Kalyaanamoorthy S, Meusemann K, Yeates DK, Misof B, Jermiin LS. 2020. A minimum reporting standard for multiple sequence alignments. *NAR Genomics and Bioinformatics* 2 (2), lqaa024.

# Contents

# 1. Nomenclature

**Completely-specified characters**

Multiple sequence alignments may contain a variety of completely-specified characters. Here a completely-specified character is defined as a letter that has an unambiguous meaning (otherwise, the character is incompletely specified and will have several meanings, as described in the next sub-section).

The following completely-specified nucleotides are available: A, C, G, and T. In addition, the letter U may replace the letter T (to accommodate Uracil in RNA sequences).

The following completely-specified amino acids are available: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y.

These alphabets—the 4-state alphabet for DNA and the 20-state alphabet for protein—are identical to the IUPAC's codes for nucleotides and amino acids (IUPAC-IUB Joint Commission on Biochemical Nomenclature 1984; Cornish-Bowden 1985).

**Incompletely specified characters**

Multiple sequence alignments may contain incompletely-specified characters. An incompletely-specified character is defined as a letter that has an ambiguous biochemical meaning. Incompletely-specified characters are used in cases where it is clear that a character is present at a given site in the data but it is unclear which molecule it is.

The following incompletely-specified nucleotides are available: B, D, H, V, R, Y, K, M, S, W, and N. B represents C or G or T, implying that a nucleotide was found at a given site in the sequence or alignment but it is unclear whether it is a C, G, or T. Stated differently, it is clearly not an A. The meaning of these 11 letters is as follows: B = C or G or T, D = A or G or T, H = A or C or T, V = A or C or G, R = A or G, Y = C or T, K = G or T, M = A or C, S = C or G, W = A or T, and N = A or C or G or T.

The following four incompletely-specified amino acids are available: B, J, Z, and X. The meaning of these letters is as follows: B = N or D, J = L or I, Z = E or Q, and X = A or C or D or E or F or G or H or I or K or L or M or N or P or Q or R or S or T or V or W or Y.

**Other characters**

Often multiple sequence alignments contain a character representing an alignment gap (i.e., '–'). Alignment gaps are inserted (e.g., by multiple sequence alignment methods) to improve the homology statement for the individual sites.

The alignment gap should only be used to improve an alignment, while N (in nucleotide sequences) and X (in amino-acid sequences) should only be used to indicate incomplete information at the site.

Alignments of amino acids may sometimes contain O (Pyrrolysine) and/or U (Selenocysteine), but this is unusual for data encoded by natural genomes. When this happens, *AliStat* will treat O and U as X.

# 2. Metrics

Consider a multiple sequence alignment with $m$ sequences and $n$ sites. Given this alignment, we may compute:

**$\mathcal{C}_a$ - Completeness score for the alignment**

$$\mathcal{C}_a = \text{Number of unambiguous characters in the alignment}/(m \times n)$$

**$\mathcal{C}_r$ - Completeness score for individual sequences**

$$\mathcal{C}_r = \text{Number of unambiguous characters in a row (sequence) of the alignment}/n$$

**$\mathcal{C}_c$ - Completeness score for individual sites**

$$\mathcal{C}_c = \text{Number of unambiguous characters in a column (site) of the alignment}/m$$

**$\mathcal{C}_{ij}$ - Completeness score for pairs of sequences**

$$\mathcal{C}_{ij} = \text{Number of columns with homologous pairs of unambiguous characters in sequences } i \text{ and } j/n$$

By definition, $\mathcal{C}_{ii} = 1$.

**$\mathcal{I}_{ij}$ - Incompleteness score for pairs of sequences**

$$\mathcal{I}_{ij} = \begin{cases} 0 & i = j \\ 1 - \mathcal{C}_{ij} & i \neq j \end{cases}$$

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

Let $W_{ij}$ be the set of columns with unambiguous homologous characters of sequences $i$ and $j$.

$$P_{ij} = \big(\text{Number of sites with differences between sequences } i \text{ and } j \text{ across the columns in } W_{ij}\big)/W_{ij}$$

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

# 3. Data Types and Alphabets

Seven types of sequence data are considered: single nucleotides, di-nucleotides, codons, 10-state genotypes, 14-state genotypes, amino acids, and a mix of single nucleotides and amino acids.

