seqmagick Documentation

Release 0.4.0

Matsen Group

March 24, 2014

Contents

CHAPTER 1

Motivation

We often have to convert between sequence formats and do little tasks on them, and it's not worth writing scripts for that. Seqmagick is a kickass little utility built in the spirit of imagemagick to expose the file format conversion in Biopython in a convenient way. Instead of having a big mess of scripts, there is one that takes arguments:

```
seqmagick convert a.fasta b.phy    # convert from fasta to phylip
seqmagick mogrify --ungap a.fasta  # remove all gaps from a.fasta, in place
seqmagick info *.fasta             # describe all FASTA files in the current directory
```

And more.

CHAPTER 2

Installation

First, you'll need to install BioPython. NumPy (which parts of BioPython depend on) is not required for `seqmagick` to function. Once done, install with:

```
pip install seqmagick
```

Get the bleeding edge version here, or clone our repository:

```
git clone git://github.com/fhcrc/seqmagick.git
```

CHAPTER 3

Use

Seqmagick can be used to query information about sequence files, convert between types, and modify sequence files. All functions are accessed through subcommands:

```
seqmagick <subcommand> [options] arguments
```

CHAPTER 4

Subcommands

4.1 `convert` and `mogrify`

Convert and mogrify achieve similar goals. `convert` performs some operation on a file (from changing format to something more complicated) and writes to a new file. `mogrify` modifies a file in place, and would not normally be used to convert formats.

The two have similar signatures:

```
seqmagick convert [options] infile outfile
```

vs:

```
seqmagick mogrify [options] infile
```

Options are shared between convert and mogrify.

4.1.1 Examples

Basic Conversion

`convert` can be used to convert between any file types BioPython supports (which is many). For a full list of supported types, see the BioPython SeqIO wiki page.

By default, file type is inferred from file extension, so:

```
seqmagick convert a.fasta a.sto
```

converts an existing file `a.fasta` from FASTA to Stockholm format. **Neat!** But there's more.

Sequence Modification

A wealth of options await you when you're ready to do something slightly more complicated with your sequences.

Let's say I just want a few of my sequences:

```
$ seqmagick convert --head 5 examples/test.fasta examples/test.head.fasta
$ seqmagick info examples/test*.fasta
name                      alignment  min_len  max_len  avg_len  num_seqs
examples/test.fasta       FALSE      972      9719     1573.67  15
examples/test.head.fasta  FALSE      978      990      984.00   5
```

Or I want to remove any gaps, reverse complement, select the last 5 sequences, and remove any duplicates from an alignment in place:

```
seqmagick mogrify --tail 5 --reverse-complement --ungap --deduplicate-sequences examples/test.fasta e
```

You can even define your own functions in python and use them via `--apply-function`.

Note: To maximize flexibility, most transformations passed as options to `mogrify` and `convert` are processed *in order*, so:

```
seqmagick convert --min-length 50 --cut 1:5 a.fasta b.fasta
```

will work fine, but:

```
seqmagick convert --cut 1:5 --min-length 50 a.fasta b.fasta
```

will never return records, since the cutting transformation happens before the minimum length predicate is applied.

4.1.2 Command-line Arguments

The full set of options to `mogrify` and `convert` are:

Sequence File Modification

```
--line-wrap N         Adjust line wrap for sequence strings. When N is 0,
                      all line breaks are removed. Only fasta files are
                      supported for the output format.
--sort {length-asc,length-desc,name-asc,name-desc}
                      Perform sorting by length or name, ascending or
                      descending. ASCII sorting is performed for names
```

Sequence Modification

```
--apply-function /path/to/module.py:function_name
                      Specify a custom function to apply to the input
                      sequences, specified as
                      /path/to/file.py:function_name. Function should accept
                      an iterable of Bio.SeqRecord objects, and yield
                      SeqRecords. Specify more than one to chain.
--cut start:end[,start2:end2]
                      1-indexed start and end positions for cutting sequences, : separated. Includes
                      to indicate start/end of sequence.
--relative-to ID      Apply --cut relative to the indexes of non-gap residues in sequence identified
--dash-gap            Change . and : into - for all sequences
--mask start:end[,start2:end2...]
                  Replace residues in 1-indexed slice with gap-
                  characters. If --relative-to is also specified,
                  coordinates are relative to the sequence ID provided.
--lower               Translate the sequences to lower case
--reverse             Reverse the order of sites in sequences
--reverse-complement  Convert sequences into reverse complements
--squeeze             Remove any gaps that are present in the same position
                      across all sequences in an alignment (equivalent to
                      --squeeze-threshold=1.0)
```

```
--squeeze-threshold PROP
                      Trim columns from an alignment which have gaps in
                      least the specified proportion of sequences.
--transcribe {dna2rna,rna2dna}
                      Transcription and back transcription for generic DNA
                      and RNA. Source sequences must be the correct alphabet
                      or this action will likely produce incorrect results.
--translate {dna2protein,rna2protein,dna2proteinstop,rna2proteinstop}
                      Translate from generic DNA/RNA to proteins. Options
                      with "stop" suffix will NOT translate through stop
                      codons .Source sequences must be the correct alphabet
                      or this action will likely produce incorrect results.
--ungap               Remove gaps in the sequence alignment
--upper               Translate the sequences to upper case
```

