On Multiseed Lossless Filtration*

Rasmus Pagh

IT University of Copenhagen, Denmark

Abstract. In *approximate string matching* a string $x \in \Sigma^n$ is given and preprocessed in order to support k-approximate match queries: we seek all substrings of x that differ from a query string q in at most k positions. This problem is motivated for example by biological sequence analysis where approximate occurrences of a sequence q are of interest.

Filtration is an approach to approximate string matching that aims to be efficient when most substrings of x have distance to q considerably larger than k. In these approaches a *seed* is used to extract multisets of subequences S_x and S_q from x and q, respectively, such that every k-approximate match gives rise to at least one element in $S_q \cap S_x$. (Elements in S_x are annotated with the substring position(s) they correspond to.) Thus, computing $S_q \cap S_x$ (for example using an index data structure for S_x to look up each element of S_q) gives a set of *candidate* positions for k-approximate matches. The filter is *efficient* if it generates few candidates that do not correspond to k-approximate matches. It is known that filtering can be particularly effective in high-entropy strings such as biological sequences.

In this talk we consider so-called *multiseed* methods where several sequences of sets S_x^i, S_q^i, $i = 1, 2, \ldots$ are extracted from x and q, and candidate matches are found in $\bigcup_i S_q^i \cap S_x^i$. Multiseed methods can yield better filtering efficiency, at the expense of a higher candidate generation cost. While some filtration methods allow a nonzero error probability, we focus on *lossless* methods that are guaranteed to report all k-approximate matches. We present a randomized construction of a set of roughly 2^k seeds for which a substring x' having $k+t$ mismatches with q becomes a candidate match $\Theta(2^{-t})$ times in expectation. Since the method is lossless, every x' with at most k mismatches becomes a candidate at least once. This filtering efficiency is better than previous methods with the same number of seeds for $k > 3$. Finally, we use a general transformation to present a new, improved trade-off between the number of seeds and the filtering efficiency.

* The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 614331.