The single nucleotides, di-nucleotides, codons, and mixture of nucleotides and amino acids use the nomenclature outlined above.

The 10- and 14-state genotype data are experimental in the sense that they have not yet been recognised by the wider scientific community as sound sources of genetic information pertaining to genomes of diploid and triploid species. Simulation-based research, led by LSJ, is currently under way to examine the usefulness of these types of data. How these types of data are generated may be revealed in due course.

When the sequences are regarded as

- single nucleotides, the data represent a 4-state alphabet (i.e., it comprises four letters: A, C, G, and T).
- di-nucleotides, the data represent a 16-state alphabet (i.e., it comprises 16 pairs of letters: AA, …, TT).
- codons, the data represent a 64-state alphabet (i.e., it comprises 64 triplets of letters: AAA, …, TTT).
- amino acids, the data represent a 20-state alphabet (i.e., it comprises 20 letters: A, …, Y).
- 10-state genotype data, the data represent a 10-state alphabet: A, C, G, T, R, Y, K, M, S, and W.
- 14-state genotype data, the data represent a 14-state alphabet: A, C, G, T, R, Y, K, M, S, W, B, D, H, and V.

When nucleotide sequences are interpreted as single nucleotides, the 4-state alphabet may be recoded into 3- and 2-state alphabets (Vera-Ruiz et al. 2014). Given a 4-state alphabet (A|C|G|T), the following 3- and 2-state alphabets may be recoded as (with different expressions of the same alphabet listed to the right):

- RCT = R|C|T A or G|C|T AG|C|T
- YAG = Y|A|G C or T|A|G CT|A|G
- KAC = K|A|C G or T|A|C GT|A|C
- MGT = M|G|T A or C|G|T AC|G|T
- SAT = S|A|T C or G|A|T CG|A|T
- WCG = W|C|G A or T|C|G AC|C|G
- RY = R|Y A or G|C or T AG|CT
- KM = K|M G or T|A or C GT|AC
- SW = S|W C or G|A or T CG|AT
- AB = A|B A|C or G or T A|CGT
- CD = C|D C|A or G or T C|AGT
- GH = G|H G|A or C or T G|ACT
- TV = T|V T|A or C or G T|ACG

In the list above, the character '|' is used to distinguish different units in the alphabet. Note that the order of the letters is flexible, implying, for example, that RCT = CRT.

## 4. Installation of AliStat

The software is written in C++, and has been tested under the Linux and MacOS platforms. You need to have C++ compiler installed in your computer to compile the source code. The compilation steps are shown as follows:

```
$ tar -zxvf AliStat_1.12.tar.gz

$ cd AliStat_1.12

$ make
```

Then an executable file named *alistat* will appear.

## 5. Usage of AliStat

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

In the second of these commands, the –h flag prints the help menu to the terminal.

| | |
|---|---|
| <alignment file> | Multiple sequence alignment file in FASTA format.<br><br>The alignment gap is represented by a dash (i.e. '–')<br><br>Example of a multiple nucleotide sequence alignment: test.fas:<br><br>>Seq_1<br>CGTACCTCTATTGTATGTTTGATGCCTCTATTGTATGTTTGAT-----<br>>Seq_2<br>-----CGTACATCTGTTGCATGATTAATGGATCTGTTGCATGATTAAT<br>>Seq_3<br>-----CGTACTTG----------CTCTATTGTATGTTTGAT<br>>Seq_4<br>-----CGTACATCTGTTGCATGATTAATGG----------<br>>Seq_5<br>CGTAC-----ATCTGTTGCATGATTAATGGCTCTATTGTATGTTTGAT |

| | |
|---|---|
| <data type> | The data type of the alignment file.<br><br>1 - Single nucleotides (SN);<br>2 - Di-nucleotides (DN);<br>3 - Codons (CD);<br>4 - 10-state genotype data (10GT);<br>5 - 14-state genotype data (14GT);<br>6 - Amino acids (AA);<br>7 - Mixture of nucleotides and amino acids (NA)<br><br>For 7, the user has to specify the data type for each partition inside a partition file (see option -p) |

