Aliscore - Manual

Bernhard Misof,[1] and Patrick Kück[2*]

[1]Zoologisches Forschungsmuseum A. Koenig,
Zentrum für molekulare Biodiversitätsforschung (zmb)
Adenauerallee 160, 53113 Bonn, Germany
[2]Zoologisches Forschungsmuseum A. Koenig Bonn, Bioinformatics
Adenauerallee 160, 53113 Bonn, Germany

[*]E-mail: aliscore@gmx.de

22nd February 2012

Contents

1 Introduction

Aliscore is designed to filter alignment ambiguous or randomly similar sites in multiple sequence alignments (MSA). It does not generate a generic alignment, this must be provided by the user. Aliscore reads exclusively alignments in FASTA format independently of suffices (.fas .txt .fts etc.). Aliscore reads the alignment and generates a hash of these sequences with taxon names as keys and simple sequence arrays as values. It works on these hash elements and uses these hash elements as the basic data. Aliscore tolerates newlines in sequences but not in taxon names. Sequences must be of similar length! Aliscore can not read sequences in interleaved format, but this does not correspond to a plain fasta file anyway. Blanks in sequences are ignored, any other sign in sequences except for these covered by the universal DNA/RNA code and '?' will chock the program. Ambiguities are understood, as are indels. Kapital or small letters are equally good as input and can be used interchangeably, RNA and DNA sequences can be used in one alignment, RNA sequences are translated into DNA sequences. Aliscore works on WindowsPCs, Macs and Linux mashines, but was written on Linux. If input files are coming from Windows make sure CRFL feeds are removed. Aliscore tries to remove them, but my not succeed in every instance. Taxon names must only include alphanumeric signs, underscores (_) and blanks, everything else might chock the program. Aliscore will issue an error prompt and die if any non-alphanumeric sign is encountered in taxon names. If used with the outgroup option avoid blanks in names as this might lead to erroneous recognition of taxon names. Aliscore will write results into folder where Aliscore is located. It will produce three files, one file with the consensus Profile, a corresponding vector graphic (.svg) and one file with a list of characters with negative scores in this profile.

Example of an input file:

>Podura aquatica 18S_1
aaagtctgtgacgttgtacggact
gcgtgtgcagctgtgacgggcgcc
>Sminthurus_sp
autgctugccguuugaucgugugc
uuggacugcgucgatcguugcgcg

2 Usage/Options

For using Aliscore open the terminal of your run system, move through your directory path to the Aliscore including folder and type the name of your Aliscore version, followed by a blank and your demand options, all in one row.

For example:

C:\Folder_of_Aliscore> Name_of_Aliscore_Version.pl -i *inputfile.fas*

Aliscores knows several options, it chocks if an unknown option is encountered. Make sure you write the input options correctly, for example "-w 4" and not "-w4" or "-w_4",

etc., likewise do not (!) use “-in infile“, or “in infile“ or “-i_infile“; these are all wrong input formats and will cause the program to die. It will still try to open an “n infile“ or “_infile“ which is hopefully not present, it will also tell you this.

2.1 -h option

-h option: with the -h option Aliscore delivers a short introduction about usage format and adaptable option codes. For detailed help on options type “help“, followed by the demand option code.

For example:

C:\Folder_of_Aliscore> Name_of_Aliscore_Version.pl help -i

For deeper explanations about output format, scoring scheme and commands type as the case maybe “help -output“, “help -scoring“ or “help -commands“.

2.2 -N option

-N option: without invoking the -N option gaps are treated as 5th character. With the -N option invoked gaps are treated as ambiguous character. Leading and tailing gaps of sequences are always interpreted as ambiguous characters with and without the -N option. Interpreting gaps as ambiguous characters results in a loss of long indel sections consistently found in the majority of taxa. This means that well aligned expansion segments in rDNA sequences, which are not present in other taxa will be lost, if not commonly found in the MSA. Interpreting gaps as 5th character interprets stretches of indels as well aligned sections. This option is currently only implemented for sequences on a nucleotide level.

2.3 -w option

-w # option: without invoking this -w # option, Aliscore will use the default window size of 6 for the sliding window. You may choose any other window size, smaller or larger, but it does not make sense to choose something smaller then 4. If you vote for a much larger window size then 6, Aliscore will become successively blind for small stretches of randomly similar sections. (See paper on Aliscore performance). If you vote for window size <4 Aliscore will start making substantial type I errors and call non-randomly similar sites randomly similar, depending on its neighbors.

