Set-min sketch: a probabilistic map for power-law distributions with applications to k-mer annotation

Yoshihiro Shibuya[1]*, Gregory Kucherov[1]

[1] *Laboratoire d'Informatique Gaspard Monge, CNRS & Université Gustave Eiffel, Marne-la-Vallée*
***Corresponding author**: yoshi.itsame@gmail.com

Abstract

Problem: Efficient storage of k-mer counting tables is a crucial part in many bioinformatics pipelines given the ubiquity of alignment-free methods. Common counting tools [1, 2, 3, 4] usually output compressed data structures containing both k-mers and their counter values. These exact representations, albeit more memory efficient than more naive solutions, remain rather large for in-memory usage even on modern commodity computers. For example, counting all 32-mers in the human reference genome with KMC [2] produces a 20 GB file, well above the 8 GB of RAM most computers have today. For a sufficiently large k, the distribution of k-mer frequencies (k-mer spectrum) of most datasets follow a power-law distribution, where most k-mers appear a small number of times and only a few "heavy hitters" have large counter values. Representing power-law distributed counters with fixed-size words can be inefficient because only few k-mers will effectively have a counter using all allocated bits. Recent solutions try to use small counter words for low frequencies allocating additional space only when needed [5]. In many applications, explicitly storing the k-mers alongside their counters can be avoided if the set of k-mers is static. Minimal Perfect Hash Functions (MPHFs) [6, 7, 8, 9, 10] take advantage of this intuition by producing a bijective mapping between keys and integer values from 1 to the size of the input set. Both keys and values are handled by a data structure external to the MPHF, which does not solve the problem of wasting space for small counters, and needs to be rebuilt from scratch for a new key addition or deletion.

Results: Here we present Set-min sketch, a sketching technique for associative tables between keys and labels where the distribution of the labels is power-law. In our work, we focus on the special case where keys are k-mers, labels are multiplicities, the k-mer spectrum is power-law. We show, both theoretically and experimentally, that our sketch can be more space-efficient than MPHFs and provides better error guarantees compared to equally-dimensioned Count-Min sketches [11]

References

[1] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764, March 2011.

[2] Kokot M, Dlugosz M, and Deorowicz S. KMC 3: counting and manipulating k-mer statistics, September 2017.

[3] Rizk G, Lavenier D, and Chikhi R. DSK: k-mer counting with very low memory usage, March 2013.

[4] Prashant Pandey, Michael A. Bender, Rob Johnson, and Rob Patro. A General-Purpose Counting Filter: Making Every Bit Count. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pages 775–787, Chicago, Illinois, USA, May 2017. Association for Computing Machinery.

[5] Moustafa Shokrof, C. Titus Brown, and Tamer A. Mansour. MQF and buffered MQF: Quotient filters for efficient storage of k-mers with their counts and metadata. *bioRxiv*, page 2020.08.23.263061, August 2020.

[6] Ye Yu, Djamal Belazzougui, Chen Qian, and Qin Zhang. Memory-efficient and Ultra-fast Network Lookup and Forwarding using Othello Hashing. *arXiv:1608.05699 [cs]*, November 2017. arXiv: 1608.05699.

[7] Ye Yu, Jinpeng Liu, Xinan Liu, Yi Zhang, Eamonn Magner, Erik Lehnert, Chen Qian, and Jinze Liu. SeqOthello: querying RNA-seq experiments at scale. *Genome Biology*, 19(1):167, October 2018.

[8] Emmanuel Esposito, Thomas Mueller Graf, and Sebastiano Vigna. RecSplit: Minimal Perfect Hashing via Recursive Splitting. *arXiv:1910.06416 [cs]*, November 2019. arXiv: 1910.06416.

[9] Ingo Müller, Peter Sanders, Robert Schulze, and Wei Zhou. Retrieval and Perfect Hashing Using Fingerprinting. In Joachim Gudmundsson and Jyrki Katajainen, editors, *Experimental Algorithms*, Lecture Notes in Computer Science, pages 138–149, Cham, 2014. Springer International Publishing.

[10] Antoine Limasset, Guillaume Rizk, Rayan Chikhi, and Pierre Peterlongo. Fast and scalable minimal perfect hashing for massive key sets. *arXiv:1702.03154 [cs]*, February 2017. arXiv: 1702.03154.

[11] Graham Cormode and S. Muthukrishnan. Summarizing and mining skewed data streams. pages 44–55, 2005. 5th SIAM International Conference on Data Mining, SDM 2005 ; Conference date: 21-04-2005 Through 23-04-2005.