**Other options:**

| | |
|---|---|
| –b | Print a brief summary of completeness scores to the terminal.<br><br>Output format:<br><br>`File, #seqs, #sites, Ca, Cr_max, Cr_min, Cc_max, Cc_min, Cij_max, Cij_min`<br><br>This option does not work with the following options: –o, –t, –r, –m, –i, –d. |
| –c <coding type> | The coding type in the alignment.<br><br>This option is only valid when <data type> = 1.<br><br>The nucleotides in the alignment will be converted into the corresponding characters according to the coding type selected by the user.<br><br>The following coding types are available:<br><br>`0 - A, C, G, T [default];`<br>`1 - C, T, R (i.e. C, T, AG);`<br>`2 - A, G, Y (i.e. A, G, CT);`<br>`3 - A, T, S (i.e. A, T, CG);`<br>`4 - C, G, W (i.e. C, G, AT);`<br>`5 - A, C, K (i.e. A, C, GT);`<br>`6 - G, T, M (i.e. G, T, AC);`<br>`7 - K, M (i.e. GT, AC);`<br>`8 - S, W (i.e. GC, AT);`<br>`9 - R, Y (i.e. AG, CT);`<br>`10 - A, B (i.e. A, CGT);`<br>`11 - C, D (i.e. C, AGT);`<br>`12 - G, H (i.e. G, ACT);`<br>`13 - T, V (i.e. T, ACG);` |
| –o <output prefix> | Prefix for output files.<br>(default: <alignment file> w/o .ext) |
| –n <FILE> | Only consider sequences with names listed in FILE.<br><br>Example of the file with a sequence name on each line:<br><br>`Seq_2`<br>`Seq_4`<br>`Seq_5` |

| | |
|---|---|
| –t <$n_1$,$n_2$,...> | Only print the tables $n_1$, $n_2$, ...<br><br>1 - Completeness scores for individual sequences ($\mathcal{C}_a$)<br>2 - Completeness scores for individual sites ($\mathcal{C}_r$)<br>3 - Distribution of completeness scores for individual sites ($\mathcal{C}_c$)<br>4 - Matrix with completeness scores for pairs of sequences ($\mathcal{C}_{ij}$)<br>(together with the heatmaps for $\mathcal{C}_{ij}$ scores)<br>5 - Matrix with incompleteness scores for pairs of sequences ($\mathcal{I}_{ij}$)<br>6 - Table with completeness scores ($\mathcal{C}_{ij}$) and incompleteness scores ($\mathcal{I}_{ij}$)<br><br>By default, AliStat does not output any table or heat map. For details about the tables, please refer to the section "Output of *AliStat*". |
| –p <FILE> | Specify the partitions in the *partition file.*<br><br>For <data type> = 1 - 6, the partition file format must be as follows:<br><br>"<partition name 1>=<start pos>-<end pos>, <start pos>-<end pos>, …"<br>"<partition name 2>= ... "<br><br>Example:<br><br>`part1=1-50,60-100`<br>`part2=101-200`<br><br>For <data type> = 7, the partition file format must be as follows:<br><br>"<SN/DN/CD/10GT/14GT/AA>, <partition name 1>=<start pos>-<end pos>, ..."<br>"<SN/DN/CD/10GT/14GT/AA>, <partition name 2>= ..."<br><br>Example:<br><br>`SN,part1=1-50,60-100`<br>`AA,part2=101-200`<br>`CD,part3=201-231` |
| –s <$n_3$,$n_4$> | Sliding window analysis: window size = $n_3$; step size = $n_4$<br><br>The first sliding window: [1, $n_3$]<br>The second sliding window: [1+$n_3$, $n_3$+$n_4$]<br>……<br><br>Each sliding window will be regarded as a partition.<br><br>This option cannot be used simultaneously with the –p option. |
| –r <row\|col\|both> | Reorder the rows/columns (or both) of the alignment according to the $\mathcal{C}_r$ or $\mathcal{C}_c$ scores.<br><br>All the tables will be displayed according to the reordered alignment.<br><br>To output the reordered alignment, please also use the –m option. |