2.4 -r option

-r # option: if -r is used without an argument 4*N random pairs are compared, checking for replications (which are avoided). If -r is used with an argument, this number of randomly selected pairs is analysed and used to infer the consensus profile, if -r used used with an argument which is beyond the maximal number of possible non-overlapping pairs, only the maximal number of pairs is compared. If the -r option and the -t option are not used, random pairs are compared as default, with 4*N selection of pairs.

2.5 -t option

-t “treefile“ option: -t must be used with a tree file in Newick format, rooted or unrooted. The tree file should be in the same folder as the sequence file (not mandatory). If there are more than one tree in the tree file, only the first one will be read, all other trees will be ignored. The tree file does not contain any polytomic splits. Otherwise Aliscore breaks up yet. Aliscore will read the tree and store as a hash with node levels as keys and taxa as values for each node. Aliscores uses this tree to work through the MSA from tips to bottom of tree. First, sister groups of terminal taxa are identified (node lists, level 1 as key) and compared, these taxa are then replaced by consensus sequences using the ambiguity code. Consensus sequences represent now the new set of terminal taxa with which Aliscore proceeds. This process is repeated until every possible pair of sequences within the tree is evaluated. Make sure that your tree does not contain CRFL from Windows if working on Linux!

2.6 -l option

-l # option: -l can be used to restrict iterating through the tree to a specific node level, specified with the argument at the -l option. If -l 1 is used only primary sister group relationships are used to infer the consensus profile. If there are less node levels then arguments, Aliscore iterates through the tree and stops.

2.7 -s option

-s option: -s option can be used to generate a strict profile from all single comparisons. This profile will be very conservative because it scores every site as negative which exhibits a negative score in one single profile already. This option does not make to much sense, do not use it on purpose!

2.8 -o option

-o “taxon,taxon,..“ option: the -o option is used with a list of taxa separated by commatas. These taxa will be compared with all other taxa not in this list, but not with each other. It can be used to assess the range of randomness between outgroup taxa and ingroup taxa, or between every two groups of interest, if the alignment is restricted to ingroup taxa only before analysis.

3 Internals

Details and comments are given in order of its appearance in code.

3.1 Input

Input arguments are collected into a 1-dimensional array and grep is used to retrieve options plus arguments; white spaces are cleaned off, and array is created by splitting input string at -; If you use taxon names with white spaces in -o option you might run into problems.

For example:

```
our ($file)=grep /^i.+/,@INPUT; $file=~s/(^i)//;
```

3.2 Reading FASTA file

Fasta file is read and stored as a hash with taxon names as keys and references to sequence arrays as values. Sequences are stored as flat list, each position constituting an element. Only references to these hash elements are returned from the subroutine. The reference to the hash is used as a global variable indicated by "our", only the file name is used as argument for the subroutine to open and read the file; will die if file has not bee found. Aliscore understands DNA ambiguity code, there is no need to replaces these. Aliscore does not accept any sign except letters and indels or '?' in sequences. It will die if anything else is encountered in sequences.

Command:

```
our ($ref_FASTA)=Alignment_alpha::readFASTA_simple($file);
```

number of taxa and taxon names are collected into an array for later comparison Aliscore attempts to estimate the data type, either nucleotide or amino acid data. Aliscore considers sequences with an ACTG content of > 0.8 (without counting indels and N) as nucleotide sequences, if less then 0.8 as amino acid data. It estimates data property from every sequence, if two sequences are considered of different data type, Aliscore stops. Aliscore might stop if a single nucleotide sequence contains more then 0.2 ambiguities. In almost every case, Aliscore will correctly estimate data type, if it does not, it will stop and report on the problem. If the data contains sequences of more then 0.2 ambiguities, it might be advisable to recode ambiguities as N's or remove the particular sequence. RNA sequences will be recoded to DNA sequences. Nucleotide data can be a mix of RNA/DNA data.

3.3 Reading data type and scoring matrix

Reads data type and generates accordingly scoring matrix. In case of nucleotide data, the scoring matrix is a simple match mismatch scoring matrix, in case of ambiguous characters the mismatch is optimistically interpreted, thus Aliscore can also handle RY-encoded data on nucleotide level. NOTE: If the amount of RY in nucleotide data exceeds X% Aliscore understands the alignment as aminoacid data. This behavior will be changed in future editions. If an alignment is RY-encoded (containing $\geq 80\%$ R, Y, without N, -, ?), Aliscore interprets the alignment as nucleotide (RY recoded) data. In this case, scoring is currently only possible based on random pairs. If indels are considered 5th characters, they are scored in a mismatch/match pattern. A BLOSUM62 is used for the amino acid scoring with indels and X scoring 0. For aminoacid scoring, a Monte Carlo approach is used to generate a threshhold value, given the actual window size and aminoacid composition of the data.

