Inferring large jumbled patterns from small patterns

Dénes Bartha and Péter Burcsi,

Faculty of Informatics
ELTE - Eötvos Loránd University, Budapest

denesb@gmail.com
bupe@inf.elte.hu

In jumbled pattern matching or abelian pattern matching the following problem is considered. A pattern and a text are given as inputs and we are looking for factors of the input text which contain exactly the same letters with the same multiplicity as the pattern, but not necessarily in the same order. For example, the pattern "anna" occurs in "banana" in a jumbled way, matching the "nana" part. The original motivation for this problem comes from the analysis of mass spectrometry data.

More formally, for a string w over a σ-letter (ordered) alphabet Σ, the Parikh vector of w, denoted by $\mathbf{pv}(w)$ is the vector of length σ whose jth coordinate is the number of occurrences of the jth letter of Σ in w. For $w = abcaaabcc$ over the ternary alphabet $\{a, b, c\}$, we have $\mathbf{pv}(w) = (4, 2, 3)$. A Parikh vector p occurs in w if it is the Parikh vector of a factor (i.e. a substring of consecutive letters) of w. For example, if w is as above, then $(2, 1, 0)$ occurs in w, since we have a factor aab, but $(1, 2, 0)$ does not occur in w. The Parikh set $\Pi(w)$ is the set of all Parikh vectors occuring in w. Using this notation, in jumbled pattern matching we test an input pattern Parikh vector for membership in $\Pi(w)$.

The set $\Pi_k(w) \subseteq \Pi(w)$ consists of those Parikh vectors whose weight (the sum of coordinates) is k. That is, these are the Parikh vectors of factors of length exactly k.

In the present paper, we investigate the following problem: given $\Pi_k(w)$ for some values of k and an unknown string w, what can be said about $\Pi_j(w)$ for other values j? We present some theoretical and empirical results.

References

[1] Burcsi, Péter and Cicalese, Ferdinando and Fici, Gabriele and Lipták, Zsuzsanna: *Algorithms for Jumbled Pattern Matching in Strings,* International Journal of Foundations of Computer Science, **23**, 2011.