Fragment assembly through minimal forbidden words

Gabriele Fici
Filippo Mignosi
Antonio Restivo
Marinella Sciortino

Dipartimento di Matematica ed Applicazioni - Università di Palermo
Motivations

We want to reconstruct a finite word starting from a set of its fragments.
Motivations

We want to reconstruct a finite word starting from a set of its fragments.

The reconstruction must be done in linear time on the size of the given set of fragments.
Motivations

We want to reconstruct a finite word starting from a set of its fragments.

The reconstruction must be done in linear time on the size of the given set of fragments.

This is a mathematical formalization of the problem of the DNA-reconstruction from a shotgun sequencing.
Minimal forbidden words

Given a word \( w \), a minimal forbidden word for \( w \) is a finite word \( v = a_1 a_2 \ldots a_n \) such that:

- \( v \) is not a factor of \( w \),
- the prefix \( a_1 a_2 \ldots a_{n-1} \) and the suffix \( a_2 a_3 \ldots a_n \) are factors of \( w \).
Given a word $w$, a minimal forbidden word for $w$ is a finite word $v = a_1a_2 \ldots a_n$ such that:

- $v$ is not a factor of $w$,
- the prefix $a_1a_2 \ldots a_{n-1}$ and the suffix $a_2a_3 \ldots a_n$ are factors of $w$.

The set of all the m.f.w. for $w$ is noted by $\mathcal{MF}(w)$. 

Given a word \( w \), a **minimal forbidden word** for \( w \) is a finite word \( v = a_1 a_2 \ldots a_n \) such that:

- \( v \) is not a factor of \( w \),
- the prefix \( a_1 a_2 \ldots a_{n-1} \) and the suffix \( a_2 a_3 \ldots a_n \) are factors of \( w \).

The set of all the m.f.w. for \( w \) is noted by \( \mathcal{MF}(w) \).

The length of the longest m.f.w. for \( w \) is noted by \( m(w) \).
Minimal forbidden words

Given a word $w$, a minimal forbidden word for $w$ is a finite word $v = a_1a_2\ldots a_n$ such that:

- $v$ is not a factor of $w$,
- the prefix $a_1a_2\ldots a_{n-1}$ and the suffix $a_2a_3\ldots a_n$ are factors of $w$.

The set of all the m.f.w. for $w$ is noted by $\mathcal{MF}(w)$. The length of the longest m.f.w. for $w$ is noted by $m(w)$.

Example. $w = abbab$

$\mathcal{MF}(w) = \{aba, bbb, aa, babb\}$, \hspace{1cm} m(w) = 4.
The Fragment Assembly problem

Given a set of fragments $I = \{i_1, \ldots, i_n\}$ (finite words over a finite alphabet $A$) find a $I$-compatible word, i.e. a finite word $w$ over $A$ such that:

- $I \subseteq \text{Fact}(w)$,

- Every "short" factor of $w$ (shorter than $m(w)$) is contained in $I$. 

Our first result is that the reconstruction cannot be ambiguous:

**Theorem.** Given a finite set of fragments $I$, there exists at most one $I$-compatible word.
Our first result is that the reconstruction cannot be ambiguous:

**Theorem.** Given a finite set of fragments $I$, there exists at most one $I$-compatible word.

The proof is combinatorial and based on the properties of the m.f.w. in relation with the DAGs.
In a previous paper we showed that one can reconstruct a finite word $w$ starting from the set of its minimal forbidden words $\mathcal{MF}(w)$ in linear time on the size of $\mathcal{MF}(w)$. 
Reconstruction

In a previous paper we showed that one can reconstruct a finite word $w$ starting from the set of its minimal forbidden words $\mathcal{MF}(w)$ in linear time on the size of $\mathcal{MF}(w)$.

So our goal is to retrieve the set $\mathcal{MF}(w)$ starting only from the set $I$. 
In a previous paper we showed that, for a set $I$ for which there exists a $I$-compatible word $w$, the knowledge of the value $m(w)$ allows to find in linear time the set $M\mathcal{F}(w)$ from the set $I$. So:

$$I + \text{existence of } w + m(w) \rightarrow w$$
Our results

Our main improvement consists in eliminating the knowledge of the value $m(w)$, preserving the linearity of the whole procedure.
Our results

Our main improvement consists in eliminating the knowledge of the value $m(w)$, preserving the linearity of the whole procedure.

Moreover, we give a procedure to decide whether, for an arbitrary set of finite words $I$, there exists a (unique) $I$-compatible word $w$. 
Our results

Our main improvement consists in eliminating the knowledge of the value $m(w)$, preserving the linearity of the whole procedure.

Moreover, we give a procedure to decide whether, for an arbitrary set of finite words $I$, there exists a (unique) $I$-compatible word $w$. So:

$$I \rightarrow \exists w \quad \text{or} \quad \# w$$
Remark

Given a set $I$ for which there exists a (unique) $I$-compatible word $w$, we proved that $w$ is solution of the Shortest Superstring Problem (that is known to be NP-hard) for the set $I$. 

Conclusion and open problems

Our algorithm solves in linear time an ideal formalization of the reconstruction of genomic sequences starting from a set of fragments.

We hope to generalize our results to a more realistic case, considering for example reading errors, orientation of the fragments, etc.