Journal of Mathematical Biology manuscript No.
(will be inserted by the editor)

James W. Anderson, Keith R. Fox and Graham A. Niblo

A fast algorithm for the construction of universal footprinting templates in DNA

Received: date / Revised version: date –

Abstract. We introduce and give a complete description of a new graph to be used for DNA sequencing questions. This graph has the advantage over the classical de Bruijn graph that it fully accounts for the double stranded nature of DNA, rather than dealing with single strands. Technically, our graph may be thought of as the quotient of the de Bruijn graph under the natural involution of sending a DNA strand to its complementary strand. However, this involution has fixed points, and this complicates the structure of the quotient graph which we have therefore modified herein.

As an application and motivating example, we give an efficient algorithm for constructing universal footprinting templates for n-mers. This problem may be formulated as the task of finding a shortest possible segment of DNA which contains every possible sequence of base pairs of some fixed length n. Previous work by Kwan et al has attacked this problem from a numerical point of view and generated minimal length universal footprinting templates for $n = 2$, 3, 5, 7, together with unsubstantiated candidates for the case $n = 4$. We show that their candidates for $n = 4$ are indeed minimal length universal footprinting templates.

1. Introduction and statements of results

Many compounds, ranging from small molecules to proteins, bind to double stranded DNA in a sequence specific fashion, recognizing unique combinations of DNA base pairs, which they access from either the major or minor groove. Several techniques are widely used for studying the strength and specificity of these interactions, including bandshift assays (EMSA) and footprinting [2], [3], [4]. However these methods generally use DNA substrates of 200 base pairs or shorter and therefore require that the preferred binding sites are already known, so as to ensure that they are present within

First and Third Authors: School of Mathematics, University of Southampton, Southampton SO17 1BJ, ENGLAND

Second Author: School of Biological Sciences, University of Southampton, Southampton SO16 7PX, ENGLAND

Send offprint requests to: James W. Anderson

Key words: DNA sequencing – universal footprinting template – de Bruijn graph – Eulerian graphs

the fragments. This is not a problem for agents that only recognise one or two base pairs but, as the selectivity of a compound increases, the chance of finding the best target sites in a given DNA fragment becomes more remote. For a target site of n base pairs there are $\frac{1}{2}4^n$ combinations if n is odd and $\frac{1}{2}(4^n + 4^{n/2})$ combinations if n is even. As a result, there are 10 different dinucleotides, 32 trinucleotides, 136 tetranucleotides, 512 pentanucleotides and 2080 hexanucleotides. Even if the preferred binding site for a ligand is present within a given fragment, a proper analysis of its specificity should examine the binding to related sequences, which differ by one or two bases, and these may not be present. The problem is such that many potential drug or protein binding sites may therefore be overlooked or ignored. We are therefore interested in designing DNA fragments that contain as many different nucleotide sequences of length n as possible, and ideally wish to have an effective algorithm to determine shortest possible fragments that can contain all possible nucleotide sequences of length n. The purpose of this note is to define and discuss the basic properties of a new graph that is of interest to researchers working on DNA, for attacking this question and others. As an application, we use this graph to formulate an effective algorithm for producing universal footprinting templates (UFTs) for n-mers.

There have been two previous empirical attempts to design such DNA segments. The first approach [6] manually produced a DNA segment of length 166 containing all 136 tetranucleotide sequences, while at the same time minimising the occurrence of oligopurine/oligopyrimidine tracts and long blocks of AT or GC base pairs. The second approach [5] used random computer searches to generate DNA segments containing all possible 2-, 3-, 4-, 5-, and 7-mer targets. Minimal length segments containing odd numbers of base pairs were easily found, and were produced within 0.01 (for $n = 3$), 0.3 (for $n = 5$) and 104 (for $n = 7$) min of CPU time, respectively. In contrast, for $n = 4$, no 139 base pair solution was found in more than 150 h of CPU time, though segments that were a few bases longer (144 base pairs) were easily generated. Our approach allows us to explain this dichotomy between the cases of n even and n odd below. We note that our graph theoretic approach combines the strengths of the two previous attacks on this problem. It yields a fast algorithm which can generate many different minimal universal footprinting templates, while at the same time allowing some control over the internal structure of the UFTs.

We do not need the details of the chemical structure of DNA in order to describe our algorithm. For our purposes, it suffices to view a segment of DNA as two intertwined complementary strings of the nucleotides adenine (A), cytosine (C), guanine (G), and thymine (T), so that a segment of DNA is described as a pair of strings in the alphabet $\mathcal{A} = \{A, T, C, G\}$. The two strings are related by a complementing operation which pairs A with T and which pairs C with G. Since the natural reading directions along the two strings making up a segment of DNA are opposite to one another, the complementing operation also reverses the order of letters in a string.

Formally, given a letter X in the alphabet $\mathcal{A}$, we denote its complement by X^c, so that $A^c = T$, $T^c = A$, $C^c = G$ and $G^c = C$. This complementing operation extends to strings in the alphabet $\mathcal{A}$ in the following way: given a string $W = X_1X_2 \dots X_n$, where each X_j is a letter in the alphabet $\mathcal{A}$, its complementary string is $W^c = X_n^c \dots X_2^c X_1^c$. In order to simplify notation, if a string W consists of a single letter X repeated n times, we sometimes write $W = X^n$.

Unless otherwise stated, we write strings so that the $5'$ end is to the left and the $3'$ end is to the right, so that the reading direction is left to right.