| | |
|---|---|
| –m <$n_5$> | Mask the alignment; $0 \leq n_5 \leq 1$<br><br>Output the alignment with columns $\mathcal{C}_c \geq n_5$ in the file 'Mask.fst', the alignment with columns $\mathcal{C}_c < n_5$ in the file 'Disc.fst', and the alignment with an extra row in the first line to indicate whether the column is masked in the file 'Stat.fst'.<br><br>Special case: if no <*$n_5$*> is supplied, whole alignment is outputted in the file 'Mask.fst'; if <*$n_5$*> is 0, the alignment with columns $\mathcal{C}_c \geq 0$ is outputted in the file 'Mask.fst'.<br><br>If the –r option is used, the rows/columns (or both) of the outputted alignment are reordered according to the $\mathcal{C}_r$ or $\mathcal{C}_c$ scores. |
| –i <$n_6$> | Generate heat map for $\mathcal{C}_{ij}$ scores of sequence pairs.<br><br>1 - Triangular heat map<br>2 - Rectangular heat map<br>3 - Both<br><br>If no <*$n_6$*> is provided, then both triangular and rectangular heat map files are outputted. |
| –d | Report the p-distances between sequences in the file with extension '.p-dist.csv'<br>By default, this option is disabled.<br><br>Note: Computation of p-distances may take long time for large number of sequences.<br><br>For more information, please refer to the Section “Output of AliStat”. |
| –u | Colour scheme of the heatmaps<br><br>1 - Default colour scheme, suitable for colour-blind persons.<br>2 - Alternative colour scheme (not suitable for colour-blind persons). |
| –h | The help page. |

# 6. Output of AliStat

## 6.1 Table 1 - completeness scores for individual sequences ($\mathcal{C}_r$)

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

$\mathcal{C}_r$ is printed for each sequence in the alignment or, if partitions are used, the relevant positions of each sequence.

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

| Seq ID | Sequence | Sites | Cr |
|---|---|---|---|
| 1 | Seq_1 | 43 | 0.896 |
| 2 | Seq_2 | 43 | 0.896 |
| 3 | Seq_3 | 26 | 0.542 |
| 4 | Seq_4 | 25 | 0.521 |
| 5 | Seq_5 | 43 | 0.896 |

Note: The table is printed in csv format, with commas used as delimiter inside the file.

## 6.2 Table 2 - completeness scores for individual sites ($\mathcal{C}_c$)

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

$\mathcal{C}_c$ is printed for each column in an alignment or, if partitions are used, the relevant positions in the partitions specified.

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

```
Site ID       Cc
1             0.4
2             0.4
3             0.4
4             0.4
5             0.4
6             0.8
7             0.8
8             0.8
9             0.8
10            0.8
.....
```

Note: The table is outputted in csv format, with commas used as delimiter inside the file.

## 6.3 Table 3 - Distribution of C scores for individual sites ($\mathcal{C}_c$)

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 and cumulative distribution of $\mathcal{C}_c$ scores.

Referring to the alignment example before, the table 3 is organised as follows (X represents the number of ambiguous characters at a given site):

```
X      Cc     Sites  Proportion    Cumulative
0      1      3      0.063         0.063
1      0.8    35     0.729         0.792
2      0.6    5      0.104         0.896
3      0.4    5      0.104         1
4      0.2    0      0             1
5      0      0      0             1
```

Note: The table is outputted in csv format, with commas used as delimiter inside the file.

