Consensus Strings with Small Maximum Distance and Small Distance Sum

Laurent Bulteau[1], **Markus L. Schmid**[2]

[1] Université Paris-Est, France
[2] Trier University, Germany

MFCS 2018

Consensus String Problems

Input:	A set (multi-set) of strings.
Output:	A string that is a good **consensus** of the input strings.

Consensus String Problems

Input: A set (multi-set) of strings.
Output: A string that is a good **consensus** of the input strings.

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a

Consensus String Problems

Input: A set (multi-set) of strings.

Output: A string that is a good **consensus** of the input strings.

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a
s	c	b	c	a	b	c	a

Consensus String Problems

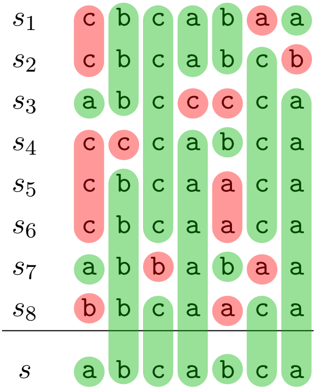
Input: A set (multi-set) of strings.
Output: A string that is a good **consensus** of the input strings.

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a
s	c	b	c	a	b	c	a

Consensus String Problems

Input: A set (multi-set) of strings.

Output: A string that is a good **consensus** of the input strings.



Basic Notations

Standard string notations:

Σ	finite alphabet
Σ^*	words over Σ
Σ^n	$\{w \in \Sigma^* \mid \lvert w \rvert = n\}$
$\Sigma^{\leq n}$	$\bigcup_{i=0}^{n} \Sigma^i$
$\mathsf{d_H}(u, v)$	Hamming distance
$\preceq$	substring relation, $u \preceq v \Leftrightarrow v = xuy$

Basic Notations

Standard string notations:

Σ	finite alphabet
Σ^*	words over Σ
Σ^n	$\{w \in \Sigma^* \mid \lvert w \rvert = n\}$
$\Sigma^{\leq n}$	$\bigcup_{i=0}^{n} \Sigma^i$
$\mathsf{d_H}(u, v)$	Hamming distance
$\preceq$	substring relation, $u \preceq v \Leftrightarrow v = xuy$

For multi-set $S \subseteq \Sigma^\ell$ and $v \in \Sigma^\ell$:

$\mathsf{r_H}(v, S) = \max\{\mathsf{d_H}(v, u) \mid u \in S\}$	**radius** of S (w.r.t. v)
$\mathsf{s_H}(v, S) = \sum_{u \in S} \mathsf{d_H}(v, u)$	**distance sum** of S (w.r.t. v)

The Closest String Problem

(r, s)-Closest String

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^\ell$, $\ell \in \mathbb{N}$, $d_r, d_s \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^\ell$ with $\mathsf{r_H}(s, S) \leq d_r$ and $\mathsf{s_H}(s, S) \leq d_s$?

The Closest String Problem

(r, s)-Closest String

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^\ell$, $\ell \in \mathbb{N}$, $d_r, d_s \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^\ell$ with $\mathsf{r_H}(s, S) \leq d_r$ and $\mathsf{s_H}(s, S) \leq d_s$?

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a

$k = 8$
$\ell = 7$
$d_r = 2$
$d_s = 20$

The Closest String Problem

(r, s)-Closest String

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^\ell$, $\ell \in \mathbb{N}$, $d_r, d_s \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^\ell$ with $\mathsf{r_H}(s, S) \leq d_r$ and $\mathsf{s_H}(s, S) \leq d_s$?

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a
s	a	b	c	a	b	c	a

$k = 8$
$\ell = 7$
$d_r = 2$
$d_s = 20$

The Closest String Problem

(r, s)-Closest String

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^\ell$, $\ell \in \mathbb{N}$, $d_r, d_s \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^\ell$ with $\mathsf{r_H}(s, S) \leq d_r$ and $\mathsf{s_H}(s, S) \leq d_s$?