Record Selection

```
--deduplicate-sequences
                      Remove any duplicate sequences by sequence content,
                      keep the first instance seen
--deduplicated-sequences-file FILE
                      Write all of the deduplicated sequences to a file
--deduplicate-taxa    Remove any duplicate sequences by ID, keep the first
                      instance seen
--exclude-from-file FILE
                      Filter sequences, removing those sequence IDs in the
                      specified file
--include-from-file FILE
                      Filter sequences, keeping only those sequence IDs in
                      the specified file
--head N              Trim down to top N sequences
--max-length N        Discard any sequences beyond the specified maximum
                      length. This operation occurs *before* all length-
                      changing options such as cut and squeeze.
--min-length N        Discard any sequences less than the specified minimum
                      length. This operation occurs *before* all length-
                      changing options such as cut and squeeze.
--min-ungapped-length N
                      Discard any sequences less than the specified minimum
                      length, excluding gaps. This operation occurs *before*
                      all length-changing options such as cut and squeeze.
--pattern-include regex
                      Filter the sequences by regular expression in name
--pattern-exclude regex
                      Filter out sequences by regular expression in name
--prune-empty         Prune sequences containing only gaps ('-')
--seq-pattern-include regex
                      Filter the sequences by regular expression in sequence
--seq-pattern-exclude regex
                      Filter out sequences by regular expression in sequence
--tail N              Trim down to bottom N sequences
```

Sequence ID Modification

```
--first-name          Take only the first whitespace-delimited word as the
                      name of the sequence
--name-suffix SUFFIX  Append a suffix to all IDs.
--name-prefix PREFIX  Insert a prefix for all IDs.
--pattern-replace search_pattern replace_pattern
                      Replace regex pattern "search_pattern" with
                      "replace_pattern" in sequence ID
--strip-range         Strip ranges from sequences IDs, matching </x-y>
```

Format Options

By default, file format is inferred from extension:

```
--input-format Format
                      Input file format (default: determine from extension)
--output-format Format
                      Output file format (default: determine from extension)
```

4.2 backtrans-align

Given a protein alignment and unaligned nucleotides, align the nucleotides using the protein alignment. Protein and nucleotide sequence files must contain the same number of sequences, in the same order, with the same IDs.

```
usage: seqmagick backtrans-align [-h] [-o destination_file]
                                 [-t {standard-ambiguous,vertebrate-mito,standard}]
                                 protein_align nucl_align

positional arguments:
  protein_align         Protein Alignment
  nucl_align            FASTA Alignment

optional arguments:
  -h, --help            show this help message and exit
  -o destination_file, --out-file destination_file
                        Output destination. Default: STDOUT
  -t {standard-ambiguous,vertebrate-mito,standard}, --translation-table {standard-ambiguous,vertebrat
                        Translation table to use. [Default: standard]
```

4.3 extract-ids

`seqmagick extract-ids` is extremely simple - all the IDs from a sequence file are printed to stdout (by default) or the file of your choosing:

```
positional arguments:
  sequence_file         Sequence file

optional arguments:
  -h, --help            show this help message and exit
  -o OUTPUT_FILE, --output-file OUTPUT_FILE
                        Destination trimmed file
```

```
  --source-format SOURCE_FORMAT
  -d, --include-description
                        Include the sequence description in output [default: False]
```

4.4 info

`seqmagick info` describes one or more sequence files

4.4.1 Example

```
seqmagick info examples/*.fasta
```

```
name                        alignment  min_len  max_len  avg_len  num_seqs
examples/aligned.fasta      TRUE       9797     9797     9797.00  15
examples/dewrapped.fasta    TRUE       240      240      240.00   148
examples/range.fasta        TRUE       119      119      119.00   2
examples/test.fasta         FALSE      972      9719     1573.67  15
examples/wrapped.fasta      FALSE      120      237      178.50   2
```

Output can be in comma-separated, tab-separated, or aligned formats. See `seqmagick info -h` for details.

