**Consensus Patterns** parameterized by input string length is W[1]-hard.

Laurent Bulteau

February 28, 2017

We consider the *Consensus Patterns* problem, where, given a set of input strings, one is asked to extract a long enough pattern which appears (with some errors) in all strings. Formally, the problem is defined as follows:

**Consensus Patterns**

**Input:** Strings $S_1, \ldots, S_n$ of length at most $\ell$, integers $m$ and $d$.

**Output:** Length-$m$ string $S$ and integers $(j_1, \ldots, j_n)$ such that $\sum_{i=1}^n \text{Ham}(S, S_i[j_i..j_i+m-1]) \leq d$

Where $\text{Ham}()$ denotes the Hamming distance and $S[a..b]$ is the substring of $S$ starting in $a$ and ending in $b$. This problem is one of many variations of the well-studied *Consensus String* problem. It is similar to *Consensus Substring* in that the target string must be close to a substring of each input string (rather than the whole string). However, in the latter problem the distance to each input string is bounded, rather than the sum of the distances in our case.

We look at this problem from the parameterized complexity viewpoint, more precisely for parameter $\ell$. Recall that *Consensus Substring* is FPT for parameter $k$ [3]. See [1] for an overview of the variants of *Consensus String*, and [4] for recent advances on parameterized aspects of *Consensus Substring* and *Consensus Patterns*. We prove the following result.

**Theorem 1.** *Consensus Patterns*($\ell$) is W[1]-hard.

By reduction from *Multi-Colored Clique*. We are given a graph $G = (V, E)$, with a partition (coloring) $V = V_1 \cup V_2 \cup \ldots \cup V_k$, such that no edge has both endpoints of the same color. Assume that $|V_h| = n$ for all $h \in [k]$. Write $V_h = \{v_{h,1}, v_{h,2}, \ldots, v_{h,n}\}$, i.e. each vertex has an index depending both on its color and its rank within its color. Let $m = |E|$. *Multi-Colored Clique* is W[1]-hard for parameter $k$ [2]. See Figure [1] for an example of the reduction.

We build an alphabet $\Sigma$ containing $V$ (i.e., one symbol per vertex) and two special characters $\$$ and $\circ$.

Define string $V_i = $\$$v_{1,i}v_{2,i} \ldots v_{k,i}$. Let $e = (v_{h,i}, v_{h',i'})$ be the $j$th edge of $E$, $j \in [m]$. Define $E_j$ as the string starting with $\$$, followed by $k+1$ characters: all $\circ$, except for two positions: $E_j[k + h + 1] = v_{h,i}$ and $E_j[h' + 2] = v_{h',i'}$. 

1
Repeat $N = 36$ times

$\mathcal{V}_1 : \$ a c e
\mathcal{V}_2 : \$ b d f
\mathcal{E}_1 : \$ o a c o
\mathcal{E}_2 : \$ o a d o
\mathcal{E}_3 : \$ o a e o
\mathcal{E}_4 : \$ o b c o
\mathcal{E}_5 : \$ o b e o
\mathcal{E}_6 : \$ o c f o
\mathcal{E}_7 : \$ o d e o

$S : \$ a d e

Figure 1: Illustration of the parameterized reduction from an instance of $k$-Colored Clique (left) to Consensus Patterns using the string length as a parameter (right). An optimal solution $S = ade$ and its alignment with each input string is given (positions producing a match are in bold). Note that vertices $\{a, c, e\}$ form a clique in $G$.

Let $N = m(k + 2) + 1$. The instance $\mathcal{I}$ of Consensus Patterns contains $N$ occurrences of strings $\mathcal{V}_i$, $i \in [n]$, and one occurrence of strings $\mathcal{E}_j$, $j \in [m]$. The target length is $m = k + 1$.

Note that due to the large value of $N$, any solution $S$ must have a minimal distance to the set of strings $\{\mathcal{V}_i \mid i \in [n]\}$. Otherwise, (if it is, say, at the minimum distance plus one), the distance to the whole instance $\mathcal{I}$ increases by at least $N$, which cannot be compensated by the remaining strings $\mathcal{E}_j$ (which have size $m(k + 2) < N$). Hence we first enumerate the optimal solutions for the set $\{\mathcal{V}_i \mid i \in [n]\}$.

**Lemma 1.** The Consensus Patterns of $\{\mathcal{V}_i \mid i \in [n]\}$ (i.e., the strings of length $k + 1$ at minimum total distance from strings $\mathcal{V}_i$) are the strings of the form $S = v_{i_1,i_1} \ldots v_{i_k,i_k}$ with $i_1, \ldots, i_k \in [n]$. Such a string has a total distance of $(n - 1)k$.