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a
s	a	b	c	a	b	c	a

$k = 8$
$\ell = 7$
$d_r = 2$
$d_s = 20$

$\mathsf{r_H}(s, S) = 2$
$\mathsf{s_H}(s, S) = 16$

The "Substring Variant"

(r,s)-Closest Substring

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^{\leq \ell}$, $\ell \in \mathbb{N}$, $d_r, d_s, m \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^m$, multi-set $S' = \{s'_i \mid s'_i \preceq s_i, 1 \leq i \leq k\} \subseteq \Sigma^m$ with $\mathsf{r}_\mathsf{H}(s, S') \leq d_r$ and $\mathsf{s}_\mathsf{H}(s, S') \leq d_s$?

The "Substring Variant"

(r, s)-Closest Substring

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^{\leq \ell}$, $\ell \in \mathbb{N}$, $d_r, d_s, m \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^m$, multi-set $S' = \{s'_i \mid s'_i \preceq s_i, 1 \leq i \leq k\} \subseteq \Sigma^m$ with $\mathsf{r_H}(s, S') \leq d_r$ and $\mathsf{s_H}(s, S') \leq d_s$?

s_1	a	a	c	b	c	a	b	a	a		
s_2	b	c	b	c	a	b	c	b			
s_3	a	a	b	c	c						
s_4	c	c	c	a	b	c	a	c			
s_5	c	c	b	c	a	a	c	a			
s_6	a	c	b	c	a	a					
s_7	a	a	b	b	a	b	a	a			
s_8	b	b	b	c	a	a	c	a	c	c	b

$k = 8$
$\ell = 11$
$m = 4$
$d_r = 3$
$d_s = 7$

The "Substring Variant"

(r, s)-Closest Substring

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^{\leq \ell}$, $\ell \in \mathbb{N}$, $d_r, d_s, m \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^m$, multi-set $S' = \{s'_i \mid s'_i \preceq s_i, 1 \leq i \leq k\} \subseteq \Sigma^m$ with $\mathsf{r_H}(s, S') \leq d_r$ and $\mathsf{s_H}(s, S') \leq d_s$?

s_1	a	a	c	b	c	a	b	a	a		
s_2	b	c	b	c	a	b	c	b			
s_3	a	a	b	c	c						
s_4	c	c	c	a	b	c	a	c			
s_5	c	c	b	c	a	a	c	a			
s_6	a	c	b	c	a	a					
s_7	a	a	b	b	a	b	a	a			
s_8	b	b	b	c	a	a	c	a	c	c	b
s	a	b	c	a							

$k = 8$
$\ell = 11$
$m = 4$
$d_r = 3$
$d_s = 7$

The "Substring Variant"

(r, s)-Closest Substring

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^{\leq \ell}$, $\ell \in \mathbb{N}$, $d_r, d_s, m \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^m$, multi-set $S' = \{s'_i \mid s'_i \preceq s_i, 1 \leq i \leq k\} \subseteq \Sigma^m$ with $\mathsf{r_H}(s, S') \leq d_r$ and $\mathsf{s_H}(s, S') \leq d_s$?

s_1	a	a	c	b	c	**a**	**b**	**a**	**a**		
s_2	b	c	b	c	**a**	**b**	**c**	**b**			
s_3	a	**a**	**b**	**c**	**c**						
s_4	c	c	c	**a**	**b**	**c**	**a**	c			
s_5	c	c	b	c	**a**	**a**	**c**	**a**			
s_6	a	**c**	**b**	**c**	**a**	a					
s_7	a	a	b	b	**a**	**b**	**a**	**a**			
s_8	b	**b**	**b**	**c**	**a**	a	c	a	c	c	b
s	a	b	c	a							

$k = 8$
$\ell = 11$
$m = 4$
$d_r = 3$
$d_s = 7$

The "Substring Variant"

(r, s)-Closest Substring

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^{\leq \ell}$, $\ell \in \mathbb{N}$, $d_r, d_s, m \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^m$, multi-set $S' = \{s'_i \mid s'_i \preceq s_i, 1 \leq i \leq k\} \subseteq \Sigma^m$ with $\mathsf{r_H}(s, S') \leq d_r$ and $\mathsf{s_H}(s, S') \leq d_s$?