Given a string W of length n, we call the pair $\{W, W^c\}$ of the string and its complement an *n-mer*. There is a one-to-one correspondence between n-mers and distinct DNA segments consisting of n base pairs. We note that generally the pair of strings in an n-mer is unordered, so that $\{W, W^c\}$ and $\{W^c, W\}$ are the same n-mer. If a string W of length n is *self-complementary* (so that $W^c = W$), then its corresponding n-mer $\{W, W^c\}$ contains only a single string of length n, which we sometimes write $\{W\}$. (In [5], these strings are referred to as *palindromic*.)

As an example, consider the string

$$S = AAATCCGTGCCCTATGGTAACAGAGTCGCTTCAA.$$

Its complement is the string

$$S^c = TTGAAGCGACTCTGTTACCATAGGGCACGGATTT.$$

Given this pairing through the complementing operation, the entire DNA segment is determined by either of the two strings comprising it. Hence, the data we need to specify a DNA segment is a string of letters in the alphabet $\mathcal{A}$. It is through these strings of letters that we describe the algorithm. The *length* of a string in the alphabet $\mathcal{A}$ is the number of letters it contains.

As described in the introduction, the task of interest is to construct DNA segments which allow experimentalists to simultaneously examine the behaviour of multiple DNA segments of a given length n. Such a DNA segment should satisfy the following two conditions: Every possible DNA segment of length n appears at least once in the segment; and the DNA segment is as short as possible among all DNA segments containing all DNA segments of length n. Any DNA segment containing all DNA segments of length n is known as a *universal footprinting template (UFT) for n-mers*. (In [5], these strings are referred to as *n-complete DNA sequences*.) If the DNA segment is shortest possible among all universal footprinting templates for n-mers, we shall call it a *minimal universal footprinting template for n-mers*.

As the essential data needed to describe a DNA segment is a string in the alphabet $\mathcal{A}$, we reformulate this condition in terms of such strings. A string in the alphabet $\mathcal{A}$ is a *universal footprinting template for n-mers* if it contains at least one string from each n-mer. A string in the alphabet $\mathcal{A}$ is a *minimal universal footprinting template for n-mers* if it contains at least one string from each n-mer, and has minimal length among all such strings.

Consider the case $n = 1$. There are two 1-mers, namely $\{A, T\}$ and $\{C, G\}$. There are two DNA segments of length 1: namely, $A \leftrightarrow T$ and $C \leftrightarrow G$, where $\leftrightarrow$ indicates the bond between the two complementary nucleotides making up a base pair. Hence, the string $ACGT$ corresponds to a universal footprinting template for 1-mers: it corresponds to the DNA segment

$$\begin{array}{c} 5' - A\ C\ G\ T - 3' \\ \quad\ \updownarrow\ \updownarrow\ \updownarrow\ \updownarrow \\ 3' - T\ G\ C\ A - 5' \end{array}$$

However, this DNA segment contains each of the two 1-mers twice, and so is not shortest possible. A shortest possible such DNA segment, and hence a minimal universal footprinting template for 1-mers, is

$$\begin{array}{c} 5' - A\ C - 3' \\ \quad\ \updownarrow\ \updownarrow \\ 3' - T\ G - 5' \end{array}$$

This DNA segment corresponds to the string AC (or its complement, the string GT). Hence, the string AC is a minimal universal footprinting template for 1-mers. The other minimal UFTs for 1-mers are the strings GA or its complement TC , CA or TG, and AG or CT.

Fix $n \geq 1$. A string S of length $k \geq n$ contains $k - (n - 1)$ substrings of length n, where the substrings of length n are formed by taking blocks of n contiguous letters in S. It follows that if we let K_n denote the number of distinct n-mers, and if we assume that the string W contains at least one string from each n-mer, then the length, length(W), of W satisfies length$(W) \geq K_n + (n - 1)$. Hence a minimal UFT for n-mers must have length at least $K_n + n - 1$.

Consider the case $n = 2$. There are $K_2 = 10$ distinct 2-mers: $\{AA, TT\}$, $\{AC, GT\}$, $\{CC, GG\}$, $\{CG\}$, $\{GC\}$, $\{CA, TG\}$, $\{GA, TC\}$, $\{AG, CT\}$, $\{AT\}$, and $\{TA\}$. The string $W = AACCGCAGATA$ contains exactly one string from each of the ten possible 2-mers, and hence gives rise to a universal footprinting template for 2-mers. To see that W is a minimal universal footprinting template for 2-mers, we note that length$(W) = 11 = K_2 + 1 = K_2 + (2 - 1)$, and so W has the minimal possible length among all strings containing each string of length 2 or its complement.

In order to determine the length of a minimal universal footprinting template for n-mers, we compute the number K_n of distinct n-mers for each n. (Compare [5].) There are 4^n distinct strings of length n in the alphabet $\mathcal{A} = \{A, C, G, T\}$. When n is odd, each string W of length n is distinct from its complement, since W^c differs from W in their middle letters as no letter is its own complement. Since no string of odd length is self-complementary, each n-mer contains two strings, and so the number of n-mers is equal to

half the number of strings of length n. Hence, there are $K_n = \frac{1}{2}4^n$ n-mers when n is odd.

When n is even, a string can equal its complement, and so some n-mers contain only a single string. (For example, $(AT)^c = T^c A^c = AT$.) A self-complementary string must take the form $W = UU^c$ where U is a string of length $\frac{1}{2}n$, and so the number of self-complementary strings is equal to the number of strings of length $\frac{1}{2}n$, which is $4^{n/2}$. These yield $4^{n/2}$ n-mers each consisting of a single self-complementary string. The remaining $4^n - 4^{n/2}$ strings come in complementary pairs, and so yield $\frac{1}{2}(4^n - 4^{n/2})$ distinct n-mers. Hence there are $K_n = \frac{1}{2}(4^n + 4^{n/2})$ n-mers when n is even.

