Design-Connectivity-Maher-2019

Maher Yann Alain

April 2019

1 Introduction

A faire par Maher

Definitions

Let $n \in \mathbb{N}, n \geq 2$.
Let Σ be an alphabet, $|\Sigma| \geq 2$. (We are especially interested in the case $\Sigma = \{A, U, G, C\}$).
Let $\mathcal{F}$ be the set of forbidden motifs.
Let $\mathcal{L}_{\mathcal{F},n}$ be the set of words in Σ^n that do not contain any motif in $\mathcal{F}$.
Let $\mathcal{L}_{\mathcal{F}}$ be the set of words in Σ^* that do not contain any motif in $\mathcal{F}$.
Let $H(w, w')$ be the Hamming distance between two words w, w' in Σ^n.

Let $m(\mathcal{F}) \stackrel{\text{def}}{=} max_{f in \mathcal{F}}|f|$.
We assume that $n \geq m(\mathcal{F})$. (Otherwise some forbidden motifs would never be problematic).
Hence any forbidden motif f in $\mathcal{F}$ of length $l < m(\mathcal{F})$ is equivalent to the set: $\bigcup_{i=0}^{n-l} \Sigma^i f \Sigma^{n-l-i}$ of forbidden motifs, all of legnth $m(F)$.
Thus we define $\widetilde{\mathcal{F}}$ the set of forbidden motifs - all of length $m(\mathcal{F})$ - equivalent to $\mathcal{F}$.

General problem

Input: $n \geq 2$, $\mathcal{F}$ a set of forbidden motifs, $\delta : \mathcal{L}_{\mathcal{F},n} \to \mathcal{L}_{\mathcal{F},n}$ a neighborhood function on $\mathcal{L}_{\mathcal{F},n}$
Question: The graph $G = (\mathcal{L}_{\mathcal{F},n}, \delta)$ is strongly connected.

2 Results

2.1 With the k-Hamming neighborhood

Definition 1. *Given $k \in \mathbb{N}^*$, we define δ_k the k-Hamming neighborhood as follows:*

$$\forall w \in \mathcal{L}_{\mathcal{F},n}, \delta_k(w) = \{w' \in \mathcal{L}_{\mathcal{F}} \mid H(w, w') \leq k\}.$$

With $k = n$, any $w \in \mathcal{L}_{\mathcal{F},n}$ can be changed into any other $w' \in \mathcal{L}_{\mathcal{F},n}$ in one step. Hence $G = (\mathcal{L}_{\mathcal{F},n}, \delta_n)$ is always strongly connected. Thus with the k-Hamming neighborhood a variant of the general problem can be considered:

General problem with the k-Hamming neighborhood

Input: $n \geq 2$, $\mathcal{F}$ a set of forbidden motifs
Question: the minimal $k \in \mathbb{N}^*$ such that the graph $G = (\mathcal{L}_{\mathcal{F},n}, \delta_k)$ is strongly connected.

2.1.1 One motif

Consider the case where $\mathcal{F}$ contains a single motif: $\mathcal{F} = \{f\}$.
Then $k = 1$ is sufficient to guarantee strong connectivity.

Result 2. *$\forall f \in \Sigma^+$, $G = (\mathcal{L}_{\{f\}}, \delta_1)$ is strongly connected.*

Proof. Let w and w' be two words in $\mathcal{L}_{\{f\}}$ (of length n).
As $f \neq \epsilon$, f can be written letter by letter as follows: $f = f_1...f_{|f|}$.
Since $|\Sigma| \geq 2$, let $a \in \Sigma$ such that $a \neq f_1$.
We show that there is a path from w to a^n.
To do so, from left to right we replace each letter in w by a (or keep it the same if already an a).
Formally, from $w = w_1...w_n$ we define the sequence $(u_i)_{0 \leq i \leq n}$ of intermediate words:

$$\forall 0 \leq i \leq n, u_i = a^i w_{i+1}...w_n.$$

Then:

- $\forall 0 \leq i \leq n-1, H(u_i, u_{i+1}) \leq 1$,
- for every i in $[1..n]$ we must prove that u_i is in $\mathcal{L}_{\{f\}}$. By contradiction, suppose that f appears in a u_i. Let j be the position in u_i of the leftmost letter of this occurrence of f.
  - if $j \leq i$: then the leftmost letter of f would be a, which is not by definition of a.
  - if $j > i$: then this occurrence of f would be a factor of w, which it cannot be since $w \in \mathcal{L}_{\{f\}}$.