s_1	a	a	c	b	c	a	b	a	a		
s_2	b	c	b	c	a	b	c	b			
s_3	a	a	b	c	c						
s_4	c	c	c	a	b	c	a	c			
s_5	c	c	b	c	a	a	c	a			
s_6	a	c	b	c	a	a					
s_7	a	a	b	b	a	b	a	a			
s_8	b	b	b	c	a	a	c	a	c	c	b
s	a	b	c	a							

$k = 8$
$\ell = 11$
$m = 4$
$d_r = 3$
$d_s = 7$

$\mathsf{r_H}(s, S') = 1$
$\mathsf{s_H}(s, S') = 7$

The "Outlier Variant"

(r, s)-Closest String with Outliers

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^\ell$, $\ell \in \mathbb{N}$, $d_r, d_s, t \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^\ell$, $S' \subseteq S$ with $|S'| = k - t$ with $\mathsf{r}_\mathsf{H}(s, S') \leq d_r$ and $\mathsf{s}_\mathsf{H}(s, S') \leq d_s$?

The "Outlier Variant"

(r, s)-Closest String with Outliers

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^\ell$, $\ell \in \mathbb{N}$, $d_r, d_s, t \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^\ell$, $S' \subseteq S$ with $|S'| = k - t$ with $\mathsf{r_H}(s, S') \leq d_r$ and $\mathsf{s_H}(s, S') \leq d_s$?

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a

$k = 8$
$\ell = 7$
$t = 2$
$d_r = 2$
$d_s = 8$

The “Outlier Variant”

(r, s)-Closest String with Outliers

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^\ell$, $\ell \in \mathbb{N}$, $d_r, d_s, t \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^\ell$, $S' \subseteq S$ with $|S'| = k - t$ with $\mathsf{r_H}(s, S') \leq d_r$ and $\mathsf{s_H}(s, S') \leq d_s$?

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
[redacted]							
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
[redacted]							
s_8	b	b	c	a	a	c	a

$k = 8$
$\ell = 7$
$t = 2$
$d_r = 2$
$d_s = 8$

The "Outlier Variant"

(r,s)-Closest String with Outliers

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^\ell$, $\ell \in \mathbb{N}$, $d_r, d_s, t \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^\ell$, $S' \subseteq S$ with $|S'| = k - t$ with $\mathsf{r}_\mathsf{H}(s, S') \leq d_r$ and $\mathsf{s}_\mathsf{H}(s, S') \leq d_s$?

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
[redacted]							
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
[redacted]							
s_8	b	b	c	a	a	c	a
s	c	b	c	a	b	c	a

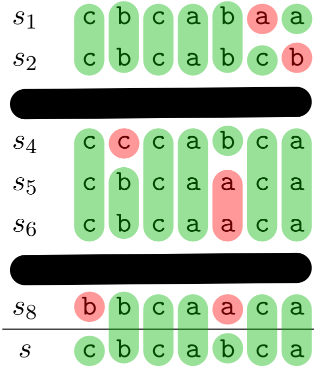
$k = 8$
$\ell = 7$
$t = 2$
$d_r = 2$
$d_s = 8$

The "Outlier Variant"

(r, s)-Closest String with Outliers

Instance: Multi-set $S = \{s_i \mid 1 \leq i \leq k\} \subseteq \Sigma^\ell$, $\ell \in \mathbb{N}$, $d_r, d_s, t \in \mathbb{N}$.

Question: Is there an $s \in \Sigma^\ell$, $S' \subseteq S$ with $|S'| = k - t$ with $\mathsf{r_H}(s, S') \leq d_r$ and $\mathsf{s_H}(s, S') \leq d_s$?



$k = 8$
$\ell = 7$
$t = 2$
$d_r = 2$
$d_s = 8$

$\mathsf{r_H}(s, S) = 2$
$\mathsf{s_H}(s, S) = 7$

The "r- and s-Variants"

(r)-Closest String:	(r, s)-Closest String without d_s bound
(s)-Closest String:	(r, s)-Closest String without d_r bound

Likewise for substring- and outlier-variants.

