Jellyfish: A fast k-mer counter

G. Marcais

April 23, 2011
Version 1.1

Abstract

Jellyfish is a software to count k-mers in DNA sequences.

1 Synopsis

jellyfish count [**-o***prefix*] [**-m***merlength*] [**-t***threads*] [**-s***hashsize*] [**--both-strands**] *fasta* [*fasta* ...]
jellyfish merge hash1 hash2 ...
jellyfish dump hash
jellyfish stats hash
jellyfish histo [**-h***high*] [**-l***low*] [**-i***increment*] *hash*
jellyfish query hash

2 Description

Jellyfish is a k-mer counter based on a multi-threaded hash table implementation.

To count k-mers, use a command like:

```
jellyfish count -m 22 -o output -c 3 -s 10000000 -t 32 input.fasta
```

This will count the the 22-mers in species.fasta with 32 threads. The counter field in the hash uses only 3 bits and the hash has at least 10 million entries. Let the size of the table be $s = 2^l$ and the max reprobe value is less than 2^r, then the memory usage per entry in the hash is (in bits, not bytes) $2k - l + r + 1$.

To save space, the hash table supports variable length counter, i.e. a k-mer occurring only a few times will use a small counter, a k-mer occurring many times will used multiple entries in the hash. The **-c** specify the length of the small counter. The tradeoff is: a low value will save space per entry in the hash but will increase the number of entries used, hence maybe requiring a larger hash. In practice, use a value for **-c** so that most of you k-mers require only 1 entry. For example, to count k-mers in a genome, where most of the sequence

is unique, use **-c***1* or **-c***2*. For sequencing reads, use a value for **-c** large enough to counts up to twice the coverage.

When the orientation of the sequences in the input fasta file is not known, e.g. in sequencing reads, using **--both-strands** (**-C**) makes the most sense.

The following subcommand are used to look at the result: histo, dump, stats.

3 Options

3.1 count

Count k-mers or qmers in fasta or fastq files
Usage: jellyfish count [OPTIONS]... [file.f[aq]]...

-h,--help Print help and exit

--full-help Print help, including hidden options, and exit

-V,--version Print version and exit

-m,--mer-len=*INT* Length of mer (mandatory)

-s,--size=*LONG* Hash size (mandatory)

-t,--threads=*INT* Number of threads (default=1)

-o,--output=*STRING* Output prefix (default=mer_counts)

-c,--counter-len=*Length* in bits Length of counting field (default=7)

--out-counter-len=*Length* in bytes Length of counter field in output (default=4)

-C,--both-strands Count both strand, canonical representation (default=off)

-p,--reprobes=*INT* Maximum number of reprobes (default=62)

-r,--raw Write raw database (default=off)

-q,--quake Quake compatibility mode (default=off)

--quality-start=*INT* Starting ASCII for quality values (default=64)

--min-quality=*INT* Minimum quality. A base with lesser quality becomes an N (default=0)

-L,--lower-count=*LONG* Don't output k-mer with count ¡ lower-count

-U,--upper-count=*LONG* Don't output k-mer with count ¿ upper-count

--matrix=*Matrix* file Hash function binary matrix

--timing=*Timing* file Print timing information

3.2 histo

Create an histogram of k-mer occurrences
Usage: jellyfish histo [OPTIONS]... [database.jf]...

--help Print help and exit

-V,--version Print version and exit

-s,--buffer-size=*Buffer* length Length in bytes of input buffer (default=10000000)

-l,--low=*LONG* Low count value of histogram (default=1)

-h,--high=*LONG* High count value of histogram (default=10000)

-i,--increment=*LONG* Increment value for buckets (default=1)

-t,--threads=*INT* Number of threads (default=1)

-o,--output=*STRING* Output file (default=/dev/fd/1)

3.3 dump

Dump k-mer counts
Usage: jellyfish stats [OPTIONS]... [database.jf]...

-h,--help Print help and exit

-V,--version Print version and exit

-c,--column Column format (default=off)

-t,--tab Tab separator (default=off)

-L,--lower-count=*LONG* Don't output k-mer with count ¡ lower-count

-U,--upper-count=*LONG* Don't output k-mer with count ¿ upper-count

-o,--output=*STRING* Output file (default=/dev/fd/1)

3.4 stats

Statistics
Usage: jellyfish stats [OPTIONS]... [database.jf]...

-h,--help Print help and exit

--full-help Print help, including hidden options, and exit

-V,--version Print version and exit

-L,--lower-count=*LONG* Don't output k-mer with count ¡ lower-count

-U,--upper-count=*LONG* Don't output k-mer with count ¿ upper-count

-v,--verbose Verbose (default=off)

-o,--output=*STRING* Output file (default=/dev/fd/1)

3.5 merge

Merge jellyfish databases
Usage: jellyfish merge [OPTIONS]... [database.jf]...

-h,--help Print help and exit

-V,--version Print version and exit

-s,--buffer-size=*Buffer* length Length in bytes of input buffer (default=10000000)

-o,--output=*STRING* Output file (default=mer_counts_merged.jf)

--out-counter-len=*INT* Length (in bytes) of counting field in output (default=4)

--out-buffer-size=*LONG* Size of output buffer per thread (default=10000000)

-v,--verbose Be verbose (default=off)

3.6 cite

How to cite Jellyfish's paper
Usage: jellyfish cite [OPTIONS]...

-h,--help Print help and exit

-V,--version Print version and exit

-b,--bibtex Bibtex format (default=off)

-o,--output=*STRING* Output file (default=/dev/fd/1)

4 Version

Version: 1.1 of April 23, 2011

5 Bugs

- *jellyfish merge* has not been parallelized and is very slow.

6 Copyright & License

7 Authors

Guillaume Marcais
University of Maryland
`gmarcais@umd.edu`
Carl Kingsford
University of Maryland
`carlk@umiacs.umd.edu`