Hence the naive estimate, naive(n), for the length of a minimal universal footprinting template for n-mers is

$$\text{naive}(n) = K_n + n - 1 = \begin{cases} \frac{1}{2}4^n + n - 1 & \text{for } n \text{ odd,} \\ \frac{1}{2}(4^n + 4^{n/2}) + n - 1 & \text{for } n \text{ even.} \end{cases}$$

If there exists a string containing one and only one string from each n-mer, then this string will have length equal to naive(n). Note that this estimate holds for the two cases considered so far, namely that $n = 1$ with naive(1) = 2, and $n = 2$ with naive(2) = 11. In both cases, there is a universal footprinting template for n-mers of length naive(n), and so such a string must be a minimal universal footprinting template for n-mers.

Consider the case $n = 3$. We calculate that there are thirty-two 3-mers, and so the shortest possible string containing a string from every 3-mer must have length at least naive(3) = $\frac{1}{2}4^3 + 3 - 1 = 34$. The following string of length 34 contains one and only one string from every 3-mer, and therefore represents a minimal universal footprinting template for 3-mers:

$$AAATCCGTGCCCTATGGTAACAGAGTCGCTTCAA$$

Our first main result is that this behavior persists for all odd n.

Theorem 1. *For n odd (and $n \geq 3$), there exists a minimal universal footprinting template for n-mers of length* naive(n) = $\frac{1}{2}4^n + n - 1$.

We will give the proof of Theorem 1 in Section 5.

When n is even the problem is considerably less tractable. As evidence for this claim we point to the experimental results of [5], as described in the introduction. That experimental approach found no UFT for 4-mers of length less than 144 depite the naive lower bound being 139. Using our technique we will show that in fact a minimal universal footprinting template for 4-mers must have length 144. Our algorithm is sufficient to construct such a string; indeed, the following example of a minimal universal footprinting template for 4-mers was constructed by hand using the algorithm:

$$ACGTACCGACCTCACGCACTCCCGCGCCTGAAGTAA$$
$$GGAATGACGGCCGGACTAACTGCCCTATCAAGCGAG$$
$$AGCTAGCAAATATAAAGAACCAACGAAAACAATTAA$$
$$TCCACACCCCAGACAGCCATGCATACATCGATCTAC$$

The analysis in the case of n even is complicated by the existence of self-complementary n-mers. While it should be possible to give an explicit formula for the length of a minimal universal footprinting template for n-mers for n even using the algorithm described below, it is more efficient to construct them for each n by a combination of our Eulerian path methods and heuristic search. The complexity of the search is bounded by the following result.

Theorem 2. *For n even (and $n \geq 4$), the length of a minimal universal footprinting template for n-mers is at least* $\min(n) = \frac{1}{2}4^n + 4^{n/2} + n - 4$ *and at most* $\max(n) = \frac{1}{2}4^n + n\,4^{n/2-1}$.

We complete the proof of Theorem 2 in Section 6.

The *gap*, which we define to be the difference between the upper and lower bounds in Theorem 2, is $(n-4)(2^{n-2}-1)$ so for $n = 2$ or $n = 4$ the gap is 0 and the bounds are tight. In particular, any universal footprinting template for 4-mers, such as the one constructed above (or the examples constructed in [5]) of length 144 is then a minimal universal footprinting template for 4-mers. We note in passing that for $n = 6$ the gap is 30, while for $n = 8$ the gap is only 252, which is small compared with a lower length of 32903 for a minimal UFT for 8-mers.

We prove Theorem 1 and Theorem 2 using the same basic technique, by converting the problem of finding a universal footprinting template for n-mers or a minimal universal footprinting template for n-mers into the question of the existence of Eulerian paths in a certain graph. Of primary importance is the observation that the graphical solution to this problem yields a fast algorithm for constructing many distinct minimal universal footprinting template for n-mers.

The paper is organised as follows. In Section 2, we present the basic graph theoretic terminology and the machinery we employ.

In Section 3, we define our basic object of study, the graph $\mathcal{G}_n$ whose edges are n-mers, whose vertices are $(n-1)$-mers, and in which universal footprinting templates for n-mers correspond to paths that cover all edges in the graph. We will show how to interpret DNA segments as paths in the graph $\mathcal{G}_n$ satisfying an admissability criterion, so that minimal universal footprinting templates for n-mers correspond to admissable paths covering all edges of $\mathcal{G}_n$ having minimal length. We close Section 3 with a discussion of the connection between $\mathcal{G}_n$ and the classical de Bruijn graph.

In Section 4, we discuss the structure of $\mathcal{G}_n$ at its vertices. We state and prove Proposition 1, characterizing standard loops and lollipop loops in $\mathcal{G}_n$. We state Proposition 2, discussing in detail the structure of $\mathcal{G}_n$ at its vertices for odd n, and give a detailed illustration of the possibilities arising for the case $n = 3$; the proof of Proposition 2 is given in Section 7. We state Proposition 3, discussing in detail the structure of $\mathcal{G}_n$ at its vertices for odd n, and give a detailed illustration of the possibilities arising for the case $n = 4$; the proof of Proposition 3 is given in Section 8.

