Translate fasta headers

Translate long fasta headers to short - and back!

Your alignment program X doesn't allow strings longer than n characters, but all your info is in the fasta headers of your file. What to do?

Use `translate_fasta_headers.pl` on your fasta file to create short labels and a translation table. Run your program X, and then back-translate your fasta headers by running `translate_fasta_headers.pl` again!

And if you created a tree with the short (or long) labels, try to back-translate using `replace_taxon_labels_in_newick.pl`!

If you only wish to transform your long fasta headers to short, without keeping the information about how they where translated, the quick solution might be to use `awk`:

```
$ awk '/>/{$0=">Seq_"++n}1' long.fas
```

But, if you want to be able to back-translate, read on!

Description

Replace fasta headers with headers taken from tab delimited file. If no tab file is given, the (potentially long) fasta headers are replaced by short labels "Seq_1", "Seq_2", etc, and the short and original headers are printed to a translation file.

If you wish, you may choose your own prefix (instead of `Seq_`). This could be handy if, for example, you wish to concatenate files.

The script for translating labels in Newick trees is somewhat limited in capacity due to the restrictions and/or peculiarities of the Newick tree format. Use with caution.

Usage

```
$ translate_fasta_headers.pl [options] <fasta file>
$ replace_taxon_labels_in_newick.pl [options] <newick file>
```

Examples

From long to short labels:

```
$ translate_fasta_headers.pl --out=short.fas long.fas
```

And back, using a translation table:

```
$ translate_fasta_headers.pl --tabfile=short.fas.translation.tab short.fas
```

Slightly shorter version (see note about the `--out` option below):

```
$ translate_fasta_headers.pl long.fas > short.fas
$ translate_fasta_headers.pl -t long.fas.translation.tab short.fas
```

Use your own prefix:

```
$ translate_fasta_headers.pl --prefix='Own_' long.fas
```

Translate short seq labels in Newick tree to long:

```
$ replace_taxon_labels_in_newick.pl -t long.fas.translation.tab short.fas.phy
```

Print seq labels in Newick tree:

```
$ replace_taxon_labels_in_newick.pl -l short.fas.phy
```

Options

Script `translate_fasta_headers.pl`

- `-t, --tabfile=<filename>` – Specify tab-separated translation file with unique "short" labels to the left, and "long" names to the right. Translation will be from left to right.
- `-o, --out=<filename>` – Specify output file for the fasta sequences. **Note**: If `--out=<filename>` is specified, the translation file will be named `<filename>.translation.tab`. This simplifies back translation. If, on the other hand, `--out` is not used, the translation file will be named after the infile!
- `-i, --in=<filename>` – Specify name of fasta file. Can be skipped as script reads files from STDIN.
- `-n, --notab` – Do not create a translation file.
- `-p, --prefix=<string>` – User your own prefix (default is `Seq_`). A numerical will be added to the labels (e.g. `Own_1`, `Own_2`, ...)
- `-v, --version` – Print version number and quit.
- `-h, --help` – Show this help text and quit.

Script `replace_taxon_labels_in_newick.pl`

- `-t, --tabfile=<translation.tab>` – File with table describing what will be translated with what.
- `-l,-p, --labels` – Print taxon labels in tree. Option does not require a translation table.
- `--no-quotemeta` – Turn off escaping of special symbols in the replacements.
- `-o, --out=<out.file>` – Print to outfile `out.file`, else to STDOUT.
- `-v, --version` – Print version number and quit.
- `-h, --help` – Help text.

Author

Johan.Nylander

Files

- `translate_fasta_headers.pl` – Perl script
- `replace_taxon_labels_in_newick.pl` – Perl script
- `data/long.fas` – Example file with long fasta headers
- `data/short.fas.translation.tab` – Example translation table
- `data/short.fas` – Example output with short fasta headers
- `data/short.fas.phy` – Example Newick tree with short labels
- `README.md` – Documentation, markdown format
- `README.pdf` – Documentation, PDF format

License and Copyright

Copyright (c) 2013-2024 Johan Nylander

LICENSE