NAME

ExtractFromSequenceFiles.pl - Extract data from sequence and alignment files

SYNOPSIS

ExtractFromSequenceFiles.pl SequenceFile(s) AlignmentFile(s)...

ExtractFromSequenceFiles.pl [-h, --help] [-i, --IgnoreGaps yes | no] [-m, --mode SequenceID | SequenceNum | SequenceNumRange] [-o, --overwrite] [-r, --root rootname] [-s, --Sequences "SequenceID, [SequenceID,...]" | "SequenceNum, [SequenceNum,...]" | "StartingSeqNum, EndingSeqNum"] [--SequenceIDMatch Exact | Relaxed] [-w, --WorkingDir dirname] SequenceFile(s) AlignmentFile(s)...

DESCRIPTION

Extract specific data from *SequenceFile(s) and AlignmentFile(s)* and generate FASTA files. You can extract sequences using sequence IDs or sequence numbers.

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 or gap columns during during generation of new sequence or alignment file(s). Possible values: *yes or no*. Default value: *yes*.

In order to remove gap columns, length of all the sequence must be same; otherwise, this option is ignored.

-m, --mode *SequenceID* | *SequenceNum* | *SequenceNumRange*

Specify how to extract data from sequence files: extract sequences using sequence IDs or sequence numbers. Possible values: *SequenceID* | *SequenceNum* | *SequenceNumRange*. Default: *SequenceNum* with value of 1.

The sequence numbers correspond to position of sequences starting from 1 for first sequence in *SequenceFile(s) and AlignmentFile(s)*.

-o, --overwrite

Overwrite existing files.

-r, --root *rootname*

New sequence file name is generated using the root: <Root><Mode>.<Ext>. Default new file: <SequenceFileName><Mode>.<Ext>. This option is ignored for multiple input files.

-s, --Sequences *"SequenceID,[SequenceID,...]"* | *"SequenceNum,[SequenceNum,...]"* | *"StartingSeqNum,EndingSeqNum"*

This value is -m, --mode specific. In general, it's a comma delimites list of sequence IDs or sequence numbers.

For *SequenceID* value of -m, --mode option, input value format is: *SequenceID,...*. Examples:

```
ACHE_BOVIN
ACHE_BOVIN,ACHE_HUMAN
```

For *SequenceNum* value of -m, --mode option, input value format is: *SequenceNum,...*. Examples:

```
2
1,5
```

For *SequenceNum* value of -m, --mode option, input value format is: *StaringSeqNum,EndingSeqNum*. Examples:

```
2,4
```

--SequenceIDMatch *Exact* | *Relaxed*

Sequence IDs matching criterion during *SequenceID* value of -m, --mode option: match specified

sequence ID exactly or as sub string against sequence IDs in the files. Possible values: *Exact* | *Relaxed*. Default: *Relaxed*. Sequence ID match is case insenstitive during both options.

--SequenceLength *number*

Maximum sequence length per line in sequence file(s). Default: *80*.

-w --WorkingDir *text*

Location of working directory. Default: current directory.

EXAMPLES

To extract first sequence from Sample1.fasta sequence file and generate Sample1SequenceNum.fasta sequence file, type:

```
% ExtractFromSequenceFiles.pl -o Sample1.fasta
```

To extract first sequence from Sample1.aln alignment file and generate Sample1SequenceNum.fasta sequence file without any column gaps, type:

```
% ExtractFromSequenceFiles.pl -o Sample1.aln
```

To extract first sequence from Sample1.aln alignment file and generate Sample1SequenceNum.fasta sequence file with column gaps, type:

```
% ExtractFromSequenceFiles.pl --IgnroreGaps No -o Sample1.aln
```

To extract sequence number 1 and 4 from Sample1.fasta sequence file and generate Sample1SequenceNum.fasta sequence file, type:

```
% ExtractFromSequenceFiles.pl -o -m SequenceNum --Sequences 1,4
  -o Sample1.fasta
```

To extract sequences from sequence number 1 to 4 from Sample1.fasta sequence file and generate Sample1SequenceNumRange.fasta sequence file, type:

```
% ExtractFromSequenceFiles.pl -o -m SequenceNumRange --Sequences
  1,4 -o Sample1.fasta
```

To extract sequence ID "Q9P993/104-387" from sequence from Sample1.fasta sequence file and generate Sample1SequenceID.fasta sequence file, type:

```
% ExtractFromSequenceFiles.pl -o -m SequenceID --Sequences
  "Q9P993/104-387" --SequenceIDMatch Exact -o Sample1.fasta
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

AnalyzeSequenceFilesData.pl, InfoSequenceFiles.pl

COPYRIGHT

Copyright (C) 2022 Manish Sud. All rights reserved.

This file is part of MayaChemTools.

MayaChemTools is free software; you can redistribute it and/or modify it under the terms of the GNU Lesser General Public License as published by the Free Software Foundation; either version 3 of the License, or (at your option) any later version.