Hardness of Optimal Spaced Seed Design

François Nicolas and Éric Rivals

Laboratoire d'Informatique, de Robotique
et de Microélectronique de Montpellier
CNRS - Université Montpellier II
{nicolas, rivals}@lirmm.fr
http://www.lirmm.fr/~rivals

Context

d_{H} denotes the *Hamming distance*.

Data:
A text T, a pattern P of length $|P| = m$ and an integer $k \geq 0$.

Problem:
Find all *k-approximate occurrences* of P in T
= starting positions of substrings S of T such that $d_{\mathrm{H}}(S, P) \leq k$.

Hypotheses:

1. direct algorithms are slow,
2. number of solutions $\ll$ text length.

Filtration by subtrings

q denotes a positive integer.

1. step: **Filtration**.

Compute the positions in T of all q-long substrings common to P and T (the *hits*).

2. step: **Checking**.

Find all k-approximate occurrences of P in T only around the hits.

False positive/False negative

$$\begin{aligned}
T &= \texttt{XXX}\overbrace{\texttt{A}\underline{\texttt{T}}\texttt{GGATT}}^{S_1}\texttt{XXX}\overbrace{\texttt{ACG}\underline{\texttt{T}}\texttt{AT}\underline{\texttt{C}}}^{S_2}\texttt{XXX}\overbrace{\texttt{A}\underline{\texttt{T}}\texttt{CGGAT}\underline{\texttt{C}}}^{S_3}\texttt{XXX}\overbrace{\texttt{ACG}\underline{\texttt{A}}\texttt{ATT}}^{S_4}\texttt{XXX} \\
P &= \texttt{ACGGATT} \\
m &= 7 \\
k &= 1
\end{aligned}$$

- For $q = 3$, S_2 and S_3 are *false positives*.
- For $q = 4$, S_3 is a false positive, but S_2 is not.
- For $q = 4$, S_4 is a *false negative*.

Filtration balance

Conclusion:

q large	$\Rightarrow$	*more*	false negatives	and	*less*	false positives
q small	$\Rightarrow$	*less*	false negatives	and	*more*	false positives

Our setup: Lossless filtration $=$ no false negative.

Example: $m = 7, k = 1, q = 3$.

Counter-example: $m = 7, k = 1, q = 4$.

Seed

Definition: A *seed* is word over $\{\#, -\}$ whose 1st and last letters are a "#".

"#" is a *match*
"−" is a *joker*.

Definition: The *weight* of a seed g, denoted weight(g), is its number of #.

Example: $g := \#\#-\#--\#$ is a seed and weight$(g) = 4$.

$\oplus$ Better filtration [Ma et al. 02, Burkhardt, Kärkkäinen 03, Kucherov et al. 04]

$\ominus$ Requires seed design

Example of grams

Let $g :=$ ##−#−−# and $T :=$ `ACGGATTAC`.

g-grams of T are

- $G_0 :=$ `ACGT`,
- $G_1 :=$ `CGAA`, and
- $G_2 :=$ `GGTC`.

G_i is the g-gram of T at position i.

T	$=$	A	C	G	G	A	T	T	A	C
G_0	$=$	A	C	−	G	−	−	T		
G_1	$=$		C	G	−	A	−	−	A	
G_2	$=$			G	G	−	T	−	−	C

Lossless filtration with a gapped seed

Consider:

$$
\begin{aligned}
T &= \texttt{XXX}\overbrace{\texttt{A\underline{T}GGATT}}^{S_1}\texttt{XXX}\overbrace{\texttt{ACG\underline{T}AT\underline{C}}}^{S_2}\texttt{XXX}\overbrace{\texttt{A\underline{T}GGA\underline{C}T}}^{S_3}\texttt{XXX}\overbrace{\texttt{ACG\underline{A}ATT}}^{S_4}\texttt{XXX} \\
P &= \texttt{ACGGATT} \\
m &= 7 \\
k &= 1
\end{aligned}
$$

and let:

$$g := \#\#-\#\#$$

One achieves lossless filtration with a seed of weight 4.

Similarity

Let S be word of length $|P| = m$.

