NAME

AnalyzeSequenceFilesData.pl - Analyze sequence and alignment files

SYNOPSIS

AnalyzeSequenceFilesData.pl SequenceFile(s) AlignmentFile(s)...

AnalyzeSequenceFilesData.pl [-h, --help] [-i, --IgnoreGaps yes | no] [-m, --mode PercentIdentityMatrix | ResidueFrequencyAnalysis | All] [--outdelim comma | tab | semicolon] [-o, --overwrite] [-p, --precision number] [-q, --quote yes | no] [--ReferenceSequence SequenceID | UseFirstSequenceID] [--region "StartResNum, EndResNum, [StartResNum, EndResNum...]" | UseCompleteSequence] [--RegionResiduesMode AminoAcids | NucleicAcids | None] [-w, --WorkingDir dirname] SequenceFile(s) AlignmentFile(s)...

DESCRIPTION

Analyze *SequenceFile(s) and AlignmentFile(s)* data: calculate pairwise percent identity matrix or calculate percent occurrence of various residues in specified sequence regions. All the sequences in the input file must have the same sequence lengths; otherwise, the sequence file is ignored.

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

-h, --help

Print this help message.

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

Ignore gaps during calculation of sequence lengths and specification of regions during residue frequency analysis. Possible values: *yes or no*. Default value: *yes*.

-m, --mode *PercentIdentityMatrix* | *ResidueFrequencyAnalysis* | *All*

Specify how to analyze data in sequence files: calculate percent identity matrix or calculate frequency of occurrence of residues in specific regions. During *ResidueFrequencyAnalysis* value of -m, --mode option, output files are generated for both the residue count and percent residue count. Possible values: *PercentIdentityMatrix, ResidueFrequencyAnalysis, or All*. Default value: *PercentIdentityMatrix*.

--outdelim *comma* | *tab* | *semicolon*

Output text file delimiter. Possible values: *comma, tab, or semicolon*. Default value: *comma*.

-o, --overwrite

Overwrite existing files.

-p, --precision *number*

Precision of calculated values in the output file. Default: up to *2* decimal places. Valid values: positive integers.

-q, --quote *yes* | *no*

Put quotes around column values in output text file. Possible values: *yes or no*. Default value: *yes*.

--ReferenceSequence *SequenceID* | *UseFirstSequenceID*

Specify reference sequence ID to identify regions for performing *ResidueFrequencyAnalysis* specified using -m, --mode option. Default: *UseFirstSequenceID*.

--region *StartResNum,EndResNum,[StartResNum,EndResNum...]* | *UseCompleteSequence*

Specify how to perform frequency of occurrence analysis for residues: use specific regions indicated by starting and ending residue numbers in reference sequence or use the whole reference sequence as one region. Default: *UseCompleteSequence*.

Based on the value of -i, --IgnoreGaps option, specified residue numbers *StartResNum,EndResNum* correspond to the positions in the reference sequence without gaps or with gaps.

For residue numbers corresponding to the reference sequence including gaps, percent occurrence of various residues corresponding to gap position in reference sequence is also calculated.

--RegionResiduesMode *AminoAcids* | *NucleicAcids* | *None*

Specify how to process residues in the regions specified using --region option during *ResidueFrequencyAnalysis* calculation: categorize residues as amino acids, nucleic acids, or simply ignore residue category during the calculation. Possible values: *AminoAcids, NucleicAcids or None*. Default value: *None*.

For *AminoAcids* or *NucleicAcids* values of --RegionResiduesMode option, all the standard amino acids or nucleic acids are listed in the output file for each region; Any gaps and other non standard residues are added to the list as encountered.

For *None* value of --RegionResiduesMode option, no assumption is made about type of residues. Residue and gaps are added to the list as encountered.

-r, --root *rootname*

New sequence file name is generated using the root: <Root><Mode>.<Ext> and <Root><Mode><RegionNum>.<Ext>. Default new file name: <SequenceFileName><Mode>.<Ext> for *PercentIdentityMatrix* value m, --mode option and <SequenceFileName><Mode><RegionNum>.<Ext> for *ResidueFrequencyAnalysis*. The csv, and tsv <Ext> values are used for comma/semicolon, and tab delimited text files respectively. This option is ignored for multiple input files.

-w --WorkingDir *text*

Location of working directory. Default: current directory.

EXAMPLES

To calculate percent identity matrix for all sequences in Sample1.msf file and generate Sample1PercentIdentityMatrix.csv, type:

```
% AnalyzeSequenceFilesData.pl Sample1.msf
```

To perform residue frequency analysis for all sequences in Sample1.aln file corresponding to non-gap positions in the first sequence and generate Sample1ResidueFrequencyAnalysisRegion1.csv and Sample1PercentResidueFrequencyAnalysisRegion1.csv files, type:

```
% AnalyzeSequenceFilesData.pl -m ResidueFrequencyAnalysis -o
  Sample1.aln
```

To perform residue frequency analysis for all sequences in Sample1.aln file corresponding to all positions in the first sequence and generate TestResidueFrequencyAnalysisRegion1.csv and TestPercentResidueFrequencyAnalysisRegion1.csv files, type:

```
% AnalyzeSequenceFilesData.pl -m ResidueFrequencyAnalysis --IgnoreGaps
  No -o -r Test Sample1.aln
```

To perform residue frequency analysis for all sequences in Sample1.aln file corresponding to non-gap residue positions 5 to 10, and 30 to 40 in sequence ACHE_BOVIN and generate Sample1ResidueFrequencyAnalysisRegion1.csv, Sample1ResidueFrequencyAnalysisRegion2.csv, SamplePercentResidueFrequencyAnalysisRegion1.csv, and SamplePercentResidueFrequencyAnalysisRegion2.csv files, type:

```
% AnalyzeSequenceFilesData.pl -m ResidueFrequencyAnalysis
  --ReferenceSequence ACHE_BOVIN --region "5,15,30,40" -o Sample1.msf
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

ExtractFromSequenceFiles.pl, InfoSequenceFiles.pl

COPYRIGHT