4.5 backtrans-align

Given a protein alignment and unaligned nucleotides, align the nucleotides using the protein alignment. **Protein and nucleotide sequence files must contain the same number of sequences, in the same order, with the same IDs.**

```
usage: seqmagick backtrans-align [-h] [-o destination_file]
                                 [-t {standard-ambiguous,vertebrate-mito,standard}]
                                 protein_align nucl_align


positional arguments:
  protein_align         Protein Alignment
  nucl_align            FASTA Alignment

optional arguments:
  -h, --help            show this help message and exit
  -o destination_file, --out-file destination_file
                        Output destination. Default: STDOUT
  -t {standard-ambiguous,vertebrate-mito,standard}, --translation-table {standard-ambiguous,vertebrat
                        Translation table to use. [Default: standard]
```

4.6 primer-trim

`primer-trim` trims an alignment to a region defined by a set of forward and reverse primers. Usage is as follows:

```
positional arguments:
  source_file           Source alignment file
  output_file           Destination trimmed file
  forward_primer        The forward primer used
```

```
  reverse_primer        The reverse primer used. By default the reverse primer
                        is assumed to be a subsequence of the top strand (that
                        is, the reverse complement of an actual downstream PCR
                        primer). Use --reverse-is-revcomp if this is not the
                        case.

optional arguments:
  -h, --help            show this help message and exit
  --reverse-is-revcomp  Reverse primer is written as the reverse complement of
                        the top strand (default: False)
  --source-format SOURCE_FORMAT
                        Alignment format (default: detect from extension
  --output-format OUTPUT_FORMAT
                        Alignment format (default: detect from extension
  --include-primers     Include the primers in the output (default: False)
  --max-hamming-distance MAX_HAMMING_DISTANCE
                        Maximum Hamming distance between primer and alignment
                        site (default: 1). IUPAC ambiguous bases in the primer
                        matching unambiguous bases in the alignment are not
                        penalized
  --prune-action {trim,isolate}
                        Action to take. Options are trim (trim to the region
                        defined by the two primers, decreasing the width of
                        the alignment), or isolate (convert all characters
                        outside the primer-defined area to gaps). default:
                        trim
```

4.7 quality-filter

`quality-filter` truncates and removes sequences that don't match a set of quality criteria. The subcommand takes a FASTA and quality score file, and writes the results to an output file:

```
positional arguments:
  input_fasta           Input fasta file
  input_qual            The quality scores associated with fasta_file
  output_file           Output file. Format determined from extension.

optional arguments:
  -h, --help            show this help message and exit
  --min-mean-quality QUALITY
                        Minimum mean quality score for each read [default: 25]
  --min-length LENGTH   Minimum length to keep sequence [default: None]
  --quality-window WINDOW_SIZE
                        Window size for truncating sequences. When set to a
                        non-zero value, sequences are truncated where the mean
                        mean quality within the window drops below --min-mean-
                        quality. [default: 0]
  --ambiguous-action {truncate,drop}
                        Action to take on ambiguous base in sequence (N's).
                        [default: no action]
```

CHAPTER 5

Supported File Extensions

By default, `seqmagick` infers the file type from extension. Currently mapped extensions are:

Extension	Format
.afa	fasta
.aln	clustal
.fa	fasta
.faa	fasta
.fas	fasta
.fasta	fasta
.fastq	fastq
.ffn	fasta
.fna	fasta
.frn	fasta
.gb	genbank
.gbk	genbank
.needle	emboss
.phy	phylip
.phylip	phylip
.phyx	phylip-relaxed (**note:** requires building BioPython from the master branch until v1.58 is released)
.qual	qual
.sff	sff
.sth	stockholm
.sto	stockholm

If an extension is not listed, you can either rename the file to a supported extension, or specify it manually via `--input-format` or `--output-format`.

CHAPTER 6

Acknowledgements

seqmagick is written and maintained by the Matsen Group at the Fred Hutchinson Cancer Research Center.

CHAPTER 7

Contributing

We welcome contributions! Simply fork the repository on GitHub and send a pull request.