Problem Description: Algorithms for String Matching with Gaps

S. Tudent T. Hesis

Problem definition Given integers a and b, $0 \leq a \leq b$, a *variable length gap* $g\{a, b\}$ is an arbitrary string over Σ of length between a and b, both inclusive. A *variable length gap pattern* (abbreviated VLG pattern) P is the concatenation of a sequence of strings and variable length gaps, that is, P is of the form

$$P = P_1 \cdot g\{a_1, b_1\} \cdot P_2 \cdot g\{a_2, b_2\} \cdots g\{a_{k-1}, b_{k-1}\} \cdot P_k \ .$$

A VLG pattern P *matches* a substring S of T iff $S = P_1 \cdot G_1 \cdots G_{k-1} \cdot P_k$, where G_i is any string of length between a_i and b_i, $i = 1, \ldots, k-1$. Given a string T and a VLG pattern P, the *variable length gap problem* (VLG problem) is to find all ending positions of substrings in T that match P.

Variable length gaps are frequently used in computational biology applications. For instance, the PROSITE data base supports searching for proteins specified by VLG patterns.

Previous work The simplest approach to solve the VLG problem is to translate P into a regular expression and then use an algorithm for regular expression matching. The regular expression R corresponding to P has length $\Omega(B\sigma + m)$, where $B = \sum_{i=1}^{k-1} b_i$ is the sum of the upper bounds of the gaps in P. Using Thompson's textbook regular expression matching algorithm [4] this leads to an algorithm for the VLG problem using $O(n(B\sigma + m))$ time.

Navarro and Raffinot [3] gave an automata algorithm using $O(n(\frac{m+B}{w} + 1))$ time, where w is the number of bits in a memory word.

Bille and Thorup [1] gave an automata algorithm using $O(n(k\frac{\log w}{w} + \log k) + m \log m + A)$ time and $O(m + A)$ space, where $A = \sum_{i=1}^{k-1} a_i$ is the sum of the lower bounds on the lengths of the gaps.

Morgante et al. [2] gave an algorithm using $O(n \log k + m + \alpha)$ time, where α is the total number of occurrences of the k strings $P_1, \ldots, P_k$ within T. However, unlike the automata based algorithm that only use $O(m + A)$ space, this algorithm use $\Theta(m + \alpha)$ space. Since α typically increases with the length of T, the space usage of this algorithms is likely to quickly become a bottleneck for processing large biological data bases.

Aim of project The goal of this thesis is to construct new and improved algorithms for the VLG problem. The algorithms should be analyzed theoretically and empirically verified. We will implement the algorithm by Morgante et al and a simple automata based algorithm to compare with.

References

[1] P. Bille and M. Thorup. Regular expression matching with multi-strings and intervals. In *Proc. 21st SODA*, 2010.

[2] M. Morgante, A. Policriti, N. Vitacolonna, and A. Zuccolo. Structured motifs search. *J. Comput. Bio.*, 12(8):1065–1082, 2005.

[3] G. Navarro and M. Raffinot. Fast and simple character classes and bounded gaps pattern matching, with applications to protein searching. *J. Comput. Bio.*, 10(6):903–923, 2003.

[4] K. Thompson. Regular expression search algorithm. *Commun. ACM*, 11:419–422, 1968.