DSK: k-mer counting with very low memory usage
Guillaume Rizk, Dominique Lavenier, Rayan Chikhi

To cite this version:
Guillaume Rizk, Dominique Lavenier, Rayan Chikhi. DSK: k-mer counting with very low memory usage. Bioinformatics, Oxford University Press (OUP), 2013, 29 (5), pp.652-653. 10.1093/bioinformatics/btt020. hal-00778473

HAL Id: hal-00778473
https://hal.archives-ouvertes.fr/hal-00778473
Submitted on 20 Jan 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
DSK: \( k \)-mer counting with very low memory usage

Guillaume Rizk 1,* , Dominique Lavenier 2 and Rayan Chikhi 2

1 Algorizk, 75013 Paris, France
2 ENS Cachan Brittany / IRISA, Campus de Beaulieu, 35700 Rennes, France

Received on XXXXX; revised on XXXXX; accepted on XXXXX

ABSTRACT

Summary: Counting all the \( k \)-mers (substrings of length \( k \)) in DNA/RNA sequencing reads is the preliminary step of many bioinformatics applications. However, state of the art \( k \)-mer counting methods require that a large data structure resides in memory. Such structure typically grows with the number of distinct \( k \)-mers to count.

We present a new streaming algorithm for \( k \)-mer counting, called DSK (disk streaming of \( k \)-mers), which only requires a fixed, user-defined amount of memory and disk space. This approach realizes a memory, time and disk trade-off. The multi-set of all \( k \)-mers present in the reads is partitioned and partitions are saved to disk. Then, each partition is separately loaded in memory in a temporary hash table. The \( k \)-mer counts are returned by traversing each hash table. Low-abundance \( k \)-mers are optionally filtered.

DSK is the first approach that is able to count all the \( k \)-mers of a whole-genome dataset using only 4.0 GB of memory and moderate memory, time and disk trade-off. The multi-set of all \( k \)-mers of a whole-genome dataset using only 4.0 GB of memory and moderate memory, time and disk trade-off. The multi-set of all \( k \)-mers of a whole-genome dataset using only 4.0 GB of memory and moderate memory, time and disk trade-off.

In the Methods section, we describe a fixed-memory and fixed-disk space streaming algorithm, DSK, and its worst-case complexity is analyzed in function of the desired memory and disk usage. In the Results section, DSK is used to count all the \( 27 \)-mers of a whole-genome human dataset (Illumina reads). Furthermore, the trade-off between memory and disk space is analyzed on two smaller datasets.

Finally, we discuss some advantages of DSK over related methods, and show a situation where a parallel implementation significantly improved the running time.

2 METHODS

Algorithms 1 describes the DSK \( k \)-mer counting algorithm. The hash function \( h(\cdot) \) maps a \( k \)-mer to a numeric value in \([0; H]\), where \( H \) is a large integer (typically \( 2^{64} \)). In the following analysis, we make a simplifying assumption. Let \( d \) be the total number of distinct \( k \)-mers in the input, we assume that the number of distinct \( k \)-mers having a given hash value \( x \in [0; H] \) is at most \( [d/H] \). In other words, the set of distinct \( k \)-mer values can be uniformly partitioned by this hash function. Each \( k \)-mer is encoded using the classical 2 bits representation in the smallest available integer type, i.e. using \( 2^{\lceil \log_2(2k) \rceil} \) bits. The abundance of each \( k \)-mer is encoded using \( b \) bits, ensuring that the maximal disk usage of the algorithm is \( D \) bits.