The "r- and s-Variants"

(r)-Closest String: (r, s)-Closest String without d_s bound
(s)-Closest String: (r, s)-Closest String without d_r bound

Likewise for substring- and outlier-variants.

The (r)- and (s)-variants are intensively investigated:

our terminology	common in literature
(r)-Closest String	Closest String
(r)-Closest Substring	Closest Substring
(s)-Closest Substring	Consensus Patterns

The "r- and s-Variants"

(r)-Closest String: (r, s)-Closest String without d_s bound
(s)-Closest String: (r, s)-Closest String without d_r bound

Likewise for substring- and outlier-variants.

The (r)- and (s)-variants are intensively investigated:

our terminology	common in literature
(r)-Closest String	Closest String
(r)-Closest Substring	Closest Substring
(s)-Closest Substring	Consensus Patterns

Hardness

All these problems are NP-hard
(except (s)-Closest String, which is trivial).

Parameters

k	number of input strings
ℓ	length of input strings
d_r	radius bound
d_s	distance sum bound
$\|\Sigma\|$	alphabet size
m	substring length ((r, s)-Closest Substring)
t	number of outliers ((r, s)-Closest String-wo)
$k - t$	number of inliers ((r, s)-Closest String-wo)

Parameters

k	number of input strings
ℓ	length of input strings
d_r	radius bound
d_s	distance sum bound
$\lvert\Sigma\rvert$	alphabet size
m	substring length ((r, s)-Closest Substring)
t	number of outliers ((r, s)-Closest String-wo)
$k - t$	number of inliers ((r, s)-Closest String-wo)

Notation:
(r, s)-Closest String($p_1, p_2, \ldots$) means (r, s)-Closest String parameterised by $p_1, p_2, \ldots$.

Parameters

k	number of input strings
ℓ	length of input strings
d_r	radius bound
d_s	distance sum bound
$\|\Sigma\|$	alphabet size
m	substring length ((r, s)-Closest Substring)
t	number of outliers ((r, s)-Closest String-wo)
$k - t$	number of inliers ((r, s)-Closest String-wo)

Notation:
(r, s)-Closest String($p_1, p_2, \dots$) means (r, s)-Closest String parameterised by $p_1, p_2, \dots$.

fixed-parameter tractable: $\exists$ algorithm with running time $f(k) \times p(|x|)$ for recursive f and polynomial p (x is input and k the parameter).

W[1]-hardness $\Rightarrow$ *not* fixed parameter tractable.

State of the Art

Previous literature:

(r)-CLOSEST STRING
(s)-CLOSEST STRING
(r)-CLOSEST SUBSTRING
(s)-CLOSEST SUBSTRING
((s)-CLOSEST SUBSTRING(ℓ, m))

State of the Art

Previous literature:

(r)-Closest String
(s)-Closest String
(r)-Closest Substring
(s)-Closest Substring
((s)-Closest Substring(ℓ, m))

(r, s)-Closest String
(r, s)-Closest Substring

State of the Art

Previous literature:

(r)-Closest String
(s)-Closest String
(r)-Closest Substring
(s)-Closest Substring
((s)-Closest Substring(ℓ, m))

(r, s)-Closest String
(r, s)-Closest Substring

(r)-Closest String-wo

(s)-Closest String-wo
(r, s)-Closest String-wo

State of the Art

Previous literature:

(r)-Closest String
(s)-Closest String
(r)-Closest Substring
(s)-Closest Substring
((s)-Closest Substring(ℓ, m))

(r, s)-Closest String
(r, s)-Closest Substring

(r)-Closest String-wo

(s)-Closest String-wo
(r, s)-Closest String-wo

Our Contribution:

(r, s)-Closest String
(r, s)-Closest Substring

(s)-Closest Substring(ℓ, m)

State of the Art

Previous literature:

(r)-Closest String
(s)-Closest String
(r)-Closest Substring
(s)-Closest Substring
((s)-Closest Substring(ℓ, m))

