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 $\widetilde{\widetilde{\mathcal{F}}}$ 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 $G_{\mathcal{F},n,k} \stackrel{\text{def}}{=} (\mathcal{L}_{\mathcal{F},n}, \delta_k)$ is strongly connected.

Remark 2. *Since δ_k is symmetric for any $1 \le k \le n$, $G_{\mathcal{F},n,k}$ is connected iff it is strongly connected. Thus we will use "connected" and "strongly connected" interchangeably when considering k-Hamming neighborhoods.*

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 3. *$\forall f \in \Sigma^+$, $G_{\{f\},n,1}$ 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 $|\Sigma| \ge 2$, 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 $(u_i)_{0 \le i \le n}$ of intermediate words:

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

Then:

- $\forall 0 \le i \le n-1, H(u_i, u_{i+1}) \le 1$,
- 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 $j \le i$: 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.

Remark 4. *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:

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

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

These examples heavily rely on the fact that $|\Sigma| = 2$. There might be a way to get around this issue when $|\Sigma| \ge 3$ 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

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

Idea. *Give an arbitrary order on the letters in Σ and take as the representative of each connected component their smallest element w.r.t the lexical order?*

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 5. *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 6. *If $|\mathcal{F}| < |\Sigma|$, then $G_{\mathcal{F},n,1}$ 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 7. *If $k = n-1$ and $|\Sigma| = 2$, then:*
u and v are disconnected in $G = (\mathcal{L}_{\mathcal{F},n}, \delta_{n-1})$ iff:

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

Proof. ($\Leftarrow$) As u and v are opposite, $H(u,v) = n$. Hence u and v are not neighbors and they are the only elements in G.

($\Rightarrow$)

- 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.

$\square$

Then the only possible disconnected graph here is with two nodes that are opposite sequences, which is very restrictive. This makes for a simple example to study the impact of $\mathcal{F}$ on the strong connectivity of $G_{\mathcal{F},n,n-1}$.

Proposition 8. *If P_1 and P_2 are in $\mathcal{L}_{\mathcal{F},|P_1|}$ and disconnected in $G_{\mathcal{F},|P_1|,k}$ (for some $k \in [1..|P_1|]$), then: $\forall (S_1, S_2)$ couple of words of the same length: (P_1S_1 and P_2S_2 are in $\mathcal{L}_{\mathcal{F},|P_1|}$) $\Rightarrow$ (P_1S_1 and P_2S_2 are disconnected in $G_{\mathcal{F},|P_1S_1|,k}$).*

In other words, if we find two words P_1, P_2 that are disconnected in $G_{\mathcal{F},i,k}$ for some $i \in \mathbb{N}^*$, then any couple of words P_1S_1, P_2S_2 in $G_{\mathcal{F},j,k}$ (for some $j > i$) that have them as their prefixes are disconnected as well.

Proof. Let P_1 and P_2 be two words in $\mathcal{L}_{\mathcal{F},|P_1|}$ that are disconnected in $G_{\mathcal{F},|P_1|,k}$.
By contradiction: if there exist S_1, S_2 such that P_1S_1 and P_2S_2 are connected in $G_{\mathcal{F},|P_1S_1|,k}$,
then there is a path $(P_1S_1 = u_0) \to u_1 \to [...] \to u_i \to (u_{i+1} = P_2S_2)$ in $G_{\mathcal{F},|P_1S_1|,k}$.
We know then that: $\forall 0 \leq j \leq i, H(u_j, u_{j+1}) \leq k$.
For $0 \leq j \leq i+1$ let P'_j be the prefix of length $|P_1|$ in u_j.
Since we take the prefixes, we still have: $\forall 0 \leq j \leq i, H(P'_j, P'_{j+1}) \leq k$.
Hence $(P_1 = P'_0) \to P'_1 \to [...] \to P'_i \to (P'_{i+1} = P_2)$ is a valid path in $G_{\mathcal{F},|P_1|,k}$.
Hence P_1 and P_2 would be connected in $G_{\mathcal{F},|P_1|,k}$.
Contradiction. □

Idea. *Use a De Bruijn graph to know when a word is a prefix of arbitrary long allowed words. See Remark* 11. *.*

2.2 De Bruijn graphs

2.2.1 Properties

Definition 9. *Given $\mathcal{F}$, we define $\mathcal{DB}_{\mathcal{F}}$ the De Bruijn graph of $\mathcal{F}$ the following way:*

- *Vertices: $\complement\widetilde{\mathcal{F}}$ the allowed substrings of length $m(\mathcal{F})$.*
- *Edges: if $u \in \Sigma^{m(\mathcal{F})-1}$, if $a, b \in \Sigma$, then: there is an edge from au to ub iff au and ub are both in $\complement\widetilde{\mathcal{F}}$.*

Proposition 10. *Let $w = w_1...w_n$ be a word of length n. Then:*

$$w \in \mathcal{L}_{\mathcal{F},n} \text{ iff } w_1...w_{m(F)} \to w_2...w_{m(F)+1} \to [...] \to w_{n-m(F)+1}...w_n \text{ is a valid path in } \mathcal{DB}_{\mathcal{F}}.$$

Proof. w is an allowed word iff every factor of length $m(\mathcal{F})$ in w is allowed. □

Remark 11. *There are arbitrarily long allowed words iff there is a cycle in $\mathcal{DB}_{\mathcal{F}}$.*

Result 12. *Let $u_1 \to [...] \to u_i$ be a path in $\mathcal{DB}_{\mathcal{F}}$ ($i \geq 3$).*
If u_i is a neighbor of u_1, then $(u_2, [...], u_{i-1})$ is a cycle in $\mathcal{DB}_{\mathcal{F}}$.

