What is this tool used for?

In a multiple sequence alignment (MSA) sometimes appear sequences that phylogenetically diverge markedly from the rest. Those sequences can influence negative on the MSA. This tool evaluate the alignment by means of determining the weight each sequence have over the quality of the whole MSA.

In short alignments is easy to identify those divergent sequences. They may stand out from the others for having a high number of gaps, inducing more gaps in the rest of sequences or having mismatches.

This tool not only scores specifically each sequence, but also identifies those who induce more gaps in the others and those having higher number of gaps.

Prerequisites

In Linux, the program interacts with the R statistical language in order to print the results. We should have installed the R modules needed for graphic representations to have the results presented properly. In Windows, due to some packages incompatibilities, graphs are represented with the GD::Graph perl module.

Usage instructions

Input Data

When running the program, the user must introduce some data:

1. MSA file path. If the file is located in the same directory where the program is executing just the name of the file is necessary. The program uses the AligIO package from BioPerl, so usual MSA formats are supported (FASTA, PHYLIP, ClustalW,...). You can take advice of the different formats supported by looking at http://www.bioperl.org/wiki/Multiple_alignment_formats.
2. Scoring matrix file path. If the file is located in the same directory where the program is executing just the name of the file is necessary. The file must be a plain-text file, with the information displayed as showed:

“

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3
```

```
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4
”
```

Pay attention to the fact that each row (except the first one) started with the abbreviation of the corresponding aminoacid or nucleotide. Besides, the white spaces between each numeric vale are important (2 spaces before positive values and 1 before negative ones). In case of having any doubt, just examine the scoring matrix files provided with this material.

3. Reference sequences (outgroups). Name of the sequences that the program must obviate in the analysis, separated by commas. By default, none sequence is set.

4. Gap penalty. A numeric value used for scoring each gap position. If none is set, the program will calculate a value that is equal to the minimum score found in the scoring matrix minus the standard deviation of all the scores of this matrix.

5. Gap threshold. A numeric value, between 0 and 1. Is the relative percentage of gaps that must be in a column of the MSA in order the program identifies the sequences that in this position have an aminoacid/nucleotide as being “gap generators”. For example, if 0.66 is set in all the columns with 66% or more percentage of gaps the tool will identify those sequences with aminoacid/nucleotides as gap generators, and will penalize them. The default value is 0.66.

On the other hand, the input data can be given to the program as execution arguments:

./EvalMSA [MSA_path] [matrix_path] [OPTIONAL ARGUMENTS] *(Linux)*
EvalMSA.exe [MSA_path] [matrix_path] [OPTIONAL ARGUMENTS] *(Windows)*

ARGUMENTS:

-o GENE1,GENE2,GENE3...
Each of the reference genes, separated by commas.
-p VALUE
Gap penalty value
-t VALUE
Gap threshold value

If the optional arguments are not specified, the program will use the default values.

Output Data

Before the execution, three different files are created

- *MSA_file*_results.txt: Include the pre-analysis data, the alignment analysis data and the

potential errors.

- Pre-analysis: Shows the average, median, standard deviation, quartiles, percentiles and the outliers taking as data the original length of the sequences.
- Alignment analysis: Shows information related with SP method and the alignment score calculated. Points out the less scored sequence in the alignment, the sequence with higher number of gaps and the sequence who induces more gaps in the rest of the sequences.
- Warning: In case of finding the symbols *, ?, or any other that is not included in the scoring matrix, list the sequence and the position associated.

• MSA_file_output.pdf: Contains some plots generated with R, in the Linux version.

- **Pre-analysis boxplot**: A boxplot generated with the original length of the sequences (before aligning).
- **Weight histogram:** A histogram that represents the scores distribution of the sequences. In yellow is identified the place where the sequence with a higher number of gaps is set. Also in magenta is identified the place of the sequence who induces more gaps in the rest.
- **Normalized weigtht plot:** A line plot with the weight values sorted and normalized, relative to the maximum weight score. If sequences have similar weights, a straight line is expected. On the other hand, if there are sequences with different influences in the alignment quality, some turning point would appear. Differences in the slope of the line may be caused by more than one cluster of sequences with similar weight.
- **Normalized quality variation plot:** Shows the variation of the alignment quality (normalized) by removing an increasing number of sequences, sort by their weight value (low value first). Again, we expected a constant decay. If the inclination of the slope varies, we can assume that there are sequences whose influence over the alignment are different to the rest of the sequences. It could be also produced by clusters of sequences with a good intra-cluster alignment but bad inter-cluster as well. The maximum score is set to 1, and the green point is the original relative score of the alignment (without removing any sequence).
- **Gapiness value histogram:** Similar to the weight histogram, but with the gapiness value represented.
- **Gapiness plot:** A line plot with the gapiness values sorted.

• PNG graph files: Several .png files containing some plots generated with GD::Graph, in the Windows version.

- **Pre-analysis boxplot**: A boxplot generated with the original length of the sequences (before aligning).
- **Normalized weight plot:** A line plot with the weight values sorted and normalized, relative to the maximum weight score. If sequences have similar weights, a straight line is expected. On the other hand, if there are sequences with different influences in the alignment quality, some turning point would appear. Differences in the slope of the line may be caused by more than one cluster of sequences with similar weight.
- **Normalized quality variation plot:** Shows the variation of the alignment quality (normalized) by removing an increasing number of sequences, sort by their weight value (low value first). Again, we expected a constant decay. If the inclination of the slope varies, we can assume that there are sequences whose influence over the alignment are different to the rest of the sequences. It could be also produced by clusters of sequences with a good intra-cluster alignment but bad inter-cluster as well. The maximum score is set to 1, and the green point is the original relative score of the alignment (without

removing any sequence).

- **Gapiness plot:** A line plot with the gapiness values sorted.

• *MSA_files*_values.csv. A .csv file that contains the number of gaps, the raw weight value and the gapiness value of all the sequences of the alignment.