Enabling FM-Index for Elastic-Degenerate Strings via a new Min/Max Wavelet Tree

Simone Faro[§], Dominik Köppl[†], Thierry Lecroq[‡] and Francesco Pio Marino[§,‡]
[§]University of Catania [†]University of Yamanashi [‡]University of Rouen

Elastic-degenerate (ED) strings generalize classic strings by allowing each position to store a set of up to h strings of arbitrary lengths [2]. An ED string is a restricted version of a regular expression whose expressiveness still causes a combinatorial explosion of possible *resolutions* (i.e., the members of its language), making classical pattern matching and indexing techniques either inefficient or inapplicable. A position in an ED string is called *solid* if it stores a single symbol, otherwise *elastic*. We introduce the Min/Max wavelet tree (MM-WT), a variant of the wavelet tree supporting semantics-aware rank and select on ED strings. Each leaf stores two arrays per symbol c, min_c and max_c, giving for each position the minimum and maximum symbol count across all alternatives. Prefix sums bound the number of occurrences in any resolution of a prefix, so $\mathsf{sm\text{-}rank}(c, i)$ returns an interval $[r_{\min}, r_{\max}]$ of possible ranks, and $\mathsf{sm\text{-}select}(c, j)$ returns an interval for the j-th occurrence of c. On classic strings, *min* and *max* collapse to the same counts and the structure becomes a standard wavelet tree. The structure uses the same space as a subset wavelet tree [1] internally, plus $\mathcal{O}(n\sigma \log h)$ bits for the min/max arrays, and supports both queries in $\mathcal{O}(\log \sigma)$ time.

On top of the MM-WT we devise a modified FM-index for ED strings. Instead of enumerating all complete resolutions of an ED string, we consider two types of suffixes: (1) solid-origin suffixes, which start at solid positions and replace subsequent elastic regions with $*$, and (2) elastic-origin suffixes, which start inside elastic positions and continue within the chosen alternative from that entry point, followed by the remainder of the string represented as before. All suffixes end with a sentinel \$ for correct lexicographic ordering. We compute the Burrows–Wheeler Transform of this collection and store its L column in the MM-WT. Backward search proceeds as in a classical FM-index, but interval updates adopt semantics-aware bounds, yielding a candidate range. Auxiliary arrays track both the depth inside elastic regions and the offset within alternatives, enabling efficient verification. This preserves the efficiency of a traditional FM-index while enabling practical search on ED strings. The MM-WT and the resulting FM-index provide a foundation for scalable indexing of highly variable genomic data and other domains requiring nondeterministic string models.

Ackn.: JSPS KAKENHI `JP25K21150` and `JP23H04378`.

[1] J. N. Alanko, E. Biagi, S. J. Puglisi, and J. Vuohtoniemi. Subset wavelet trees. In *Proc. SEA*, volume 265 of *LIPIcs*, pages 4:1–4:14, 2023.

[2] C. S. Iliopoulos, R. Kundu, and S. P. Pissis. Efficient pattern matching in elastic-degenerate strings. *Inf. Comput.*, 279:104616, 2021.