We present the algorithm for constructing universal foootprinting templates for n-mers for the cases of n odd and n even separately. We begin with the case n odd, which contains fewer technical complications. In Section 5, we present an algorithm for constructing minimal universal footprinting template for n-mers for n odd, including a detailed discussion of the implementation of the algorithm in the case $n = 3$. We also give the proof of Theorem 1.

In Section 6, we present the extra step needed to implement the algorithm for the construction of universal footprinting templates for n-mers in the case of n even, including a detailed discussion of the implementation of this additional step in the case $n = 4$. We also give the proof of Theorem 2.

2. Graph theoretic preliminaries

In this Section, we describe the basic machinery from graph theory we will employ. We use [1] as our basic reference for graph theory facts and results.

A *graph* $\mathcal{G}$ is a mathematical abstraction, a figure containing a set of *vertices* and a set of *edges*. The vertices can be anything. The edges join the vertices. An edge can be *unordered*, so that there is no preferred direction of travel along the edge, or *ordered*, so that there is a preferred direction of travel along the edge. For the time being, we assume that the edges are unordered. We allow the possibility that an edge joins a vertex to itself (which we refer to as a *loop*), and we also allow the possibility that there is more than one edge joining a given pair of vertices.

Given a vertex v of a graph $\mathcal{G}$, define the *valency* of v to be the number of edges in $\mathcal{G}$ that have v as an endpoint (and so are *incident* to v). We adopt the convention that if e is a loop based at v (so that both endpoints of e are at v), then e contributes 2 to the valency of v.

A *path* in a graph $\mathcal{G}$ is an ordered sequence $\mathcal{P} = \{e_1, \ldots, e_n\}$ of edges of $\mathcal{G}$, so that each consecutive pair e_j and e_{j+1} of edges appearing in $\mathcal{P}$ are incident to a common vertex. Note that a given edge may appear several times in a path. Working within the restrictions of our motivating problem, we assume that all graphs appearing in this paper are *connected*, so that given any pair of distinct vertices of $\mathcal{G}$, there exists a path in $\mathcal{G}$ beginning at one and ending at the other. A path in $\mathcal{G}$ is a *circular path* if it begins and ends at the same vertex.

A path $\mathcal{P}$ in a graph $\mathcal{G}$ is *Eulerian* if $\mathcal{P}$ contains each edge of $\mathcal{G}$ exactly once. A graph $\mathcal{G}$ admits an Eulerian path precisely when it has either zero or two vertices of odd valency. If $\mathcal{G}$ has zero vertices of odd valency, then the Eulerian path can begin at any vertex of $\mathcal{G}$ and will necessarily be a circular path. If $\mathcal{G}$ has two vertices of odd valency, then the Eulerian path must begin at one of them and end at the other.

There is a standard (fast) algorithm for constructing an Eulerian path $\mathcal{P}$ in a graph $\mathcal{G}$ (necessarily containing either zero or two vertices of odd valency). We note that the complexity of this algorithm is linear in the number of edges of the graph. We will first discuss the algorithm when G is assumed to have the property that every vertex has even valency.

We first choose a pairing between the edges incident to each vertex. By this, we mean that at each vertex v of $\mathcal{G}$, we divide the edges of $\mathcal{G}$ incident to v into two equal sets, and we associate each edge in one set to one and only one edge in the other set.

Starting at any vertex v_0, choose any edge e_1 incident to v_0, and mark it as *used*. The edge e_1 is incident to the vertex v_0 and to another vertex v_1 (we allow the possibility that $v_0 = v_1$, in the case that e_1 is a loop). At v_1, the edge e_1 is paired with an edge e_2, which is incident v_1 and to another vertex v_2. If e_2 has not yet been used, we follow it to the vertex at its other endpoint, where it is paired with an edge e_3. We continue in this way for as long as possible, marking each edge as used as we cross it, until we come to an edge e_n whose pair has already been used. Because of the construction, this final edge e_n must be incident to the initial vertex v_0, where it must be paired with the intial edge e_1. (All the other edges that we have marked as used have been paired.) So, we have constructed a circular path $\mathcal{P}_1$ in $\mathcal{G}$ which crosses each edge at most once, but may not have crossed every edge.

If all of the edges of $\mathcal{G}$ are marked as used at this stage, then every edge in $\mathcal{G}$ appears once and only once in $\mathcal{P}_1$, and so $\mathcal{P}_1$ is the desired Eulerian path.

If there are unused edges in $\mathcal{G}$, choose a vertex v_0' that is incident to an unused edge, choose an unused edge e_1 incident to v_0', and repeat the construction above, crossing only unused edges to construct a second circular path.

We continue in this manner until all edges of $\mathcal{G}$ are marked as used. We are left with a collection $\mathcal{P}_1, \ldots, \mathcal{P}_k$ of circular paths in $\mathcal{G}$ which together cross each edge of the graph exactly once and each circular path may be traversed in either direction, starting at any vertex on it. We now *plumb* these circular paths together to get an Eulerian path in $\mathcal{G}$.

This plumbing operation can be described as follows. Since $\mathcal{G}$ is connected, two of the circular paths, say $\mathcal{P}_1$ and $\mathcal{P}_2$, contain edges incident to a common vertex v. Rewrite $\mathcal{P}_1$ and $\mathcal{P}_2$ so that both begin and end at the vertex v, by cyclically reordering the edges of $\mathcal{G}$ that occur in each; we again call these rewritten circular paths $\mathcal{P}_1$ and $\mathcal{P}_2$. We then construct a new path by first following $\mathcal{P}_1$, and then following $\mathcal{P}_2$. Each pair of circular paths that pass through a common vertex of $\mathcal{G}$ can be plumbed together in this way. Since $\mathcal{G}$ is connected, we are able to plumb together all of the circular paths $\mathcal{P}_1, \ldots, \mathcal{P}_k$ to obtain a single circular path $\mathcal{P}$. Since each edge of $\mathcal{G}$ appears exactly once in one and only one of the $\mathcal{P}_k$, the path $\mathcal{P}$ is an Eulerian path.