(r, s)-Closest String
(r, s)-Closest Substring

(r)-Closest String-wo

(s)-Closest String-wo
(r, s)-Closest String-wo

Our Contribution:

(r, s)-Closest String
(r, s)-Closest Substring

(s)-Closest Substring(ℓ, m)

(r)-Closest String-wo
(s)-Closest String-wo
(r, s)-Closest String-wo

Results for (r, s)-Closest String

k	d_r	d_s	$\lvert\Sigma\rvert$	ℓ	Result
p	–	–	–	–	FPT
–	**p**	–	–	–	FPT
–	–	**p**	–	–	FPT
–	–	–	2	–	NP-hard
–	–	–	–	**p**	FPT

Results for (r, s)-Closest String

k	d_r	d_s	$\|\Sigma\|$	ℓ	Result
p	–	–	–	–	FPT
–	**p**	–	–	–	FPT
–	–	**p**	–	–	FPT
–	–	–	2	–	NP-hard
–	–	–	–	**p**	FPT

Branching Algorithm

Fpt-branching algorithm for (r)-CLOSEST STRING(d_r):[1]

$d_r = 2$

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a
s	a	b	c	a	b	c	a

[1] Gramm, Niedermeier, Rossmanith, Algorithmica, 2003.

Branching Algorithm

Fpt-branching algorithm for (r)-Closest String(d_r):[1]

$d_r = 2$

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a
s	a	b	c	a	b	c	a

$s' = \texttt{c b c a b a a}$

[1]Gramm, Niedermeier, Rossmanith, Algorithmica, 2003.

Branching Algorithm

Fpt-branching algorithm for (r)-Closest String(d_r):[1]

$d_r = 2$

s_1	c	b	c	a	b	a	a
s_2	c	b	c	a	b	c	b
s_3	a	b	c	c	c	c	a
s_4	c	c	c	a	b	c	a
s_5	c	b	c	a	a	c	a
s_6	c	b	c	a	a	c	a
s_7	a	b	b	a	b	a	a
s_8	b	b	c	a	a	c	a
s	a	b	c	a	b	c	a

$s' =$ c b c a b a a

[1] Gramm, Niedermeier, Rossmanith, Algorithmica, 2003.

Extended Branching Algorithm

Goal: Extend branching algorithm to (r, s)-Closest String-wo(d_r, t).

Extended Branching Algorithm

Goal: Extend branching algorithm to (r, s)-Closest String-wo(d_r, t).

Choice of outliers can be added to the branching (t is a parameter).[2]

General branching similar:
branch by $d_r + 1$ mismatches between candidate and input string.

---

[2]Boucher and Ma, BMC Bioinformatics, 2011.

Extended Branching Algorithm

Goal: Extend branching algorithm to (r, s)-Closest String-wo(d_r, t).

Choice of outliers can be added to the branching (t is a parameter).[2]

General branching similar:
branch by $d_r + 1$ mismatches between candidate and input string.

Main problem: What is a good initial candidate string?

some input string $\rightsquigarrow$ how do we satisfy d_s bound?

some string with low distance sum $\rightsquigarrow$ how to bound depth of branching?

[2]Boucher and Ma, BMC Bioinformatics, 2011.

Extended Branching Algorithm

Majority string s_m:
pick a most frequent symbol in each column.

Example:

s_1	d	b	a	b	b	b	b
s_2	d	a	a	b	c	c	d
s_3	d	a	a	b	c	c	d
s_4	a	a	c	c	c	c	d
s_5	a	a	c	b	c	c	d
s_6	a	c	a	b	d	b	d
s_m	d	a	a	b	c	c	d

Extended Branching Algorithm

Majority string s_m:
pick a most frequent symbol in each column.

Example:

s_1	d	b	a	b	b	b	b
s_2	d	a	a	b	c	c	d
s_3	d	a	a	b	c	c	d
s_4	a	a	c	c	c	c	d
s_5	a	a	c	b	c	c	d
s_6	a	c	a	b	d	b	d
s_m	d	a	a	b	c	c	d

Lemma

$\mathsf{r_H}(s, S) \leq d_r \Rightarrow \mathsf{d_H}(s_m, s) \leq 2d_r$.