## 6.4 Table 4 - Matrix with C scores for pairs of sequences ($\mathcal{C}_{ij}$)

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

For each pair of sequences in the alignment or each partition (in case a partition file is provided), $\mathcal{C}_{ij}$ is printed in a matrix. By definition, $\mathcal{C}_{ii}$ is always 1.

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

```
5
Seq_1  1      0.792  0.438  0.521  0.792
Seq_2  0.792  1      0.542  0.521  0.792
Seq_3  0.438  0.542  1      0.167  0.438
Seq_4  0.521  0.521  0.167  1      0.417
Seq_5  0.792  0.792  0.438  0.417  1
```

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

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

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. The order of the columns is always the same as that of the rows.

## 6.5 Table 5 - Matrix with incompleteness scores for pairs of sequences ($\mathcal{I}_{ij}$)

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

The table with the matrix of incompleteness scores is derived from Table 4 (by subtracting $\mathcal{C}_{ij}$ from 1.0.

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

```
5
Seq_1  0      0.208  0.563  0.479  0.208
Seq_2  0.208  0      0.458  0.479  0.208
Seq_3  0.563  0.458  0      0.833  0.563
Seq_4  0.479  0.479  0.833  0      0.583
Seq_5  0.208  0.208  0.563  0.583  0
```

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

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

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 columns is always the same as that of the rows.

The format of this file is suitable as a distance matrix for various phylogenetic software tools.

## 6.6 Table 6 - Table with completeness scores ($\mathcal{C}_{ij}$) and incompleteness scores ($\mathcal{I}_{ij}$)

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

In this table, the values of $\mathcal{C}_{ij}$ and $\mathcal{I}_{ij}$ are listed jointly as a function of the pairs of sequences considered.

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

```
taxon 1       taxon 2       Cij    Iij
Seq_1         Seq_1         1      0
Seq_1         Seq_2         0.792  0.208
Seq_1         Seq_3         0.438  0.563
Seq_1         Seq_4         0.521  0.479
Seq_1         Seq_5         0.792  0.208
Seq_2         Seq_1         0.792  0.208
Seq_2         Seq_2         1      0
Seq_2         Seq_3         0.542  0.458
Seq_2         Seq_4         0.521  0.479
…….
```

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

## 6.7 Summary statistics of the alignment

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

This file shows the summary statistics for the alignment of the partitions listed in the partition file. The file presents all details needed to comply with the minimum reporting standard (MRS) for multiple sequence alignments.

Example of the summary statistics for the example file (test.fas):

```
Name of input file ..................................... test.fas
Type of data ........................................... Single nucleotides (SN)
Characters used ........................................ ACGTU
Number of sequences in the alignment    ................ 5
Number of sites in the alignment    .................... 48
Pairs of sequences ..................................... 10
Completeness (C) score for the alignment    (Ca) ...... 0.78
Maximum C-score for individual sequences (Cr_max) ...... 0.896
Minimum C-score for individual sequences (Cr_min) ...... 0.521
Maximum C-score for individual sites (Cc_max) ........... 1
Minimum C-score for individual sites (Cc_min) ........... 0.4
Maximum C-score for pairs of sequences (Cij_max, i!=j) ..... 0.792
Minimum C-score for pairs of sequences (Cij_min, i!=j) ..... 0.167

C scores for individual sequences (Cr) stored in ....... Table_1.csv
C scores for individual sites (Cc) stored in ........... Table_2.csv
Distribution of C scores for individual sites (Cc) in ..... Table_3.csv
Matrix with C scores for individual pairs of seqs (Cij) in .... Table_4.dis
Matrix with incompleteness scores (Iij = 1.0 - Cij) in ..... Table_5.dis
Table with C scores (Cij) and incompleteness scores (Iij) in .. Table_6.csv
```

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

This table is printed only when the –d option is specified.

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

Referring to the example before, the table with p-distances is organised as follows:

```
5
Seq_1  0      0.816  0.524  0.84   0.632
Seq_2  0.816  0      0.269  0      0.132
Seq_3  0.524  0.269  0      0.25   0.095
Seq_4  0.84   0      0.25   0      0
Seq_5  0.632  0.132  0.095  0      0
```

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

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

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. The order of the columns is always the same as that of the rows.