Now we consider the case where $\mathcal{G}$ has exactly two vertices u, v of odd valency. Since $\mathcal{G}$ is connected, we can choose a path $\mathcal{P}_1$ joining u to v which crosses each edge at most once. We remove these edges from the graph $\mathcal{G}$ to obtain a new graph $\mathcal{G}'$ which contains no vertices of odd valency. This new graph $\mathcal{G}'$ is not necessarily connected, but each connected component admits a circular Eulerian path that can be constructed as described above. Label these Eulerian paths as $\mathcal{P}_2, \ldots, \mathcal{P}_k$. We now plumb them onto the path $\mathcal{P}_1$ as before to produce an Eulerian path in $\mathcal{G}$ from u to v.

We close this Section by noting that, even if we are unable to construct an Eulerian path in a graph $\mathcal{G}$ because it does not satisfy the appropriate condition on the valencies of its vertices, we can still implement a modification of this algorithm to construct a path that crosses every edge of $\mathcal{G}$ at least once, at least in the case that there are an even number of vertices of odd valency. Begin by pairing the vertices of odd valency, and connect the vertices in each pair with a path in $\mathcal{G}$. Traverse each of these paths twice, from one end to the other and then back again, to obtain a collection of circular paths. Removing each of these short circular paths from $\mathcal{G}$, we obtain a graph $\mathcal{G}'$ in which each vertex has even valency, and therefore admits the required edge pairing. We then use the algorithm described above to produce a collection of Eulerian paths in the graph $\mathcal{G}'$, one for each connected component in $\mathcal{G}'$. Plumbing these Eulerian paths with the short circular paths joining pairs of vertices of odd valence yields a circular path in the original graph $\mathcal{G}$ which crosses each edge at least once and that crosses some edges twice. Every circular path in $\mathcal{G}$ which crosses each edge at least once can be described in this way; hence, to construct a shortest path in $\mathcal{G}$ which crosses every edge at least once, we only need to find a collection of paths joining the odd valency vertices, in pairs, of shortest total length.

3. The problem posed in graphical form

In this Section, we formulate the question of finding a (minimal) universal footprinting template for n-mers as a graph theoretic question. Fix an integer $n \geq 2$, and define the graph $\mathcal{G}_n$ as follows.

The vertices of $\mathcal{G}_n$ are the $(n-1)$-mers, so each vertex consists of a complementary pair $\{V, V^c\}$ of strings of length $n-1$ in the alphabet $\mathcal{A} = \{A, C, G, T\}$. The edges of $\mathcal{G}_n$ are the n-mers, so each edge consists of a complementary pair $\{W, W^c\}$ of strings of length n in the alphabet $\mathcal{A}$. Depending on the parity of n, the graph $\mathcal{G}_n$ will have either self-complementary vertex $(n-1)$-mers or self-complementary edge n-mers.

The incidence relation for vertices and edges is straightforward. Given a string $W = X_1 \dots X_n$ of length n, where $X_1, \dots, X_n$ are letters in the alphabet $\mathcal{A} = \{A, C, G, T\}$, there are two natural strings of length $n-1$ associated to W: its *initial substring* $X_1 \dots X_{n-1}$, consisting of its first $n-1$ letters, and its *terminal substring* $X_2 \dots X_n$, consisting of its last $n-1$ letters. So, the n-mer $\{W, W^c\}$ has naturally associated to it four strings: the initial substring $X_1 \dots X_{n-1}$ and the terminal substring $X_2 \dots X_n$ of W, and the initial substring $X_n^c \dots X_2^c$ and the terminal substring $X_{n-1}^c \dots X_1^c$ of W^c. These strings are organized into the two $(n-1)$-mers $\{X_1 \dots X_{n-1}, X_{n-1}^c \dots X_1^c\}$ and $\{X_2 \dots X_n, X_n^c \dots X_2^c\}$. In the graph $\mathcal{G}_n$, the edge n-mer $\{X_1 \dots X_n, X_n^c \dots X_1^c\}$ joins the two vertex $(n-1)$-mers $\{X_1 \dots X_{n-1}, X_{n-1}^c \dots X_1^c\}$ and $\{X_2 \dots X_n, X_n^c \dots X_2^c\}$.

We note that for each $n \geq 2$, the graph $\mathcal{G}_n$ contains loops. While $\mathcal{G}_2$ contains multiple edge 2-mers joining a given pair of vertex 1-mers, this does not occur in $\mathcal{G}_n$ for $n \geq 3$: in $\mathcal{G}_n$ for $n \geq 3$, there is at most one edge n-mer joining any given pair of vertex $(n-1)$-mers.

