Masked superstrings as a compact, indexable and dynamic representation of k-mer sets

Ondřej Sladký, Pavel Veselý, Karel Břinda

Abstract

Masked superstrings as a compact, indexable and dynamic representation of *k*-mer sets

Ondřej Sladký[1,2], Pavel Veselý[1], Karel Břinda[3]*

[1]*Charles University in Prague, Czechia*
[2]*ETH Zurich, Zurich, Switzerland*
[3]*Inria, Irisa, Univ. Rennes, Rennes, France*
***Corresponding author**: karel.brinda@inria.fr

Abstract

The exponential growth of DNA sequencing data calls for efficient approaches for their compression and search [1,2]. Modern bioinformatics increasingly uses *k*-merization as a central tokenization technique. This process enables a unified representation of various genomic data types and facilitates data reduction via techniques such as subsampling [3] and sketching [4]. *k*-mer-based methods have been successfully applied in large-scale data search [5–7], metagenomic classification [3,8], infectious disease diagnostics [9,10], and transcript abundance quantification [11,12], among other applications. As the resulting sets may encompass hundreds of billions of *k*-mers [7,13], efficient storage, querying and associated *k*-mer set operations have become critical issues in sequence bioinformatics [14–16].

Traditional information theory-based approaches offer worst-case lower bounds for storing *k*-mer sets [17]. However, much better compression can be achieved in practice by exploiting their non-independentness of *k*-mers via the so-called spectrum-like property [18]. This concept was used in unitigs and later simplitigs / Spectrum Preserving String Sets (SPSS) [19–22], followed by matchtigs / repetitive SPSS (rSPSS) [23]; all compact *k*-mer along non-branching paths in the associated de Bruijn graphs. Importantly, the resulting sequences can be efficiently compressed with general-purpose compressors and are integral to numerous *k*-mer-set data structures [24–26]. However, (r)SPSS, when applied to modern data that incorporate *k*-merization combined with subsampling [3] or sketching [4], do not achieve high compressibility due to their reliance on (*k*-1)-long overlaps, which are typically absent.

In this talk, we will present our recent results on masked superstrings as an efficient representation of unconstrained *k*-mer sets, suitable for indexing and dynamic operations [27–29]. Masked superstrings consist of an approximated shortest superstring of the *k*-mers to be represented, accompanied by a mask distinguishing the true represented *k*-mers from the false positives [27]. Masked superstrings unify and generalize all (r)SPSS [19–23] and further enhance the resulting compression efficacy [27]. Although computing optimal masked superstrings is NP-hard, they can be near-optimally approximated in linear time [27], as implemented in the KmerCamel tool (https://github.com/OndrejSladky/kmercamel). Furthermore, we have used masked superstrings to develop FMSI [29] (https://github.com/OndrejSladky/fmsi), a space- and time-efficient data structure for *k*-mer set indexing using a modified version of the Burrows Wheeler Transform (BWT) [30] called Masked BWT (MBWT). FMSI substantially improves space efficiency compared to the state of the art [25,31–33], while maintaining competitive query times. We will conclude by demonstrating that masked superstrings are amenable to dynamic operations, via an algebraic merge-based framework of so-called function-assigned masked superstrings (*f*-MS) [28]. Overall, our work demonstrates that indexes and representations based on superstrings are a highly general, efficient, and dynamic approach for modern *k*-mer sets, without imposing constraints on their structure.

References

[1] Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol.* **13**, e1002195, 2015. ⟨10.1371/journal.pbio.1002195⟩

[2] Loh, P.-R. *et al.* Compressive genomics. *Nat. Biotechnol.* **30**, 627–630, 2012. ⟨10.1038/nbt.2241⟩

[3] Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46, 2014. ⟨10.1186/gb-2014-15-3-r46⟩

[4] Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132, 2016. ⟨10.1186/s13059-016-0997-x⟩

[5] Bradley, P. *et al.* Ultrafast search of all deposited bacterial and viral genomic data. *Nat. Biotechnol.* **37**, 152–159, 2019. ⟨10.1038/s41587-018-0010-1⟩

[6] Bingmann, T. *et al.* in *String Processing and Information Retrieval* 285–303, Springer International Publishing, 2019. ⟨10.1007/978-3-030-32686-9_21⟩

[7] Karasikov, M. *et al.* Indexing All Life's Known Biological Sequences. *bioRxiv* 2020.10.01.322164, 2024. ⟨10.1101/2020.10.01.322164⟩

[8] Břinda, K. *et al. prophyle/prophyle: ProPhyle 0.3.3.1*. Zenodo, 2021. ⟨10.5281/ZENODO.1045429⟩

[9] Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **6**, 10063, 2015. ⟨10.1038/ncomms10063⟩