The maximal memory usage of the algorithm is \( M \) bits, since Steps 7-11 require no memory, and Steps 12-17 load a single partition in \( T \) which requires exactly \( M \) bits. With an open-addressing mechanism, each distinct \( k \)-mer occupies exactly \((b + 32)\) bits in \( T \). To prove that the algorithm terminates, it suffices to show that \( T \) never overflows, i.e. that strictly

---

* Corresponding author.
Table 1. Wall-clock time and memory usage for counting 27-mers in whole-genome human data

<table>
<thead>
<tr>
<th>Program</th>
<th>Time (hours)</th>
<th>Memory (GB)</th>
<th>Disk (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSK</td>
<td>17.9</td>
<td>4</td>
<td>160</td>
</tr>
<tr>
<td>DSK-SSD *</td>
<td>3.5</td>
<td>4</td>
<td>240</td>
</tr>
<tr>
<td>BFCounter</td>
<td>41.2</td>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>Jellyfish</td>
<td>3.5</td>
<td>70</td>
<td>211</td>
</tr>
</tbody>
</table>

The dataset used is the NA18507 human genome (SRX1016231), unfiltered, consisting of 1.4 billion of reads of average length 100 bp (160 GB file size). Jellyfish used 8 threads, DSK-SSD used 4 threads, DSK and BFCounter are single-threaded. The disk column indicates the temporary amount of disk space used by each method. * Executed on a desktop computer equipped with two hard drives including a SSD.

less than $M/(b + 32)$ distinct $k$-mers are inserted in $T$. At each iteration, $(v/n_i)$ $k$-mers are split into $n_i$ partitions. Each partition contains at most $v/(n_i \cdot p) \leq [v/(0.7M/(b + 32))]$ $k$-mers. In the worst case, all these $k$-mers are distinct, thus the load factor is upper-bounded by $O(1/b)$. A classical threshold above which hash table performance degrades.

The time complexity of Steps 7-11 (including the iteration loop) is $O(v^2 b/D)$. The algorithm creates $(n_i \cdot p) \leq [v/(b + 32)/(0.7M)]$ temporary hash tables, inserting at most $[0.7M/(b + 32)]$ elements in each. Hash tables accesses and insertions (Step 15) are done in constant expected time with open-addressing, as long as the load factor is strictly less than $M/(b + 32)$ distinct $k$-mers are inserted in $T$. At each iteration, $(v/n_i)$ $k$-mers are split into $n_i$ partitions. Each partition contains at most $v/(n_i \cdot p) \leq [v/(0.7M/(b + 32))]$ $k$-mers. In the worst case, all these $k$-mers are distinct, thus the load factor is upper-bounded by $O(1/b)$. (a classical threshold above which hash table performance degrades).

The time complexity of Steps 7-11 (including the iteration loop) is $O(v^2 b/D)$. The algorithm creates $(n_i \cdot p) \leq [v/(b + 32)/(0.7M)]$ temporary hash tables, inserting at most $[0.7M/(b + 32)]$ elements in each. Hash tables accesses and insertions (Step 15) are done in constant expected time with open-addressing, as long as the load factor is strictly less than $M/(b + 32)$ distinct $k$-mers are inserted in $T$. At each iteration, $(v/n_i)$ $k$-mers are split into $n_i$ partitions. Each partition contains at most $v/(n_i \cdot p) \leq [v/(0.7M/(b + 32))]$ $k$-mers. In the worst case, all these $k$-mers are distinct, thus the load factor is upper-bounded by $O(1/b)$. (a classical threshold above which hash table performance degrades).

The time complexity of Steps 7-11 (including the iteration loop) is $O(v^2 b/D)$. The algorithm creates $(n_i \cdot p) \leq [v/(b + 32)/(0.7M)]$ temporary hash tables, inserting at most $[0.7M/(b + 32)]$ elements in each. Hash tables accesses and insertions (Step 15) are done in constant expected time with open-addressing, as long as the load factor is strictly less than $M/(b + 32)$ distinct $k$-mers are inserted in $T$. At each iteration, $(v/n_i)$ $k$-mers are split into $n_i$ partitions. Each partition contains at most $v/(n_i \cdot p) \leq [v/(0.7M/(b + 32))]$ $k$-mers. In the worst case, all these $k$-mers are distinct, thus the load factor is upper-bounded by $O(1/b)$. (a classical threshold above which hash table performance degrades).