Extended Branching Algorithm

Majority string s_m:
pick a most frequent symbol in each column.

Example:

s_1	d	b	a	b	b	b	b
s_2	d	a	a	b	c	c	d
s_3	d	a	a	b	c	c	d
s_4	a	a	c	c	c	c	d
s_5	a	a	c	b	c	c	d
s_6	a	c	a	b	d	b	d
s_m	d	a	a	b	c	c	d

Lemma

$\mathsf{r_H}(s, S) \leq d_r \Rightarrow \mathsf{d_H}(s_m, s) \leq 2d_r$.

Problem: Outliers!

Extended Branching Algorithm

Majority string s_m:
pick a most frequent symbol in each column.

Example:
$t = 2$

s_1	d	b	a	b	b	b	b
s_2	d	a	a	b	c	c	d
s_3	d	a	a	b	c	c	d
s_4	a	a	c	c	c	c	d
s_5	a	a	c	b	c	c	d
s_6	a	c	a	b	d	b	d
s_m	d	a	a	b	c	c	d

Lemma

$\mathsf{r_H}(s, S) \leq d_r \Rightarrow \mathsf{d_H}(s_m, s) \leq 2d_r$.

Problem: Outliers!

Extended Branching Algorithm

Refined majority string $s_m^\diamond$:
$\exists$ symb. with at least as many occ. as majority symbol minus $t \Rightarrow$ use $\diamond$.

Example:
$t = 2$

s_1	d	b	a	b	b	b	b
s_2	d	a	a	b	c	c	d
s_3	d	a	a	b	c	c	d
s_4	a	a	c	c	c	c	d
s_5	a	a	c	b	c	c	d
s_6	a	c	a	b	d	b	d
$s_m^\diamond$	$\diamond$	a	$\diamond$	b	c	$\diamond$	d

Lemma

$\mathsf{r_H}(s, S) \leq d_r \Rightarrow \mathsf{d_H}(s_m, s) \leq 2d_r$.

Problem: Outliers!

Extended Branching Algorithm

Refined majority string $s_m^\diamond$:
$\exists$ symb. with at least as many occ. as majority symbol minus $t \Rightarrow$ use $\diamond$.

Example:
$t = 2$

s_1	d	b	a	b	b	b	b
s_2	d	a	a	b	c	c	d
s_3	d	a	a	b	c	c	d
s_4	a	a	c	c	c	c	d
s_5	a	a	c	b	c	c	d
s_6	a	c	a	b	d	b	d
$s_m^\diamond$	$\diamond$	a	$\diamond$	b	c	$\diamond$	d

Lemma

$\mathsf{r_H}(s, S) \leq d_r \Rightarrow \mathsf{d_H}(s_m, s) \leq 2d_r$.

Problem: Outliers!
Disputed columns.

Extended Branching Algorithm

Refined majority string $s_m^{\diamond}$:
$\exists$ symb. with at least as many occ. as majority symbol minus $t \Rightarrow$ use $\diamond$.

Example:
$t = 2$

s_1	d	b	a	b	b	b	b
s_2	d	a	a	b	c	c	d
s_3	d	a	a	b	c	c	d
s_4	a	a	c	c	c	c	d
s_5	a	a	c	b	c	c	d
s_6	a	c	a	b	d	b	d
$s_m^{\diamond}$	$\diamond$	a	$\diamond$	b	c	$\diamond$	d

Lemma

$\mathsf{r_H}(s, S) \leq d_r \Rightarrow \mathsf{d_H}(s_m, s) \leq 2d_r$.

Problem: Outliers!
Disputed columns.

Lemma

If the instance has a solution, then #disp. columns $\leq 4d_r$.

Extended Branching Algorithm

Refined majority string $s_m^\diamond$:
$\exists$ symb. with at least as many occ. as majority symbol minus $t \Rightarrow$ use $\diamond$.