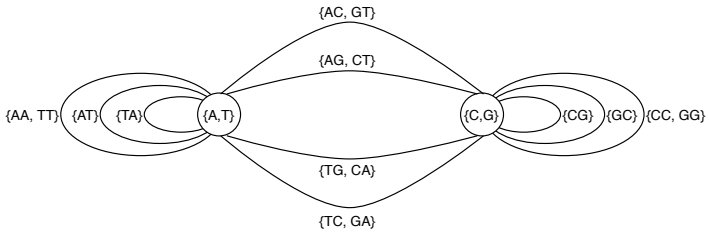

Fig. 1. The 2-mer graph $\mathcal{G}_2$

As an example, consider the graph $\mathcal{G}_2$. The edge 2-mers $\{AT\}$, $\{AA, TT\}$, and $\{TA\}$ join the vertex 1-mer $\{A, T\}$ to itself, and so are loops incident to $\{A, T\}$. The edge 2-mers $\{CG\}$, $\{GC\}$, and $\{CC, GG\}$ are loops incident to the vertex 1-mer $\{C, G\}$. The four edge 2-mers $\{AC, GT\}$, $\{AG, CA\}$, $\{CA, TG\}$, $\{GA, TC\}$ join the vertex 1-mers $\{A, T\}$ and $\{C, G\}$.

The valency of the vertex $\{A, T\}$ is 10: there are four 2-mer edges joining the vertex 1-mer $\{A, T\}$ to the vertex 1-mer $\{C, G\}$, each of which contributes 1 to the valency of $\{A, T\}$, and there are three loops at $\{A, T\}$, namely the 2-mers $\{AA, TT\}$, $\{AT\}$, and $\{TA\}$, each of which contributes 2 to the valency of $\mathcal{G}_2$ at $\{A, T\}$. Using a similar argument, the valency of $\{C, G\}$ is 10.

We spend the remainder of this Section discussing the connection between strings in the alphabet $\mathcal{A} = \{A, T, C, G\}$ and paths in $\mathcal{G}_n$.

A string S in the alphabet $\mathcal{A} = \{A, T, C, G\}$ of length $k \geq n$ gives rise to a path $\mathcal{P}_S$ in $\mathcal{G}_n$ in a natural way, where the consecutive n-mers in the path correspond to consecutive substrings of n contiguous letters in S. Specifically, set $S = X_1 \dots X_k$, where each X_j is a letter in the alphabet $\mathcal{A} = \{A, T, C, G\}$. Define W_j to be the substring $W_j = X_j X_{j+1} \dots X_{n+j-1}$ of S consisting of n contiguous letters starting at X_j. This breaks S up into a sequence of overlapping substrings of length n, so that the edge n-mers containing the strings W_j and W_{j+1} are both incident to the vertex $(n-1)$-mer containing the string $X_{j+1} \dots X_{n+j-1}$. The path $\mathcal{P}_S = \{\{W_1, W_1^c\}, \dots, \{W_{k-n+1}, W_{k-n+1}^c\}\}$ is the desired path arising from S.

Consider the case $n = 2$. The string $S = AAATCCGT$ gives rise to the following path $\mathcal{P}_S$ in $\mathcal{G}_2$:

$$\{\{AA, TT\}, \{AA, TT\}, \{AT\}, \{TC, GA\}, \{CC, GG\}, \{CG\}, \{GT, AC\}\}.$$

In this language, a string S is a universal footprinting template for n-mers if its associated path P_S covers every edge of $\mathcal{G}_n$. Furthermore, if the string S contains one and only one string from each n-mer, then its corresponding path $\mathcal{P}_S$ contains each edge n-mer of $\mathcal{G}_n$ exactly once, and so S is a minimal universal footprinting template for n-mers.

The main technical issue in this paper arises from the fact that passing from paths in $\mathcal{G}_n$ to strings in the alphabet $\mathcal{A} = \{A, T, C, G\}$ is not as

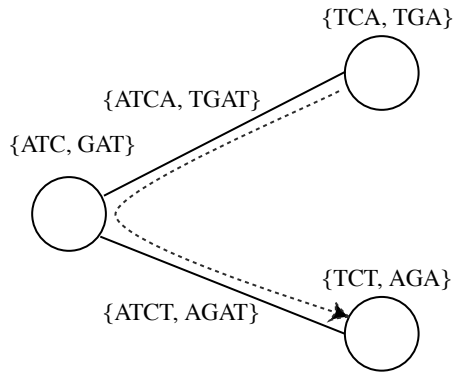

Fig. 2. A non-admissable path

transparent an operation as passing from strings to paths. The cause of this is that every edge of $\mathcal{G}_n$ is an n-mer, which is a pair of strings, rather than a single string.

Let $W = X_1 \dots X_n$ be a string of length n, and consider the edge n-mer $\{W, W^c\}$ of $\mathcal{G}_n$. The choice of one of the strings from this edge imposes a natural *direction of travel* along the edge. Specifically, the string $W = X_1 \dots X_n$ carries a natural direction of travel from the vertex $(n-1)$-mer $\{X_1 \dots X_{n-1}, X_{n-1}^c \dots X_1^c\}$ containing the substring $X_1 \dots X_{n-1}$ and towards the vertex $(n-1)$-mer $\{X_2 \dots X_n, X_n^c \dots X_2^c\}$ containing the substring $X_2 \dots X_n$. The complementary string $W^c = X_n^c \dots X_1^c$ carries the opposite direction of travel, from the the vertex $(n-1)$-mer $\{X_n^c \dots X_2^c, X_2 \dots X_n\}$ containing the substring $X_n^c \dots X_2^c$ and towards the vertex $(n-1)$-mer $\{X_{n-1}^c \dots X_1^c, X_1 \dots X_{n-1}\}$ containing the substring $X_{n-1}^c \dots X_1^c$. (We note that there are special considerations involving self-complementary n-mers appearing either as vertices or as edges.)