  Contradiction. Hence every u_i is in $\mathcal{L}_{\{f\}}$.

This proves that there is a path from w to a^n with δ_1 as the neighborhood function.
The same can be done to obtain a path from w' to a^n.
Finally, since δ_1 is symmetric this gives a path from w to w' and vice-versa. □

In addition to show that $G = (\mathcal{L}_{\{f\}}, \delta_1)$ is strongly connected, this proof gives us $2n$ as an upper bound to the diameter of G.

We could have tried to prove Result 1 by induction on $H(w, w')$ instead. But it is unclear to what extent such an induction would be feasible (at least for now). Consider the following example with $\Sigma = \{A, U\}$:

$$\mathcal{F} = \{AUA\}, u = AAAA, v = AUUA.$$

There is no way to replace one non-extremal A of u with U without getting an occurrence of AUA. Hence there is no path from u to v in G with decreasing Hamming distance, even though u and v are connected according to Result 1. The same idea gives counter-examples of arbitrary Hamming distance:

$$\forall i \in \mathbb{N}^*, i \geq 2, \text{with: } \mathcal{F} = \{AUA\}, u_i = A^{i+2}, v_i = AU^iA,$$

$$\text{then: } H(u_i, v_i) = i.$$

These examples heavily rely on the fact that $|\Sigma| = 2$. There might be a way to get around this issue when $|\Sigma| \geq 3$ and find paths with non-increasing Hamming distance, but this would have to be looked at.

Yann: Another possible direction is to choose another candidate for an intermediate word, *i.e.* instead of showing connectivity as

$$w \leftrightarrow A^n \leftrightarrow w',$$

find $u(w, w')$, depending on w and w' such that

$$w \leftrightarrow u(w, w') \leftrightarrow w'$$

2.1.2 Two or more motifs

The idea from the proof of Result 1 could be used again to treat the cases when there is an available letter to do the same trick.

Result 3. *Let F be the set of forbidden motifs.*

- *If there exists $a \in \Sigma$ such that: $\forall f \in \mathcal{F}, f[1] \neq a,$, then $G = (\mathcal{L}_{\mathcal{F},n}, \delta_1)$ is strongly connected.*
- *Same result if there exists $a \in \Sigma$ such that: $\forall f \in \mathcal{F}, f[|f|] \neq a$.*

This tells us that we need at least $|\Sigma|$ forbidden motifs to obtain a disconnected graph with δ_1. Indeed if there are less than $|\Sigma|$ motifs, then we know that at least one letter is not the first letter of any forbidden motif.

Corollary 4. *If $|\mathcal{F}| < |\Sigma|$, then $G = (\mathcal{L}_{\mathcal{F},n}, \delta_1)$ is strongly connected.*

An example with $|\Sigma|$ words that gives a disconnected graph with δ_1 is the following:

$$with : \Sigma = \{a_1, a_2, ..., a_k\}, let : \mathcal{F} = \{a_1a_2, a_2a_1, a_3, ..., a_k\}.$$

Then the only two allowed words are a_1^n and a_2^n, and there is no way to go from one word to the other.

Case $k = n - 1$, $|\Sigma| = 2$

Result 5. *If $k = n - 1$ and $|\Sigma| = 2$, if u and v are disconnected in $G = (\mathcal{L}_{\mathcal{F},n}, \delta_{n-1})$, then:*

- *u is the opposite word of v in Σ^n,*
- *$\mathcal{L}_{\mathcal{F},n} = \{u, v\}$.*

Proof.

- With $|\Sigma| = 2$ the only word in Σ^n at Hamming distance greater than $n-1$ from u is its opposite word.
- By contradiction: if any other word w were in $\mathcal{L}_{\mathcal{F},n}$, then w would be a δ_{n-1}-neighbor of both u and v, and thus u and v would be connected.

□

2.2 Algorithmic aspects