Example:
$t = 2$

s_1	d	b	a	b	b	b	b
s_2	d	a	a	b	c	c	d
s_3	d	a	a	b	c	c	d
s_4	a	a	c	c	c	c	d
s_5	a	a	c	b	c	c	d
s_6	a	c	a	b	d	b	d
$s_m^\diamond$	$\diamond$	a	$\diamond$	b	c	$\diamond$	d

Lemma

$\mathsf{r_H}(s, S) \leq d_r \Rightarrow \mathsf{d_H}(s_m, s) \leq 2d_r$.

Problem: Outliers!
Disputed columns.

Lemma

If the instance has a solution, then #disp. columns $\leq 4d_r$.

Algorithm:
start with $s_m^\diamond$
branch over $d_r + 1$ mismatches
(or declare outlier)
depth bound: $6d_r + t$.

Results for (r, s)-Closest Substring

ℓ	k	m	d_r	d_s	$\|\Sigma\|$	Result
–	–	**p**	–	–	**p**	FPT
p	**p**	–	–	–	–	FPT
p	–	–	–	**p**	–	FPT
p	–	–	–	–	**p**	FPT
p	–	**p**	**p**	–	–	W[1]-hard
–	**p**	–	**p**	**p**	**p**	W[1]-hard
–	**p**	**p**	**p**	**p**	–	W[1]-hard

Results for (r, s)-CLOSEST SUBSTRING

ℓ	k	m	d_r	d_s	$\|\Sigma\|$	Result
–	–	**p**	–	–	**p**	FPT
p	**p**	–	–	–	–	FPT
p	–	–	–	**p**	–	FPT
p	–	–	–	–	**p**	FPT
p	–	**p**	**p**	–	–	W[1]-hard
–	**p**	–	**p**	**p**	**p**	W[1]-hard
–	**p**	**p**	**p**	**p**	–	W[1]-hard

(r,s)-Closest Substring(ℓ, m)

Theorem

(s)-Closest Substring(ℓ, m) is $\mathsf{W}[1]$-hard.

(r, s)-Closest Substring(ℓ, m)

Theorem

(s)-Closest Substring(ℓ, m) is W[1]-hard.

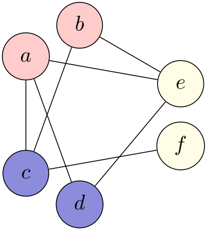
Reduction from multi-coloured clique:
Let $G = (V, E)$ be a graph with partition $V = V_1 \cup \ldots \cup V_{\mathrm{k_c}}$ such that every V_i, $1 \leq i \leq \mathrm{k_c}$, is an independent set.

(r, s)-Closest Substring(ℓ, m)

Theorem

(s)-Closest Substring(ℓ, m) is W[1]-hard.

Reduction from multi-coloured clique:
Let $G = (V, E)$ be a graph with partition $V = V_1 \cup \ldots \cup V_{k_c}$ such that every V_i, $1 \leq i \leq k_c$, is an independent set.



Repeat $N = 36$ times (rows $\mathcal{V}_1$, $\mathcal{V}_2$):

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

s : \$ a d e

Result for the Outlier Variants

Theorem

(s)-Closest String-wo($d_s, \ell, k - t$) is W[1]-hard.

Result for the Outlier Variants

Theorem

(s)-Closest String-wo($d_s, \ell, k - t$) is W[1]-hard.

We know (r)-Closest String-wo($t = 0, |\Sigma| = 2$) is NP-hard, but ...

Open Problem

(r)-Closest String-wo($|\Sigma|, k - t$),
(r)-Closest String-wo($|\Sigma|, d_r$),
(r)-Closest String-wo($|\Sigma|, d_r, k - t$).

Result for the Outlier Variants

(r, s)-Closest String-wo$(\ell, |\Sigma|) \in$ FPT (trivial).

Theorem

(r, s)-Closest String-wo$(\ell, |\Sigma|, d_r, d_s, (k-t))$ does not admit a polynomial kernel unless coNP $\subseteq$ NP/Poly.

Thank you very much for your attention.