Comparing algorithms for Structured Motifs Search

Md. Kawser Habib, Shuhana Azmin

Introduction :

Searching biological sequence(s) is a fundamental task in bioinformatics. Searching for structured motifs is important, among others, in the context of identifying conserved features in biological sequences.

A structured motif $\mathcal{M}$ is usually expressed in the following form:

$$\mathcal{M} = s_1\ [a_1, b_1]\ s_2\ [a_2, b_2]\ \ldots\ s_{k-1}[a_{k-1}, b_{k-1}]\ s_k$$

Where,

$\mathcal{S} = (s_1, \ldots, s_k)$ is a sequence of patterns.
a_i , $b_i \in Z$, $a_i \leq b_i$, for $1 \leq i < k$, gaps between consecutive seeds.

Table 1: Structured Motifs Search

Sequence ($s \in \mathcal{S}$):	GCAT**GC**G**TTA**G**CA**T**CAT**C
Structured Motif ($\mathcal{M}$):	**GC**[0,1]**TTA**[1,4]**CAT**

Occurrences of $\mathcal{M}$ in $\mathcal{S}$ are marked in bold.

Objective:

SimpliSMS[1] :

A simple, lightweight and fast tool for structured motifs search. It identifies all possible search contexts in sequences. SimpLiSMS uses an exact pattern matching algorithm (e.g., the famous KMP algorithm for exact matching) to identify all the occurrences.

sMotif[2] :

sMotif is currently popular software for structured motif search. It search patterns directly by positional joins over an inverted index. It considers variable gap constraints during the positional joins as opposed to building a constraint graph, and it handles missing components efficiently by considering them over patterns.

We compare SimpliSMS and sMotif to understand which one gives more accurate results and takes less time.

Approach:

Since sMotif has been developed for Linux and Mac OS platform and SimpliSMS is developed for Windows platform, it was difficult to compare their performance in different environment. To overcome this hindrance , we develop two different versions of sMotif in C# and C++ language. And run our experiment successfully.

Result:

We have evaluated the performance of SimpLiSMS through extensive experiments. sMotif was crashing for long motifs therefore we couldn't run the experiment for longer gap lengths, it clear from the plots that sMotif time is atregardless of the gap length of number of occurrences. We show our comparison below :

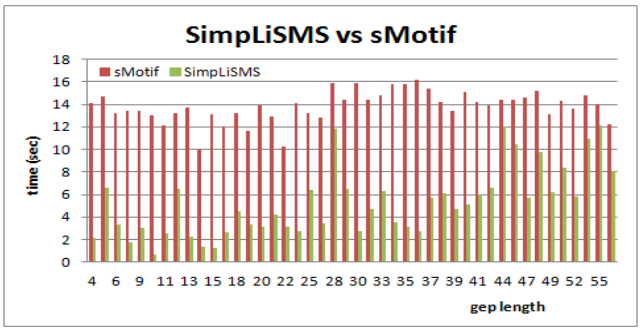

Fig. 1: Comparison of SimpLiSMS and sMotif (time vs. gaps length)

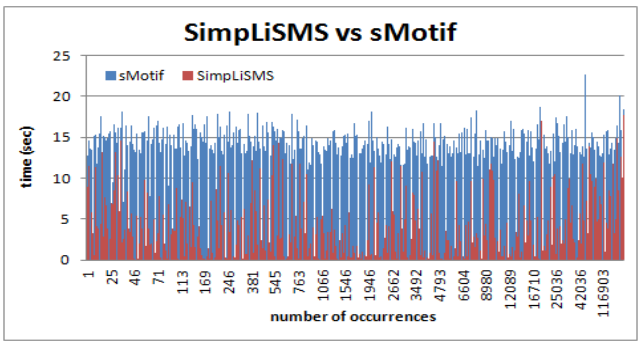

Fig. 2: Comparison of SimpLiSMS and sMotif (time vs. number of occurrences)

Conclusion:

SimpLiSMS is lightweight and faster than sMotif. It finds more number of motifs then sMotif within the same time interval and same gap length.

References:

1. SimpLiSMS: A Simple, Lightweight and Fast Approach for Structured Motifs Searching. Ali Alatabbi, Shuhana Azmin, Md. Kawser Habib, Costas S. Iliopoulos, and M. Sohel Rahma. *Bioinformatics and Biomedical Engineering Lecture Notes in Computer Science Volume 9044, 2015, pp 219-230*
2. sMotif: efficient structured pattern and profile motif search. Yongqiang Zhang and Mohammed J Zaki. *Algorithms for Molecular Biology 2006.*

Department of Computer Science and Engineering (CSE), BUET