## 6.9 Masked and discarded (reordered) alignment

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

The alignment with columns $\mathcal{C}_c \geq n_7$ will be printed in the file '*Mask.fst*' (or '*[partition name].Mask.fst*' if a partition file is used).

The alignment with columns $\mathcal{C}_c < n_7$ will be printed in the file '*Disk.fst*' (or '*[partition name].Disc.fst*' if a partition file is used).

If the –r option is used, then the rows/columns (or both) of alignment will be sorted according to the $\mathcal{C}_r$ or $\mathcal{C}_c$ (or both) scores.

The file '*Stat.fst*' (or '*[partition name].Stat.fst*' if a partition file is used) contains the original alignment with an extra first line telling users which columns were included ("I") in the masked alignment and which columns were excluded ("E") from the original alignment.

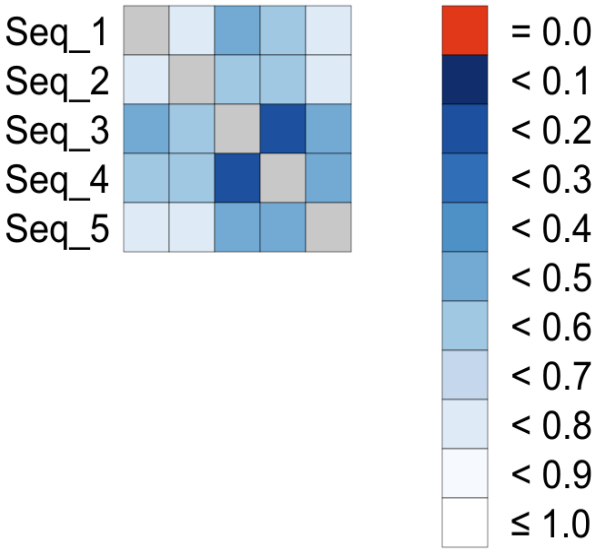
Referring to the example before, when using the –m 0.7 option, the file '*Stat.fst*' will look as follows:

```
>Mask
EEEEEIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIEEEEE
>Seq_1
CGTACCTCTATTGTATGTTTGATGCCTCTATTGTATGTTTGAT-----
>Seq_2
-----CGTACATCTGTTGCATGATTAATGGATCTGTTGCATGATTAAT
>Seq_3
-----CGTACTTG-----------------CTCTATTGTATGTTTGAT
>Seq_4
-----CGTACATCTGTTGCATGATTAATGG------------------
>Seq_5
CGTAC-----ATCTGTTGCATGATTAATGGCTCTATTGTATGTTTGAT
```

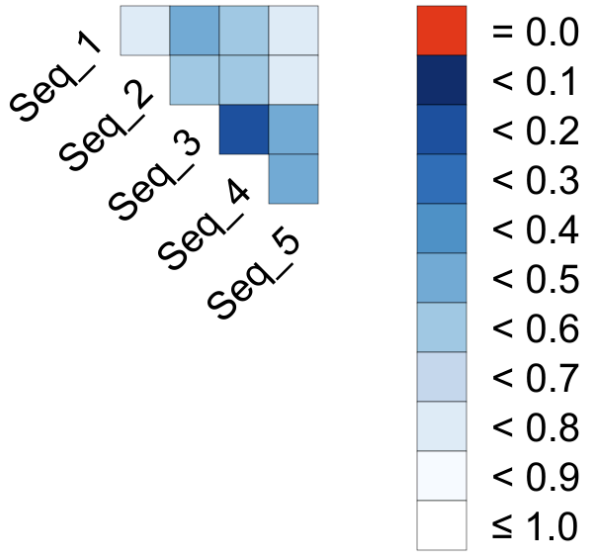
## 6.10 Heatmap

The heatmap files can visually display the $\mathcal{C}_{ij}$ 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 example before, the following image is the rectangular heatmap for table 4:




The corresponding triangular heatmap table is:



# 7. More examples

Sequence alignment file: M_Ento.fas (Source: http://doi.org/10.5061/dryad.mk8p7: Dell'Ampio et al. 2014)

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.