The time complexity of Steps 7-11 (including the iteration loop) is $O(v^2 b/D)$. The algorithm creates $(n_i \cdot p) \leq [v/(b + 32)/(0.7M)]$ temporary hash tables, inserting at most $[0.7M/(b + 32)]$ elements in each. Hash tables accesses and insertions (Step 15) are done in constant expected time with open-addressing, as long as the load factor is strictly less than $M/(b + 32)$ distinct $k$-mers are inserted in $T$. At each iteration, $(v/n_i)$ $k$-mers are split into $n_i$ partitions. Each partition contains at most $v/(n_i \cdot p) \leq [v/(0.7M/(b + 32))]$ $k$-mers. In the worst case, all these $k$-mers are distinct, thus the load factor is upper-bounded by $O(1/b)$. (a classical threshold above which hash table performance degrades).

The time complexity of Steps 7-11 (including the iteration loop) is $O(v^2 b/D)$. The algorithm creates $(n_i \cdot p) \leq [v/(b + 32)/(0.7M)]$ temporary hash tables, inserting at most $[0.7M/(b + 32)]$ elements in each. Hash tables accesses and insertions (Step 15) are done in constant expected time with open-addressing, as long as the load factor is strictly less than $M/(b + 32)$ distinct $k$-mers are inserted in $T$. At each iteration, $(v/n_i)$ $k$-mers are split into $n_i$ partitions. Each partition contains at most $v/(n_i \cdot p) \leq [v/(0.7M/(b + 32))]$ $k$-mers. In the worst case, all these $k$-mers are distinct, thus the load factor is upper-bounded by $O(1/b)$. (a classical threshold above which hash table performance degrades).

3 RESULTS

In Table 1, we compared the execution time and memory usage of DSK with Jellyfish (version 1.1.5) and BFCounter (version 0.2) on a human genome Illumina dataset. The target disk usage of DSK was 160 GB of memory, the reads file was placed on a standard hard drive (DSK-SSD). The reads file was placed on a standard hard drive (DSK-SSD). Since the algorithm relies heavily on I/O to the disk, we also tested DSK with a solid-state drive (DSK-SSD). The reads file was placed on a standard hard drive, and partitions of redundant kmers were written on a 256 GB SSD. In this configuration, we noticed the algorithm is no longer limited by disk I/O and could benefit from multi-threading. The two for loops lines 7 and 12 were parallelized using openMP (4 threads).

To further assess the trade-off between time, memory and disk usage, we executed DSK (using a standard hard drive) on two smaller E. coli and D. ananassae datasets, with various target memory and disk usage parameters. For the executions with 100 MB and 1 GB memory usage, the running time of DSK on both datasets decreases as the target disk space increases. This is a consequence of the decreasing number of iterations $n_i$. The running times reaches a plateau at roughly the reads file size (where $n_i = 1$).

The execution time generally appears to be unaffected by the target memory usage. However, at the smallest tested memory usage (10 MB), the execution time on both datasets is slightly higher, possibly due to consecutive disk writes to a large number of partitions. Note that in practice, the memory usage of DSK cannot be arbitrarily low: it is limited by the number of files that can be simultaneously opened on the system (the partitions $\{d_0, ..., d_{n_p}\}$ are all opened simultaneously). In the Drosophila dataset, DSK failed to run with 10 MB of memory and 6 GB disk space for this reason.

4 DISCUSSION

Compared to other methods, DSK has three strong points:

- **Low memory usage**: Only an arbitrarily small subset of $k$-mers is loaded in memory at any time. In contrast, BFCounter stores all the $k$-mers with count $\geq 2$ in a hash table. In principle, Jellyfish can use arbitrarily small hash tables, however storing the intermediate results requires a prohibitive amount of disk ($\geq 1$ TB for human genome reads using a hash table of size 5 GB).
- **Parameters are automatically inferred**: the only mandatory argument is the $k$-mer length. Optionally, target memory and disk usages can be specified. Jellyfish and BFCounter require the user to specify respectively a hash table size and an upper-bound on the number of distinct $k$-mers.
- **Supports arbitrarily large values of $k$**: as opposed to up to 32 for Jellyfish (unbounded for BFCounter).

ACKNOWLEDGMENT

**Funding**: ANR MAPPI, ANR-10-COSI-0004

REFERENCES