Consider the case $n = 4$. The edges $\{ATCA, TGAT\}$, $\{ATCT, AGAT\}$ of $\mathcal{G}_4$ are both incident to the vertex 3-mer $\{ATC, GAT\}$, and there is a path in the graph of length 2 running along them from the vertex $\{TCA, TGA\}$ to the vertex $\{TCT, AGA\}$. See Figure 2. However this path does not arise (in the manner described above) from an DNA segment of length 5. If it did so then the first three letters on one of its strands would have to be TCA or TGA while the last three letters on the same strand would have to be TCT or AGA. Since the third letter on the strand is the first letter of the last three, we see that the last three letters must be AGA and the strand must read $TCAGA$ or $TGAGA$. In the first case the path runs through the vertex $\{CAG, CTG\}$ and in the second case it runs through the vertex $\{GAG, CTC\}$. In neither case does it run through the required vertex $\{ATC, GAT\}$. (Note that the DNA segment described by the strand $TCAGA$ does define a path of length 2 from the vertex $\{TCA, TGA\}$ to the vertex $\{TCT, AGA\}$ but this does not run along the required edges.)

In terms of the natural direction of travel of these two n-mers at the common incident vertex 3-mer, the string $ATCA$ in the first 4-mer has its natural direction of travel from the vertex 3-mer $\{ATC, GAT\}$, since ATC is the initial substring of $ATCA$. Similarly, the string $TGAT$ has its natural direction of travel towards the vertex 3-mer $\{ATC, GAT\}$, since GAT is the terminal substring of $TGAT$. The string $ATCT$ in the first 4-mer has its natural direction of travel from the vertex 3-mer $\{ATC, GAT\}$, since ATC is the initial substring of $ATCA$. Similarly, the string $AGAT$ has its natural direction of travel towards the vertex 3-mer $\{ATC, GAT\}$, since GAT is the terminal substring of $AGAT$. Therefore, it is not possible to admissibly choose a string from the first 4-mer in $\mathcal{P}$ and a string from the second 4-mer in $\mathcal{P}$ in such a way that the natural directions of travel match up, in the sense that one is towards the vertex 3-mer and the other is from.

With this in mind, we make the following definition:

Definition 1. *An* admissable path *in the graph $\mathcal{G}_n$ is a path of the form $\mathcal{P} = \{\{W_1, W_1^c\}, \{W_2, W_2^c\}, \ldots, \{W_k, W_k^c\}\}$ such that for each $j = 1, \ldots k-1$, there exists a string U of length $n+1$ whose initial substring of length n is contained in $\{W_j, W_j^c\}$ and whose terminal substring of length n is contained in $\{W_{j+1}, W_{j+1}^c\}$.*

This is equivalent to requiring that the initial substring of one of the strings in $\{W_{j+1}, W_{j+1}^c\}$ agrees with the terminal substring of one of the strings in $\{W_j, W_j^c\}$. This definition encodes what occurs when a path $\mathcal{G}_n$ arises from a string in the alphabet $\mathcal{A} = \{A, T, C, G\}$, and this definition describes exactly what is needed to ensure that a path in $\mathcal{G}_n$ gives rise to a string in the alphabet $\mathcal{A}$. In order to show that it is always possible to construct a path in this way, we need to have a detailed description of the structure of $\mathcal{G}_n$ at its vertices.

In summary, every segment of DNA can be described by an admissable path in our graph, and every admissable path describes a segment of DNA. The two directions of travel along an admissable path are both admissable, and each transcribes one of the strands of the corresponding segment. In particular a minimal UFT corresponds to a shortest possible admissable path which crosses each edge of the graph at least once.

Now recall the classical (fast) algorithm for constructing Eulerian paths from Section 2. The first stage relies on a pairing between edges incident to each vertex, so that on arrival at a vertex along a given edge the pairing tells us how to leave the vertex along the paired edge. In this paper we are largely concerned with the problem of constructing admissable Eulerian paths and again we want a fast algorithm to construct them. We achieve this by generalising the classical algorithm as follows. We start (where possible) with a pairing of edges incident to each vertex so that each pair forms an admissable path of length 2 through the given vertex. In our example above we could pair the edges $\{TCAG, CTGA\}$ and $\{CAGA, TCTG\}$ at the vertex $\{CAG, CTG\}$, but we cannot pair the edges $\{ATCA, TGAT\}$ and $\{ATCT, AGAT\}$ at the vertex $\{ATC, GAT\}$. Such a pairing, if it exists,

will be called an *admissable pairing*. The existence of an admissable pairing is both necessary and sufficient for us to construct an admissable Eulerian path in the graph.

We will show that for n odd (or $n = 2$) the graph $\mathcal{G}_n$ does admit an admissable pairing, while for $n \geq 4$ and even, the graph $\mathcal{G}_n$ does not admit an admissable pairing.

We close this Section with a discussion of the classical de Bruijn graph and its relation to the graph $\mathcal{G}_n$. Fix a finite alphabet $\mathcal{A}$. (We consider the alphabet $\mathcal{A} = \{A, C, G, T\}$, but the basic details of the construction work for any other finite alphabet.) For $n \geq 2$, the de Bruijn graph $\mathcal{B}_n$ is defined to be the graph whose edges are strings of length n in an alphabet $\mathcal{A}$ and whose vertices are strings of length $n-1$ in $\mathcal{A}$. The edge $X_1 \dots X_n$ of length n, where each X_j is a letter in the alphabet $\mathcal{A}$, joins the vertices $X_1 \dots X_{n-1}$ and $X_2 \dots X_n$. A string S of length $k \geq n$ in the alphabet $\mathcal{A}$ corresponds to a path in $\mathcal{B}_n$, in a manner completely analogous to the manner in which S gives rise to a path in $\mathcal{G}_n$; namely, start at the vertex defined by the first $n-1$ letters of S, move along the edge defined by the first n letters of S to arrive at the vertex defined by 2^{nd} through n^{th} letters of S, et cetera. The de Bruijn graph has been used with great effect in DNA sequencing questions; see for instance [7].

