NAME

InfoSequenceFiles.pl - List information about sequence and alignment files

SYNOPSIS

InfoSequenceFiles.pl SequenceFile(s) AlignmentFile(s)...

InfoSequenceFiles.pl [-a, --all] [-c, --count] [-d, --detail infolevel] [-f, --frequency] [--FrequencyBins number | "number, number, [number,...]"] [-h, --help] [-i, --IgnoreGaps yes | no] [-l, --longest] [-s, --shortest] [--SequenceLengths] [-w, --workingdir dirname] SequenceFile(s)...

DESCRIPTION

List information about contents of *SequenceFile(s) and AlignmentFile(s)*: number of sequences, shortest and longest sequences, distribution of sequence lengths and so on. The file names are separated by spaces. All the sequence files in a current directory can be specified by **.aln*, **.msf*, **.fasta*, **.fta*, **.pir* or any other supported formats; additionally, *DirName* corresponds to all the sequence files in the current directory with any of the supported file extension: *.aln, .msf, .fasta, .fta, and .pir*.

Supported sequence formats are: *ALN/CLustalW*, *GCG/MSF*, *PILEUP/MSF*, *Pearson/FASTA*, and *NBRF/PIR*. Instead of using file extensions, file formats are detected by parsing the contents of *SequenceFile(s) and AlignmentFile(s)*.

OPTIONS

-a, --all

List all the available information.

-c, --count

List number of of sequences. This is default behavior.

-d, --detail *InfoLevel*

Level of information to print about sequences during various options. Default: *1*. Possible values: *1, 2 or 3*.

-f, --frequency

List distribution of sequence lengths using the specified number of bins or bin range specified using FrequencyBins option.

This option is ignored for input files containing only single sequence.

--FrequencyBins *number* | *"number,number,[number,...]"*

This value is used with -f, --frequency option to list distribution of sequence lengths using the specified number of bins or bin range. Default value: *10*.

The bin range list is used to group sequence lengths into different groups; It must contain values in ascending order. Examples:

```
100,200,300,400,500,600
200,400,600,800,1000
```

The frequency value calculated for a specific bin corresponds to all the sequence lengths which are greater than the previous bin value and less than or equal to the current bin value.

-h, --help

Print this help message.

-i, --IgnoreGaps *yes* | *no*

Ignore gaps during calculation of sequence lengths. Possible values: *yes or no*. Default value: *no*.

-l, --longest

List information about longest sequence: ID, sequence and sequence length. This option is ignored for input files containing only single sequence.

-s, --shortest

List information about shortest sequence: ID, sequence and sequence length. This option is ignored for input files containing only single sequence.

--SequenceLengths

List information about sequence lengths.

-w, --WorkingDir *dirname*

Location of working directory. Default: current directory.

EXAMPLES

To count number of sequences in sequence files, type:

```
% InfoSequenceFiles.pl Sample1.fasta
% InfoSequenceFiles.pl Sample1.msf Sample1.aln Sample1.pir
% InfoSequenceFiles.pl *.fasta *.fta *.msf *.pir *.aln
```

To list all available information with maximum level of available detail for a sequence alignment file Sample1.msf, type:

```
% InfoSequenceFiles.pl -a -d 3 Sample1.msf
```

To list sequence length information after ignoring sequence gaps in Sample1.aln file, type:

```
% InfoSequenceFiles.pl --SequenceLengths --IgnoreGaps Yes
  Sample1.aln
```

To list shortest and longest sequence length information after ignoring sequence gaps in Sample1.aln file, type:

```
% InfoSequenceFiles.pl --longest --shortest --IgnoreGaps Yes
  Sample1.aln
```

To list distribution of sequence lengths after ignoring sequence gaps in Sample1.aln file and report the frequency distribution into 10 bins, type:

```
% InfoSequenceFiles.pl --frequency --FrequencyBins 10
  --IgnoreGaps Yes Sample1.aln
```

To list distribution of sequence lengths after ignoring sequence gaps in Sample1.aln file and report the frequency distribution into specified bin range, type:

```
% InfoSequenceFiles.pl --frequency --FrequencyBins
  "150,200,250,300,350" --IgnoreGaps Yes Sample1.aln
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

AnalyzeSequenceFilesData.pl, ExtractFromSequenceFiles.pl, InfoAminoAcids.pl, InfoNucleicAcids.pl

COPYRIGHT