Fastcov - Fast Multiple Covariance Detector v1.03

Usage

```
Name:
  fastcov V1.03 -- Fast Multiple Covariance Detector
  http://yanlilab.github.io/fastcov

Authors:
  Yan Li   <liyan.com@gmail.com>
  Wei Shen <shenwei356@gmail.com>

Usage:
  fastcov [options] inputfile

Available Options:
  -p FLOAT                  minimum pairing purity of two sites [0.7]
  -r FLOAT                  minimum matching ratio of to the pattern [0.45]
  -n INT                    minimum residue number at each site [5]
  -c FLOAT                  minimum proportion of any sequence identical to the
                            consensus [0.33]
  -o STRING                 prefix of output files [inputfile]
  -j INT                    CPU number [CPU number of your computer]
  -h, --help                show this help message

Copyright:
  Copyright © 2015-2016, All Rights Reserved
  This software is free to distribute for academic research.
```

Positional arguments

- `inputfile` should be aligned protein sequences in FASTA format file, produced by multi sequence alignment softwares.
  Case is not sensitive.

  One-seq-per-line format could be converted to FASTA format by

  ```
  for f in *.aln; do cat -n $f | awk '{print ">"$1"\n"$2}' > $f.fas; done
  ```

Options

Main algorithm parameters

- `-p` defines the minimum pairing purity of two sites. Default is 0.7.
- `-r` defines the minimum matching ratio of to the pattern at clustering stage. Default is 0.45.

Sequences filter criteria

- `-n` is the minimum residue number at each site. Default value is 5.
- `-c` is the minimum proportion of any sequence identical to the consensus. Default value is 0.33, i.e. the number of residues identical to the that of the same position of consensus sequences should be at least one third of the length of consensus.
  Sequences that fail to reach this criteria will be discarded.

Output

- `-o` defines the prefix of output files, default value is the same as input file. e.g, for a input file `test.fa`, output files will be:

```
test.aligned.fa.pairs.txt
test.aligned.fa.clusters.txt
test.aligned.fa.patterns.txt
test.aligned.fa.seq2patterns.txt
```

Performance

- `-j` is the number of CPU. fastcov detects your computer and set the default value with the maximum CPU number. The bigger the value is, the faster fastcov runs.

Examples

Taking `examples/ABCD_RT_M.aligned.fas` for example.

Quik run:

```
fastcov ABCD_RT_M.aligned.fas
```

Terminal Output:

```
Input: ABCD_RT_M.aligned.fas

Step 1/5: Reading sequences


Done

Step 2/5: Searching candidate sites

Done

Step 3/5: Searching independent pairs
21115 / 21115
[==========================================================================
=====] 100.00 % 28s

Covariant site pairs saved to file: ABCD_RT_M.aligned.fas.pairs

Done

Step 4/5: Searching covariant patterns
52 / 52
[==========================================================================
=============] 100.00 % 0

Covariant patterns saved to file: ABCD_RT_M.aligned.fas.patterns

Done

Step 5/5: Clustering by covariant patterns
Covariant patterns assigned to sequences: ABCD_RT_M.aligned.fas.seq2patterns
Sequences clustered by covariant patterns: ABCD_RT_M.aligned.fas.clusters
```

The most time-consuming stage is step 3, so we add a process bar.

Output files:

```
ABCD_RT_M.aligned.fas.pairs.txt            # covariant pairs information, table
file, could be imported to MS Excel
ABCD_RT_M.aligned.fas.patterns.txt         # covariant patterns, table file,
could be imported to MS Excel
ABCD_RT_M.aligned.fas.clusters.txt         # sequence clusters by covariant
patterns
ABCD_RT_M.aligned.fas.seq2patterns.txt     # covariant patterns of every
sequence, table file, could be imported to MS Excel
```

Note: For windows user, please use a modern text editor to view the result files. Notepad is not recommended, [Notepad++ (https://notepad-plus-plus.org/)](https://notepad-plus-plus.org/) is a better choice.

[More examples (https://github.com/yanlilab/fastcov/tree/master/examples)](https://github.com/yanlilab/fastcov/tree/master/examples)

Errors and Solutions

1. No input file given. Please feed fastcov a aligned amino acids sequences in FASTA format.

```
$ fastcov
[Error] no input file (aligned amino acids sequences in FASTA format)
given.
type "fastcov -h" for help
```

2. Input file is not aligned.

```
[Error] sequence length not equal: 343 (AB014392_Pol-C) != 344.
input file should be aligned amino acids sequences in FASTA format
```

3. Illegal characters in sequence. FASTA parsing module of fastcov strictly check the sequences, you may check input sequence according according to the IUPAC nucleotide code (http://www.bioinformatics.org/sms2/iupac.html). It may also be caused by unmatch of sequence type (PROTEIN) and actual sequence type (DNA) in FASTA file.

```
Input: test.fa

Step 1/5: Reading sequences
error when reading AB014367_Pol-C: invalid Protein sequence:
AB014367_Pol-C
```

FAQ

Please don't hesitate to email us.

Q: What a mess when opening the result files!

A: Microsoft Windows user may open the result files by Notepad provided by the Operating system.
Please choose another moder text editor like [Notepad++ (https://notepad-plus-plus.org/)](https://notepad-plus-plus.org/).

Authors

Yan Li [liyan.com@gmail.com (mailto:liyan.com@gmail.com)](mailto:liyan.com@gmail.com), Wei Shen [shenwei356@gmail.com (mailto:shenwei356@gmail.com)](mailto:shenwei356@gmail.com)

Copyright