The -e option refers to special aminoacid scoring. if -e option is used, matching indels are penalized but not aminoacid and indel matches. This favors sections of the alignment,

in which aminoacids are indeed present, but not dominating the signal. A biological interpretation is not straightforward, but given data from EST projects or phylogenomic data, in which often parts of genes are missing, ALISCORE is less restrictive and favors information from aminoacid data.

3.4 Checks for mutiple identical sequences

Aliscore checks for potential identical sequences. It considers sequences which can be a subset of another longer sequence, ignoring N, potentially identical. Only the longer sequence, not considering N's, will be retained for the analyses. If there are mutiple potentially identical sequences, only the one with the most inclusive sequence will be retained. Aliscore does not concatenate sequences, even if potentially profitable. Results are reported to the terminal.

3.5 Checking variability

Checks for invariant sections across the alignment with an extension of >w+2 (w window size). Reports these sections and places information as an argument into subroutine later. This step improves speed, because only variable sections are actually scored for random similarity. A simple iteration through all sequence arrays is used to check variability of sites. A @temp array is used to create the list of variable sections, results are reported to terminal

3.6 Scoring of randomly selected sequence pairs

The code reads the $random parameter and if defined with a number, which should have happened in any case except for the situation where the -t option and the -l option was evoked starts the random selection process. It first generates al possible pairs from the list of taxon names. It then checks whether the -o option was provided with an argument, if this is the case it fills a pairs list with all comparisons between outgroup taxa and ingroup taxa, if the -o option was not provided with an argument, it checks the argument of the -r option and selects as many random unique pairs from the list of all possible pairs. If the argument of -r was too large it stops when all possible pairs are included.

3.6.1 Scoring

For each entry in the pairs list, it uses the two taxon names to look in the data hash for both sequences and uses the scoring type, flat list of variable characters and both sequence references as arguments. All arguments are provided as references. The scoring profile is returned as a reference. Description of the scoring process see Alignment_alpha.pm. The list of arguments must be in order, reference to the scoring type must be first.

3.7 Scoring using tree based selection of pairwise comparisons

A user provided tree, rooted or unrooted, but fully dichotomous must be provided by the user. This tree is used for selection of sequence pairs. First, terminal sister taxa are compared, then these sequence pairs are replace by one consensus sequence. Consequently,

the next set of terminal sequence pairs might contain consensus sequences and primary sequences. Consensus sequences make uses of the full ambiguity code to represent every difference in primary parent sequences. The scoring stops when the last sequence pair has been analysed.

3.7.1 Reading Tree

Tree must be in Newick format. Be careful, PAUP saves trees with basal polytomy as default. If these trees are used, Aliscore breaks up without an error message yet. Save trees without basal polytomies in PAUP and everything will be fine. Check set options in PAUP! You can use rooted trees, either rooted in PAUP or any other software package, and everything should be fine. Take care to check taxon names in trees, because only if these names correspond exactly (!) to names in sequence files, scoring will be performed. Aliscore will have its own tree reconstruction routine soon, to avoid problems of incongruent taxon names and polytomies.

3.7.2 Removing potentially identical taxa

Similarly to random pair selection, Aliscore removes potentially identical sequences in tree base selection of sequence pairs.

3.8 Generation of Consensus Profiles

From the collection of single profiles a consensus profile is generated. The consensus profile consists of medians for each site derived from site scores of all single profiles. It is thus a consensus representation of the situation in single profiles. Aliscore generates a List of all characters of the consensus profile below the 0 - base line. This list is written into a list file. Additionally, Aliscore writes a profile file in which three columns are written. First column, an enumeration of positions, second column sites with positive consensus values and third column sites with negative consensus values. Alternative consensus techniques would be conceivable, but the median certainly reflects the dominating mode among single profiles. Single profiles are collected into a temporary array, before a consensus profile will be generated. If the number of taxa > 200 and/or length of sequences > 8000 the process might crash because of RAM limits. This must be corrected in the near future, to avoid problems with very large data.

4 Copyright

5 License

e-mail: aliscore@gmx.de
Homepage: http://aliscore.zfmk.de

Randomly similar aligned sections can be discraded by ALICUT. To download ALICUT and other free tools visit: http://software.zfmk.de

6 Citation

1. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Misof B, Misof K. *Systematic Biology* **2009** Feb; 58(1):21-34. DOI: 10.1093/sysbio/syp006.

2. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wägele JW, Misof B. *Frontiers in Zoology* **2010**;710. DOI: 10.1186/1742-9994-7-10.

7 Acknowledgements

We thank Caro Greve and Karen Meusemann for comments on the Manual.