In other words: if there is a shortcut to a path in $\mathcal{DB}_{\mathcal{F}}$, then the intermediate elements form a cycle.

Proof. Since u_i is a a neighbor of u_1 in $\mathcal{DB}_{\mathcal{F}}$, we can write: $u_1 = av$ and $u_i = vb$ for some $v \in \Sigma^{m(\mathcal{F})-1}$ and $a, b \in \Sigma$.
But then we know as well that: $u_2 = vc$ and $u_{i-1} = dv$ for some $a, b \in \Sigma$.
Hence u_2 is a neighbor of u_{i-1} and $(u_2, [...], u_{i-1})$ forms a cycle in $\mathcal{DB}_{\mathcal{F}}$. □

Idea. *How to interpret Hamming edits with paths in $\mathcal{DB}_{\mathcal{F}}$?*

Definition 13. *We define $\mathcal{DB}_{\mathcal{F},n}$ the graph obtained by removing all the connected components in $\mathcal{DB}_{\mathcal{F}}$ that do not encode any word of length n.*

The goal of this notion is to only keep the meaningful part in $\mathcal{DB}_{\mathcal{F}}$ that generates the allowed words of length n. This is the same as removing all the connected components that have no path of length $\geq n - m(\mathcal{F})$.

Remark 14. *For any $n \geq m(\mathcal{F})$, $\mathcal{DB}_{\mathcal{F}}$ and $\mathcal{DB}_{\mathcal{F},n}$ have exactly the same paths of length $n - m(\mathcal{F})$.*

Lemma 15. *If we follow the same sequence of letters $a_1, a_2, [...], a_j$ from two distinct sequences u, v in $\mathcal{DB}_{\mathcal{F}}$, if $j \geq m(\mathcal{F})$, then the two subsequent paths have merged at some index $i \leq m(\mathcal{F})$.*

Proof. After $m(\mathcal{F})$ steps the resulting word is $a_1...a_{m(\mathcal{F})}$ in both paths, so the paths merged either at index $m(\mathcal{F})$ or at a smaller index. □

2.2.2 Applications

<u>**Back to the case $k = n - 1$, $|\Sigma| = 2$**</u>

Result 16. *With $|\Sigma| = 2$ (for instance $\Sigma = \{A, C\}$): $G_{\mathcal{F},n,n-1}$ is disconnected iff $\mathcal{DB}_{\mathcal{F},n}$ is either:*

(i) A...A$\circlearrowleft$, C...C$\circlearrowleft$

(ii) ACA... $\leftrightarrow$ CAC...

(iii) two "opposite" paths of length $n - m(\mathcal{F})$ with no connection: $u_1 \rightarrow [...] \rightarrow u_{n-m(\mathcal{F})+1}$, $\overline{u_1} \rightarrow [...] \rightarrow \overline{u_{n-m(\mathcal{F})+1}}$, where $\overline{u_i}$ is the opposite word of u_i.

Proof. ($\Leftarrow$) All the three cases for $\mathcal{DB}_{\mathcal{F},n}$ imply that there are exactly two paths of length $n - m(\mathcal{F})$ in $\mathcal{DB}_{\mathcal{F}}$. By *Proposition* 10., we deduce that there are only two words in $\mathcal{L}_{\mathcal{F},n}$ and they are opposite, which by *Result* 7. means that $G_{\mathcal{F},n,n-1}$ is disconnected.

($\Rightarrow$) We know from *Result* 7. that the only way to have $G_{\mathcal{F},n,n-1}$ disconnected is to have exactly two vertices and that they are opposite to each other. Using *Proposition* 9., this means that we must exactly have two paths of length $n - m(\mathcal{F})$ in $\mathcal{DB}_{\mathcal{F}}$.
Now we show that $\mathcal{DB}_{\mathcal{F},n}$ can only be of one the three proposed forms.

- If there is a 1-cycle in $\mathcal{DB}_{\mathcal{F},n}$:
  then this 1-cycle is either A...A$\circlearrowleft$ or C...C$\circlearrowleft$. In any case the other one must be included as well in order to include the path for the opposite word. There are already two paths of length $n - m(\mathcal{F})$ in the graph [A...A$\circlearrowleft$, C...C$\circlearrowleft$] and appending any element to these components would add another path of length $n - m(\mathcal{F})$, which we do not want. Thus the only graph $\mathcal{DB}_{\mathcal{F},n}$ that can have a 1-cycle is [A...A$\circlearrowleft$, C...C$\circlearrowleft$], which is case (i).
- If there is a 2-cycle in $\mathcal{DB}_{\mathcal{F},n}$:
  then this 2-cycle can only be ACA... $\leftrightarrow$ CAC..., which already encodes two words of length n. Again, appending any element to this component would add another allowed word of length n, so the only graph $\mathcal{DB}_{\mathcal{F},n}$ that can have a 2-cycle is [ACA... $\leftrightarrow$ CAC...], which is case (ii).
- If there is a (3+)-cycle in $\mathcal{DB}_{\mathcal{F},n}$:
  then there would be at least three allowed words of length n (we obtain them by starting from a different element of the cycle as the prefix and by going through the cycle). Since we only want two allowed words of length n, $\mathcal{DB}_{\mathcal{F},n}$ cannot have a (3+)-cycle.
- If there is no cycle in $\mathcal{DB}_{\mathcal{F},n}$:
  TODO

□

2.3 Algorithmic aspects