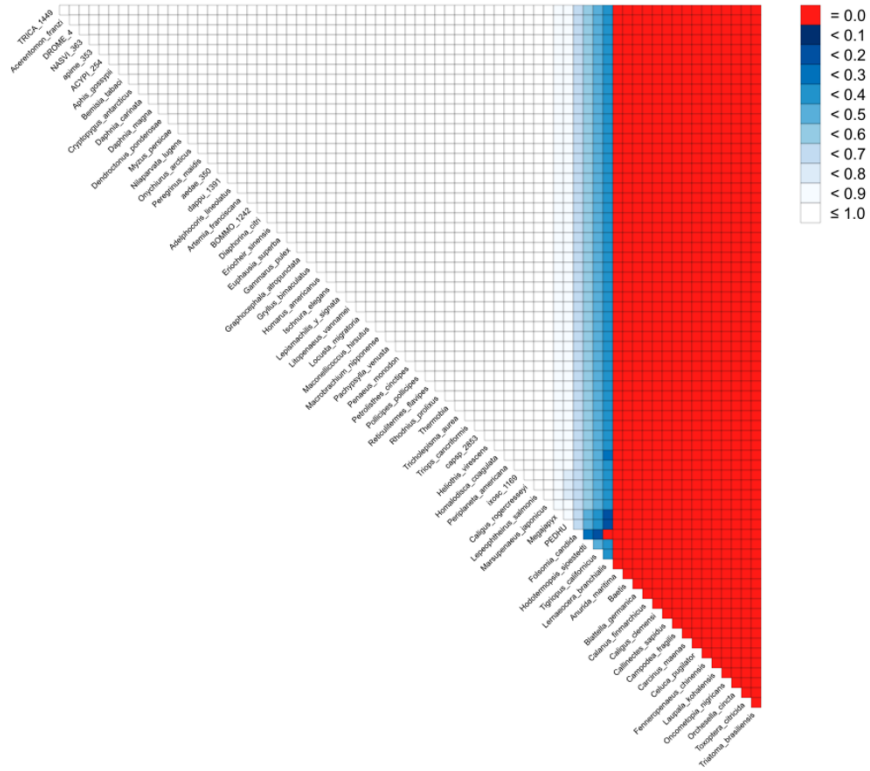
According to the summary files, the completeness scores of the partitions are:

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

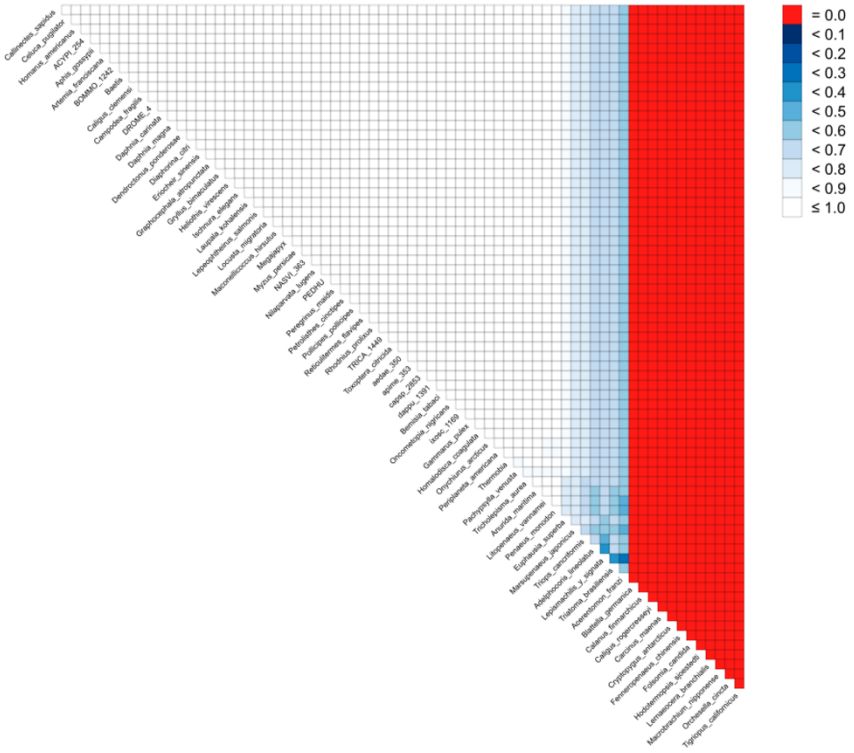
It indicates that the partition *gene409979g* has highest level of completeness, while the partition *gene409986g* has lowest level of completeness.