Definition: The *similarity* between P and S is the word $H_P^S \in \{0, 1\}^m$ defined by:

$$H_P^S[i] = \begin{cases} 1 & \text{si } S[i] = P[i] \\ 0 & \text{si } S[i] \neq P[i] \end{cases} .$$

$$\begin{array}{rcl} S_1 & = & \mathtt{ATGGATT} \\ P & = & \mathtt{ACGGATT} \\ H_P^{S_1} & = & \mathtt{1011111} \end{array} \qquad \begin{array}{rcl} S_2 & = & \mathtt{ATGGACT} \\ P & = & \mathtt{ACGGATT} \\ H_P^{S_2} & = & \mathtt{1011101} \end{array}$$

Detection of similarities

Definition: A seed g *detects* a similarity H if H admits a g-gram without 0.

$$
\boxed{\begin{array}{ccccccccc}
H_1 & = & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\
g & = & & & \# & \# & - & \# & \# \\
 & & & & 1 & 1 & & 1 & 1
\end{array}}
$$

$$
\boxed{\begin{array}{cccccccccccc}
H_2 & = & 1 & 0 & 1 & 1 & 1 & 0 & 1 & & & \\
 & & \# & \# & - & \# & \# & & & \Rightarrow & 10 & 11 \\
 & & & \# & \# & - & \# & \# & & \Rightarrow & 01 & 10 \\
 & & & & \# & \# & - & \# & \# & \Rightarrow & 11 & 01
\end{array}}
$$

g detects $H_P^S \iff S$ and P admit a common g-gram starting at the same position.

(m,k)-similarities

An (m,k)-similarity is a word in $\{\texttt{0},\texttt{1}\}^m$ with k zeros and $m-k$ ones.

Example: `1110110` is a $(7,2)$-similarity.

The number of (m,k)-similarities is $\binom{m}{k}$.

A seed g detects all (m,k)-similarities

$\Updownarrow$

for all text T and pattern P of length m,
g filters all k-approximate occurrences of P in T without loss.

The NON DETECTION problem

NON DETECTION [Kucherov et al., 04]

Instance: A seed g and two integers $0 \leq k \leq m$.

Question: Does it exist an (m,k)-similarity not detected by g?

Our contribution

Theorem: NON DETECTION is NP-complete.

Previous results

The problems of finding:

- the number of (m,k)-similarities left undetected by g
- the minimal number of positions at which any (m,k)-similarity is detected by g

are exponential in $|g|$, but Fixed Parameter Tractable

[Burkhardt, Kärkkäinen, 03] and [Kucherov et al., 04]

Our reduction

SOAPY SET COVER (SSC)

Instance: A collection $\mathcal{G}$ of subsets of $\mathbb{Z}$ and two integers N, $q \geq 0$.

Question: $\exists\, G_1, G_2, \ldots, G_q \in \mathcal{G}$, $\exists\, t_1, t_2, \ldots, t_q \in \mathbb{Z}$ such that $\bigcup_{i=1}^{q}(G_i + t_i)$ contains at least N consecutive integers?

SSC is NP-complete by reduction from SET COVER

(even if $\mathcal{G}$ is a singleton).

Our reduction (2)

Let $(\mathcal{G}, N, q)$ be an instance of SSC with $\mathcal{G} = \{G\}$ and $\min G = 0$.

We construct an instance (g, m, q) of Non Detection with

$$\begin{aligned} g &:= \text{the characteristic function of the reversal of } G \\ m &:= N - 1 + |g| \end{aligned}$$

Example:

$$(G := \{0,3,5,6\}, N := 9, q := 3) \longmapsto (g := \#\#-\#--\#, m := 15, q := 3)$$

Our reduction (example)

$$(G := \{0,3,5,6\},\ N := 9,\ q := 3) \longmapsto (g := \#\#{-}\#{-}{-}\#,\ m := 15,\ q := 3)$$

0	#	#	–	#	–	–	#								
1		#	#	–	#	–	–	#							
2			#	#	–	#	–	–	#						
3				#	#	–	#	–	–	#					
4					#	#	–	#	–	–	#				
5						#	#	–	#	–	–	#			
6							#	#	–	#	–	–	#		
7								#	#	–	#	–	–	#	
8									#	#	–	#	–	–	#

s_1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1
s_2	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1

Region Specific Optimal Seeds

RSOS [Li et al., 04]

Instance: Two integers d and p, a finite set S of similarities.

Solution: A set $\Gamma = \{g_1, g_2, \ldots, g_d\}$ of d seeds, each of weight p.

Measure: The number of similarities in S detected by Γ, to be maximized.

Theorem: RSOS is not approximable within $\frac{e}{e-1} - \varepsilon$
unless $\mathrm{P} = \mathrm{NP}$ [Li et al., 04]

Our contribution: the result holds even with a single seed (*i.e.*, $d = 1$).