MSPCounter: A disk-based fast and memory-efficient k-mer counter

January 17, 2012

Version 0.1

Abstract

MSPCounter is a disk-based software to count k-mers in DNA sequences.

1. Synopsis

```
java -jar Partition.jar -in InputPath -k kmerLength -L readLength [-NB NumberOfBlocks] [-p MinimumSubstringLength] [-t threads] [-b bufferSize]

java -jar Count32.jar -k kmerLength -NB NumberOfBlocks [-t threads] [-b bufferSize] [-c capacity]

java -jar Count64.jar -k kmerLength -NB NumberOfBlocks [-t threads] [-b bufferSize] [-c capacity]

java -jar Dump64.jar -k kmerLength -NB NumberOfBlocks [-t threads] [-b bufferSize]

java -jar Stat64.jar -k kmerLength -NB NumberOfBlocks [-m maxCount] [-b bufferSize]

java -jar Query64.jar -q queryKmer -NB NumberOfBlocks -p MinimumSubstringLength [-b bufferSize]
```

2. Description

MSPCounter is a k-mer counter based on the minimum substring partitioning technique. It will first partition the k-mers in DNA sequences into several disjoint partitions and compress consecutive k-mers to reduce I/O cost. Then it will count k-mers in each partition individually. Since each k-mer will appear in one and only one partition, there is no need to merge the counting results in different partitions later, which helps make MSPCounter a fast and memory-efficient k-mer counter.

To count k-mers in DNA sequences with MSPCounter, use the commands like:

```
java -jar Partition.jar -in input.fasta -k 31 -L 101 -NB 256 -p 6 -t 8
java -jar Count32.jar -k 31 -NB 256 -t 8
```

These two commands will count the 31-mes in input.fasta using 8 threads. Specifically speaking, the first command will partition the short reads data input.fasta (whose average read length is 101) into 256 partitions using minimum substring partitioning, with the minimum substring length being 6; and the second command will count the 31-mers in these 256 partitions with 8 threads.

3. Options

3.1 Partition

Function: Partition short reads data (in fasta format) using minimum substring partitioning

Usage: `java -jar Partition.jar [options]`

Options Available:

```
[-help]: Print Help Information and Exit
[-in InputPath]: (String) Input Short Reads Data Path (Mandatory)
[-k k]: (Integer) K-mer Length, should be odd number smaller than 64 (Mandatory)
[-L readLength]: (Integer) Average Short Read Length (Mandatory)
[-NB numOfBlocks] :  (Integer) Number Of K-mer Blocks/Partitions. Default: 256
[-p pivotLength] :  (Integer) Minimum Substring Length. Default: 6
[-t numOfThreads] :  (Integer) Number Of Threads. Default: 8
[-b bufferSize] :  (Integer) Read/Writer Buffer Size. Default: 8192
```

3.2 Count

Function: Count the k-mers in each minimum substring partitions

Usage: For k⩽32: `java -jar Count32.jar [options]`

For k⩽64: `java -jar Count64.jar [options]`

Options Available:

```
[-help]: Print Help Information and Exit
[-k k]: (Integer) K-mer Length, should be odd number smaller than 64 (Mandatory)
[-NB numOfBlocks] :  (Integer) Number Of K-mer Blocks/Partitions. (Mandatory)
[-t numOfThreads] :  (Integer) Number Of Threads. Default: 8
[-b bufferSize] :  (Integer) Read/Writer Buffer Size. Default: 8192
[-c capacity] : (Integer) Hash Table Initial Capacity. Default: 1000000
```

Note: The settings of ***k*** and ***NB*** should be in consistent with those in Partition.

3.3 Dump

Function: Dump the k-mer counts in each minimum substring partitions

Usage: `java -jar Dump64.jar [options]`

Options Available:

```
[-help]: Print Help Information and Exit
[-k k]: (Integer) K-mer Length, should be odd number smaller than 64 (Mandatory)
[-NB numOfBlocks] :  (Integer) Number Of K-mer Blocks/Partitions. (Mandatory)
[-t numOfThreads] :  (Integer) Number Of Threads. Default: 8
[-b bufferSize] :  (Integer) Read/Writer Buffer Size. Default: 8192
```

Note: The settings of ***k*** and ***NB*** should be in consistent with those in Partition.

3.4 Stat

Function: Output the histogram of k-mer occurrences in all partitions

Usage: `java -jar Stat64.jar [options]`

Options Available:

```
[-help]: Print Help Information and Exit
[-k k]: (Integer) K-mer Length, should be odd number smaller than 64 (Mandatory)
[-NB numOfBlocks] :  (Integer) Number Of K-mer Blocks/Partitions. (Mandatory)
[-m maxCount] : (Integer) Maximum Counts to Be Considered. Default: 256
[-b bufferSize] :  (Integer) Read/Writer Buffer Size. Default: 8192
```

Note: The settings of ***k*** and ***NB*** should be in consistent with those in Partition.

3.5 Query

Function: Query the k-mer occurrence of specific k-mer

Usage: `java -jar Query64.jar [options]`

Options Available:

```
[-help]: Print Help Information and Exit
[-q queryKmer]: (String) The Specific K-mer to Be Queried (Mandatory)
[-NB numOfBlocks] :  (Integer) Number Of K-mer Blocks/Partitions. (Mandatory)
[-p pivotLength] :  (Integer) Minimum Substring Length. (Mandatory)
[-b bufferSize] :  (Integer) Read/Writer Buffer Size. Default: 8192
```

Note: The settings of ***p*** and ***NB*** should be in consistent with those in Partition.

4. Version

Version: 0.1 of January 17, 2012

5. Bug Reports

For bugs or questions or comments, please write to *yangli* at *cs* dot *ucsb* dot *edu*

6. Copyright