Complexities of the Centre and Median String Problems

François Nicolas (nicolas@lirmm.fr) and Eric Rivals (rivals@lirmm.fr)

Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (France)

Outline

1. Preliminaries
   - Edit Distance
   - The Longest Common Subword problem
2. The median and centre string problems
3. Intractability of centre string and median string
4. Intractability of centre string: a sketch of proof
5. Approximating median and centre string
6. Open problems

Edit Distance

Let x and y be words.

The (non-weighted) *edit distance* (also called Levenshtein *distance*) between x et y (denoted $\mathrm{lev}(x, y)$) is the smallest number of single letter *deletions*, *insertions* and *substitutions* needed to transform x into y.
For example, $\mathrm{lev}((\mathtt{01})^n, (\mathtt{10})^n) = 2$ for all integer $n \geq 1$.

Wagner & Fisher's algorithm computes the edit distance $\mathrm{lev}(x, y)$ in polynomial time $\mathrm{O}(|x||y|)$.

The LCS problem

Let w be a word. A *subword* of w is any word obtained from w by deleting between 0 and $|w|$ letters.

Longest Common Subword (o)	
instance:	a non empty finite language L.
solution:	a word s such that for all $w \in L$, s is a subword of w.
measure:	the length of s.

For example 0^n and 1^n are the longest common subwords of the language $\{(01)^n, 1^n 0^n\}$.

LCS (D)	
instance:	a non empty finite language L and a positive integer λ.
question:	Is there a word of length λ wich is a subword of all words in L?

Binary LCS (D) is NP-complete [Maier 1978] and W[1]-hard with respect to the parameter $\# L$ [Pietrzak 2003].

Centre and Median string problem (optimisation)

We are interested in the complexity of the two following consensus problems:

CENTRE STRING (O)	
instance:	a non empty finite language X.
solution:	any word γ.
measure:	$\max_{x \in X} \operatorname{lev}(x, \gamma)$.

and

MEDIAN STRING (O)	
instance:	a non empty finite language X.
solution:	any word μ.
measure:	$\sum_{x \in X} \operatorname{lev}(x, \mu)$.

Centre and Median string problem (decision)

The decision problems associated with CENTRE STRING (O) and MEDIAN STRING (O) are

CENTRE STRING (D)	
instance:	a non empty finite language X and a positive integer d.
question:	Is there a word γ such that $\max_{x \in X} \operatorname{lev}(x, \gamma) \leq d$?

and

MEDIAN STRING (D)	
instance:	a non empty finite language X and a positive integer d.
question:	Is there a word μ such that $\sum_{x \in X} \operatorname{lev}(x, \mu) \leq d$?

Known results

MEDIAN STRING (O) can be solved in time $\mathrm{O}\left(2^{\# X} \prod_{x \in X} |x|\right)$ by dynamic programming [Sankoff-Kruskal 1983].

CENTRE STRING (D) is NP-complete even restricted to languages X with alphabet size 4 [de la Higuera - Casacuberta 2000]

MEDIAN STRING (D) is NP-complete for unbounded alphabet size (infinite alphabet) [de la Higuera - Casacuberta 2000]

MEDIAN STRING (D) is NP-complete even restricted to languages X with alphabet size 7 but the non-weighted edit distance is replaced by a conveniently weighted edit distance [Sim - Park 1999]

Our results

Regular complexity: Binary CENTRE STRING (D) and binary MEDIAN STRING (D) are both NP-complete.

Meaning: one of these problems can be solved in time $\mathrm{O}(|X|^{\alpha})$ where α is a positive constant if and only if $\mathrm{P} = \mathrm{NP}$.

Parameterized complexity: Binary CENTRE STRING (D) and binary MEDIAN STRING (D) are both W[1]-hard with respect to parameter $\# X$.

Meaning: if one of these problems can be solved in time $\mathrm{O}(f(\# X)\,|X|^{\alpha})$ where α is a positive constant and f an arbitrary function then $\mathrm{FPT} = \mathrm{W}[1]$.

Intractability of CENTER STRING (sketch of the proof)

At first we reduce binary LCS (D) to binary LCS0 (D):

LCS0 (D)	
instance:	a non empty finite language K such that all words in K share the same even length $2k$.
question:	Does there exists a word of length k which is a subword of all words in K?

Given an instance (L, λ) of binary LCS (D) we construct an instance :

$$K := \bigcup_{x \in L} \left\{ x0^{2\lambda+m-|x|}, x1^{2\lambda+m-|x|} \right\} 0^m$$

of binary LCS0 (D) where $m := \max_{x \in L} |x|$ (and $k = \lambda + n$).

Intractability of CENTRE STRING (continued)

We reduce binary LCS0 (D) to binary CENTRE STRING (D).
Given an instance K of binary LCS0 (D) we construct an instance :

$$(X, d) := (K \cup \{\varepsilon\}, k)$$

of binary CENTRE STRING (D) where k is such that all words in K are of length $2k$.

The proof relies on the following lemma :

- for all words x, y, we have $\mathrm{lev}(x, y) \geq |x| - |y|$ and,
- if $\mathrm{lev}(x, y) = |x| - |y|$ then y is a subword of x.

Approximation

CENTRE STRING (O) and MEDIAN STRING (O) are 2-approximable:
compute respectively

$$\operatorname*{argmin}_{x_0 \in X} \left(\max_{x \in X} \operatorname{lev}(x_0, x) \right)$$

and

$$\operatorname*{argmin}_{x_0 \in X} \left(\sum_{x \in X} \operatorname{lev}(x_0, x) \right)$$

There exists a P.T.A.S. (??) for MEDIAN STRING (O) [Li-Ma-Wang 2001]

Open problems

Regular Complexity: Do CENTRE STRING (D) and MEDIAN STRING (D) remain NP-complete if we replace the non-weighted edit distance by any weighted edit distance?

Parameterized complexity: What is the parameterized complexity of CENTRE STRING (D) and MEDIAN STRING (D) with respect to the distance parameter d?

Approximability: Does there exists a P.T.A.S. for CENTRE STRING (O)?