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

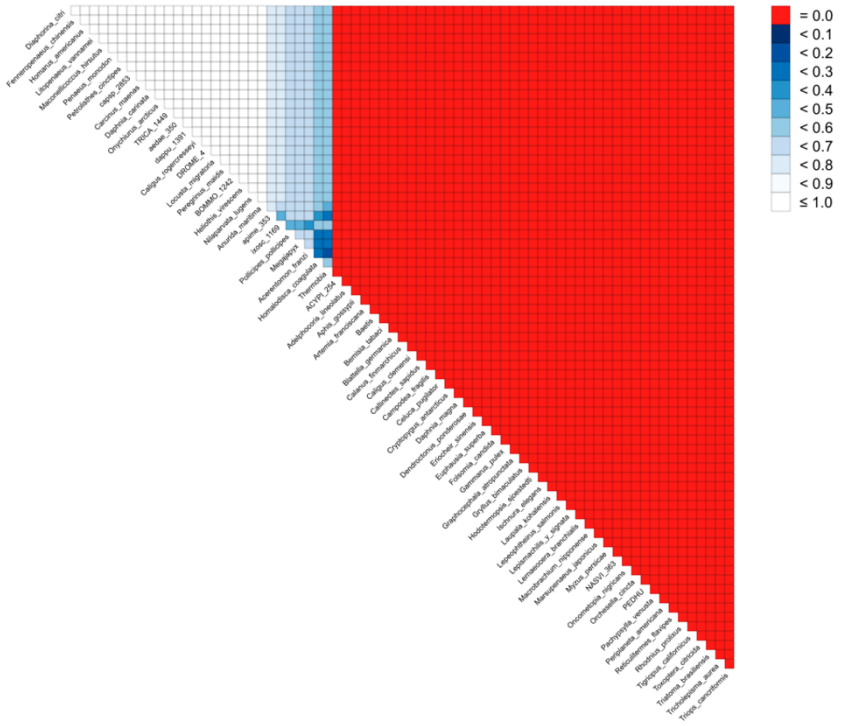
The corresponding heatmap for partition *gene409969g*:




The corresponding heatmap for partition *gene409979g*:



The corresponding heatmap for partition *gene409986g*:




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

## 8. Contacts

Dr. Lars S. Jermiin (Email: lars.jermiin@anu.edu.au)

Dr. Thomas Wong (Email: thomas.wong@anu.edu.au)

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

## 9. References

Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research* 13, 3021-3030.

Dell'Ampio E, Meusemann K, Szucsich NU, Peters RS, Meyer B, Borner J, Petersen M, Aberer AJ, Stamatakis A, Walzl MG, Minh BQ, von Haeseler A, Ebersberger I, Pass G, Misof B. 2014. Decisive datasets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. *Molecular Biology and Evolution* 31, 239-249.

IUPAC-IUB Joint Commission on Biochemical Nomenclature. 1984. Nomenclature and symbolism for amino acids and peptides: Recommendations 1983. *European Journal Biochemistry* 138, 9-37.

Vera-Ruiz VA, Lau KW, Robinson J, Jermiin LS. 2014. Statistical tests to identify appropriate types of nucleotide sequence recoding in molecular phylogenetics. *BMC Bioinformatics* 15 (Suppl 2), S8.