[10] Břinda, K. *et al.* Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nature Microbiology* **5**, 455–464, 2020. ⟨10.1038/s41564-019-0656-6⟩

[11] Bray, N. L. *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527, 2016. ⟨10.1038/nbt.3519⟩

[12] Patro, R. *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419, 2017. ⟨10.1038/nmeth.4197⟩

[13] Hunt, M. *et al.* AllTheBacteria - all bacterial genomes assembled, available and searchable. *bioRxiv* 2024. ⟨10.1101/2024.03.08.584059⟩

[14] Marchet, C. *et al.* Data structures based on k-mers for querying large collections of sequencing data sets. *Genome Res.* **31**, 1–12, 2021. ⟨10.1101/gr.260604.119⟩

[15] Marchet, C. Advances in practical k-mer sets: essentials for the curious. *arXiv [q-bio.GN]* 2024. ⟨10.48550/ARXIV.2409.05210⟩

[16] Marchet, C. Advances in colored k-mer sets: essentials for the curious. *arXiv [q-bio.GN]* 2024. ⟨10.48550/ARXIV.2409.05214⟩

[17] Conway, T. C. & Bromage, A. J. Succinct data structures for assembling large genomes. *Bioinformatics* **27**, 479–486, 2011. ⟨10.1093/bioinformatics/btq697⟩

[18] Chikhi, R. *et al.* Data Structures to Represent a Set of k-long DNA Sequences. *ACM Comput. Surv.* **54**, 1–22, 2021. ⟨10.1145/3445967⟩

[19] Břinda, K. Novel computational techniques for mapping and classification of Next-Generation Sequencing data. Preprint , 2016. ⟨10.5281/ZENODO.1045316⟩

[20] Břinda, K. *et al.* Simplitigs as an efficient and scalable representation of de Bruijn graphs. *Genome Biol.* **22**, 96, 2021. ⟨10.1186/s13059-021-02297-z⟩

[21] Rahman, A. & Medevedev, P. Representation of k-Mer sets using spectrum-preserving string sets. *J. Comput. Biol.* **28**, 381–394, 2021. ⟨10.1089/cmb.2020.0431⟩

[22] Schmidt, S. & Alanko, J. N. Eulertigs: minimum plain text representation of k-mer sets without repetitions in linear time. *Algorithms Mol. Biol.* **18**, 5, 2023. ⟨10.1186/s13015-023-00227-1⟩

[23] Schmidt, S. *et al.* Matchtigs: minimum plain text representation of k-mer sets. *Genome Biol.* **24**, 136, 2023. ⟨10.1186/s13059-023-02968-z⟩

[24] Pibiri, G. E. *et al.* Locality-preserving minimal perfect hashing of k-mers. *Bioinformatics* **39**, i534–i543, 2023. ⟨10.1093/bioinformatics/btad219⟩

[25] Pibiri, G. E. Sparse and skew hashing of K-mers. *Bioinformatics* **38**, i185–i194, 2022. ⟨10.1093/bioinformatics/btac245⟩

[26] Alanko, J. N. *et al.* Succinct k-mer Sets Using Subset Rank Queries on the Spectral Burrows-Wheeler Transform *. *bioRxiv* 2022.05.19.492613, 2022. ⟨10.1101/2022.05.19.492613⟩
[27] Sladký, O. *et al.* Masked superstrings as a unified framework for textual k-mer set representations. *bioRxiv* 2023. ⟨10.1101/2023.02.01.526717⟩
[28] Sladký, O. *et al.* Function-Assigned Masked Superstrings as a Versatile and Compact Data Type for k-Mer Sets. *bioRxiv* 2024. ⟨10.1101/2024.03.06.583483⟩
[29] Sladký, O. *et al.* FroM Superstring to Indexing: a space-efficient index for unconstrained*k*-mer sets using the Masked Burrows-Wheeler Transform (MBWT). *bioRxiv* 2024. ⟨10.1101/2024.10.30.621029⟩
[30] Burrows, M. & Wheeler, D. J. *A Block-sorting Lossless Data Compression Algorithm*. 1994.
[31] Alanko, J. N. *et al.* in *SIAM Conference on Applied and Computational Discrete Algorithms (ACDA23)* 225–236, Society for Industrial and Applied Mathematics, 2023. ⟨10.1137/1.9781611977714.20⟩
[32] Martayan, I. *et al.* Conway-Bromage-Lyndon (CBL): an exact, dynamic representation of k-mer sets. *Bioinformatics* **40**, i48–i57, 2024. ⟨10.1093/bioinformatics/btae217⟩
[33] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* btp324, 2009. ⟨10.1093/bioinformatics/btp324⟩