**Proof.** Since all strings in $\{\mathcal{V}_i \mid i \in [n]\}$ have length $k + 1$, any consensus pattern $S$ must be aligned with $\mathcal{V}_i$ from the very first character. Hence $S$ is a consensus string of $\{\mathcal{V}_i \mid i \in [n]\}$. The consensus strings of this set are obtained by taking the majority character at each position. Thus, $S[1] = \#$, and, for all $h \in [k]$, there exists $i_h$ such that $S[h + 1] = \{v_{h,i_h}\}$.

Consider now an optimal solution $S$ for $\mathcal{I}$. Let $\{i_h \mid h \in [k]\}$ be the set of indices as obtained from the lemma above. We show that the set of vertices $K = \{v_{h,i_h} \mid h \in [k]\}$ forms a clique of $G$ iff the distance is below a certain
threshold. To this end, we compute the best possible alignment between $S$ and each string $E_j$.

**Lemma 2.** Let $j \in [m]$. If both endpoints of edge $e_j$ are in $K$ then there exists an alignment of $S$ at distance $k−1$ from $E_j$, otherwise the best possible alignment has distance $k$.

**Proof.** Let $h_s, h_t, i_s, i_t$ be such that $e_j = (v_{h_s, i_s}, v_{h_t, i_t})$. There are two possible alignments of $S$ with $E_j$: $S[1]$ is aligned either with $E_j[1]$ or with $E_j[2]$. We compute the distance in both cases.

If $S[1]$ is aligned with $E_j[1]$, then there is exactly one common character, namely $S[1] = E_j[1] = \$. Indeed, for all $h \in [k]$, $S[h+1] \in V_h$ and $E_j[h+1] \in V_{h-1} \cup \{x\}$, hence these two characters are different. The distance in this case is $k$.

If $S[1]$ is aligned with $E_j[2]$, then first note that $S[1] = \$ \neq x = E_j[2]$. Consider index $h_s$. If $i_s = i_{h_s}$, then $S[h_s] = v_{h_s,i_{h_s}} = E_j[h_s+1]$, otherwise $S[h_s] = E_j[h_s+1]$. Similarly for $h_t$, $S[h_t] = E_j[h_t+1]$ iff $i_t = i_{h_t}$. For other values of $h$ (i.e. $h \in [k] \setminus \{h_s, h_t\}$), $S[h] = E_j[h+1]$. The distance is thus $k−1$ iff $i_s = i_{h_s}$ and $i_t = i_{h_t}$, it is at least $k$ otherwise.

Overall, if $i_s = i_{h_s}$ and $i_t = i_{h_t}$ the optimal alignment has distance $k−1$, otherwise the optimal alignment has distance $k$.

We can now conclude the proof. Let $S$ be an optimal solution of Consensus Pattern for instance $I$ and $K$ its corresponding set of vertices. The distance from $S$ to the $N$ copies of strings $V_i$ is $N(n−1)k$. The distance between $S$ and $E_j$ is $k−1$ if both endpoints of $e_j$ are in $K$, and $k$ otherwise. $|E(K)|$ is the number of edges with both endpoints in $K$: the total distance from $S$ to strings $E_j$ is thus $mk − |E(K)|$, and the total distance from $S$ to $I$ is $N(n−1)k + mk − |E(K)|$.

Overall, the optimal distance is at most $N(n−1)k + mk − \frac{k(k−1)}{2}$ if, and only if, $G$ contains a size-$k$ set of vertices $K$ with $|E(K)| \geq \frac{k(k−1)}{2}$, i.e. if $G$ contains a clique.

**References**

[1] Laurent Bulteau, Falk Hüffner, Christian Komusiewicz, and Rolf Niedermeier. Multivariate algorithmics for np-hard string problems. *Bulletin of the EATCS*, 114, 2014.

[2] Rodney G Downey and Michael Ralph Fellows. *Parameterized complexity*. Springer Science & Business Media, 2012.

[3] Patricia A. Evans, Andrew D. Smith, and Harold T. Wareham. On the complexity of finding common approximate substrings. *Theor. Comput. Sci.*, 306(1-3):407–430, 2003.
[4] Markus L. Schmid. Finding consensus strings with small length difference between input and solution strings. In MFCS 2015, Part II, volume 9235 of LNCS, pages 542–554. Springer, 2015.