On-Line Pattern Matching on D-Texts

Nadia Pisanti
University of Pisa, Italy

Abstract

The *Elastic Degenerate String Matching* (EDSM) problem is defined as that of finding an occurrence of a pattern P of length m in an ED-text T. A *D-text* (Degenerate text) is a string that actually represents a set of similar and aligned strings (e.g. a pan-genome [5]) by collapsing common fragments into a standard string, and representing variants with sets of alternative substrings. When such substrings are not bound to have the same size, then we talk about *elastic D-strings* (ED-strings). In [6] we gave an $O(nm^2 + N)$ time on-line algorithm for EDSM, where n is the length of T and N is its size, defined as the total number of letters. A fundamental toolkit of our algorithm is the $O(m^2+N)$ time solution of the later called *Active Prefixes* problem (AP). In [2], a $O(m^{1.5}\sqrt{\log m}+N)$ solution for AP was shown, leading to a $O(nm^{1.5}\sqrt{\log m} + N)$ time solution for EDSM. The natural open problem was thus whether the 1.5 exponent could furtherly be decreased. In [3], we prove several properties that answer this and other questions: we give a conditional $O(nm^{1.5} + N)$ lower bound for EDSM, proving that a combinatorial algorithm solving EDSM in $O(nm^{1.5-\epsilon} + N)$ time would break the Boolean Matrix Multiplication (BMM) conjecture; we use this result as a hint to devise a non-combinatorial algorithm that solves EDSM in $O(nm^{1.381} + N)$ time; we do so by successfully combining Fast Fourier Transform and properties of string periodicity.
In my talk I will overview the results above, as well as some interesting side results: the extension to a dictionary rather than a single pattern [7], the introduction of errors [4], and a notion of matching *among* D-strings with its linear time solution [1].

2012 ACM Subject Classification Mathematics of computing

Keywords and phrases pattern matching, elastic-degenerate string, matrix multiplication

Digital Object Identifier 10.4230/LIPIcs.CPM.2021.3

Category Invited Talk

Funding *Nadia Pisanti*: This work is partially supported by the EU funding schemes H2020-MSCA-ITN-2020. Project ALPACA GA 956229, and by the University of Pisa funding scheme PRA (Progetti di Ricerca di Ateneo) Project no. PRA_2020-2021_26.

References

1 Mai Alzamel, Lorraine A. K. Ayad, Giulia Bernardini, Roberto Grossi, Costas S. Iliopoulos, Nadia Pisanti, Solon P. Pissis, and Giovanna Rosone. Comparing degenerate strings. *Fundam. Informaticae*, 175(1-4):41–58, 2020. doi:10.3233/FI-2020-1947.

2 Kotaro Aoyama, Yuto Nakashima, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Faster Online Elastic Degenerate String Matching. In Gonzalo Navarro, David Sankoff, and Binhai Zhu, editors, *Annual Symposium on Combinatorial Pattern Matching (CPM 2018)*, volume 105 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 9:1–9:10, Dagstuhl, Germany, 2018. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.CPM.2018.9.

3 Giulia Bernardini, Pawel Gawrychowski, Nadia Pisanti, Solon P. Pissis, and Giovanna Rosone. Even Faster Elastic-Degenerate String Matching via Fast Matrix Multiplication. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 21:1–21:15, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.ICALP.2019.21.

4 Giulia Bernardini, Nadia Pisanti, Solon P. Pissis, and Giovanna Rosone. Approximate pattern matching on elastic-degenerate text. *Theor. Comput. Sci.*, 812:109–122, 2020. doi:10.1016/j.tcs.2019.08.012.

5 The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings Bioinform.*, 19(1):118–135, 2018. doi:10.1093/bib/bbw089.

6 Roberto Grossi, Costas S. Iliopoulos, Chang Liu, Nadia Pisanti, Solon P. Pissis, Ahmad Retha, Giovanna Rosone, Fatima Vayani, and Luca Versari. On-Line Pattern Matching on Similar Texts. In Juha Kärkkäinen, Jakub Radoszewski, and Wojciech Rytter, editors, *28th Annual Symposium on Combinatorial Pattern Matching (CPM 2017)*, volume 78 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 9:1–9:14, Dagstuhl, Germany, 2017. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.CPM.2017.9.

7 Solon P. Pissis and Ahmad Retha. Dictionary Matching in Elastic-Degenerate Texts with Applications in Searching VCF Files On-line. In Gianlorenzo D'Angelo, editor, *17th International Symposium on Experimental Algorithms (SEA 2018)*, volume 103 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 16:1–16:14, Dagstuhl, Germany, 2018. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.SEA.2018.16.