

Faster Computation of Genome Mappability with one Mismatch*

Sahar Hooshmand
dept. of Computer Science
University of Central Florida
Orlando, FL, USA
sahar@cs.ucf.edu

Paniz Abedin
dept. of Computer Science
University of Central Florida
Orlando, FL, USA
paniz@cs.ucf.edu

Daniel Gibney
dept. of Computer Science
University of Central Florida
Orlando, FL, USA
daniel.gibney@ucf.edu

Srinivas Aluru
dept. of Computer Science
Georgia Institute of Technology
Atlanta, GA, USA
aluru@cc.gatech.edu

Sharma V. Thankachan
dept. of Computer Science
University of Central Florida
Orlando, FL, USA
sharma.thankachan@ucf.edu

Abstract—The genome mappability problem refers to cataloging repetitive occurrences of every substring of length m in a genome, and its k -mappability variant extends this to approximate repeats by allowing up to k mismatches. This problem is formulated as follows: Given a sequence $S[1, n]$ of length n over the constant DNA alphabet $\Sigma = \{A, C, G, T\}$, and two integers k and $m \leq n$, output an integer array F_k , such that: $F_k[i] = |\{j \neq i \mid d_H(S[i, i + m - 1], S[j, j + m - 1]) \leq k\}|$ where $d_H(\cdot, \cdot)$ represents the hamming distance. Derrien *et al.* [PLoS one 2012] represented this problem within the framework of genome analysis. In this work we present a provably efficient algorithm for 1-mappability with $O(n \log n)$ worst case run time and $O(n)$ spece. The fundamental technique is the heavy path decomposition on the suffix tree (ST) of S , and the entire work is based on the framework by Thankachan *et al.* [RECOMB 2018]. The previous best known run time is $O(n \log n \log \log n)$ [Alzamel *et al.*, COCOA 2017].

Index Terms—Genome mappability, heavy path decomposition, Hamming distance.

ACKNOWLEDGMENT

This research is supported in part by the U.S. National Science Foundation under CCF-1704552 and CCF-1703489.

REFERENCES

- [1] M. Alzamel, P. Charalampopoulos, C. S. Iliopoulos, S. P. Pissis, J. Radoszewski, and W.-K. Sung. Faster algorithms for 1-mappability of a sequence. In *International Conference on Combinatorial Optimization and Applications*, pages 109–121. Springer, 2017.
- [2] T. Derrien, J. Estellé, S. M. Sola, D. G. Knowles, E. Raineri, R. Guigó, and P. Ribeca. Fast computation and applications of genome mappability. *PLoS one*, 7(1):e30377, 2012.
- [3] S. V. Thankachan, C. Aluru, S. P. Chockalingam, and S. Aluru. Algorithmic framework for approximate matching under bounded edits with applications to sequence analysis. In *Research in Computational Molecular Biology - 22nd Annual International Conference, RECOMB 2018, Paris, France, April 21-24, 2018, Proceedings*, pages 211–224, 2018.