However, give a segment of DNA composed of a string S and its complementary string S^c, we see that S and S^c give rise to different paths in $\mathcal{B}_n$. The graph $\mathcal{G}_n$ described above removes this ambiguity and, given the 1-1 correspondence between segments of DNA and admissible paths in $\mathcal{G}_n$, may have further utility in DNA sequencing problems.

The operation ^{c} of complementation on strings in the alphabet $\mathcal{A} = \{A, C, G, T\}$ defines an action of the cyclic group of order 2 on both the vertex set and the edge set of the de Bruijn graph $\mathcal{B}_n$. Furthermore, this involution preserves the incidence of vertices and edges, so that if an edge in $\mathcal{B}_n$ comprised of the string W of length n is incident to the vertex comprised of the string V of length $n-1$, then the edge comprised of the string W^c is incident to the vertex comprised of the string V^c. In particular, the action of the complementing operation ^{c} extends to an involution on $\mathcal{B}_n$. The quotient object may be interpreted as our graph $\mathcal{G}_n$. However, the involution on $\mathcal{B}_n$ has fixed points, namely the self-complementary strings (which are vertices if n is odd and edges if n is even). These fixed points create significant technical difficulties in passing from $\mathcal{B}_n$ to its quotient $\mathcal{G}_n$. The resolution of these technical difficulties occupies a significant portion of this paper.

4. Structure of $\mathcal{G}_n$

In this Section, we give a detailed description of the structure of the graph $\mathcal{G}_n$ at its vertices. The information in this Section feeds directly into the discussion of the implementation of the algorithm for the construction of (minimal) universal footprinting templates for n-mers. The information required includes the valencies of the vertices of $\mathcal{G}_n$, as well as a description of

how the edges incident to a given vertex can be paired, to ensure that the paths constructed are admissible. The proofs of Proposition 2 and Proposition 3 are given at the end of the paper.

We begin by showing that there are two distinct types of loops that can arise in $\mathcal{G}_n$ for $n \geq 2$, and discuss their behavior. Recall that a *loop* is an edge n-mer $\{W, W^c\}$ both of whose endpoints are the same vertex $(n-1)$-mer $\{V, V^c\}$. There are two possibilities: either both the initial and terminal substrings of W of length $n-1$ (or of W^c) are V, or the initial substring of W of length $n-1$ is V and the terminal substring of W of length $n-1$ is V^c (or vice versa).

We refer to the former type of loop, in which the initial and terminal substrings of W of length $n-1$ are the same, as a *standard loop*. Write $W = Y_1 \dots Y_n$ and $V = X_1 \dots X_{n-1}$. Since both the initial and terminal $n-1$ letters of W are $V = X_1 \dots X_{n-1}$, we get the following relationships between the Y_k and the X_j:

$$\begin{array}{ccccccc} Y_1 & Y_2 & \dots\dots & Y_{n-1} & Y_n \\ \| & \| & & \| & \| \\ X_1 & X_2 & \dots\dots & X_{n-1} & \| \\ & \| & & \| & \| \\ & X_1 & \dots\dots & X_{n-2} & X_{n-1} \end{array}$$

These relationships yield that $Y_1 = Y_2 = \dots = Y_{n-1} = Y_n$. Hence, in any $\mathcal{G}_n$, there are only two standard loops, namely $\{A^n, T^n\}$ and $\{C^n, G^n\}$.

We refer to the other type of loop, in which the initial and terminal substrings of W of length $n-1$ are complementary, as a *lollipop loop*. Write $W = Y_1 \dots Y_n$ and $V = X_1 \dots X_{n-1}$. Since the initial $(n-1)$ letters of W are $V = X_1 \dots X_{n-1}$ and the terminal $(n-1)$ letters of W are $V^c = X_{n-1}^c \dots X_1^c$, we get the following relationships between the Y_k and the X_j:

$$\begin{array}{ccccccc} Y_1 & Y_2 & \dots\dots & Y_{n-1} & Y_n \\ \| & \| & & \| & \| \\ X_1 & X_2 & \dots\dots & X_{n-1} & \| \\ & \| & & \| & \| \\ & X_{n-1}^c & \dots & X_2^c & X_1^c \end{array}$$

These relationships yield that W is self-complementary, so that $W = W^c$, and so can be written as $W = UU^c$ for some string U of length $\frac{1}{2}n$, namely $U = Y_1 \dots Y_{n/2}$. (The length of W must be even, for otherwise the middle letter of W would equal its own complement, which is impossible.) Conversely, if W is self-complementary, then it has even length; it can be written as $W = UU^c$ for some string U of length $\text{length}(U) = \frac{1}{2}\text{length}(W)$; and it is necessarily a lollipop loop. By way of illustration, the self-complementary edge 4-mer $\{ACGT\}$ in $\mathcal{G}_4$ is a lollipop loop joining the vertex 3-mer $\{ACG, CGT\}$ to itself.

So we have established the following Proposition:

Proposition 1. *If n is odd (and $n \geq 3$), then $\mathcal{G}_n$ contains two standard loops $\{A^n, T^n\}$ and $\{C^n, G^n\}$, and $\mathcal{G}_n$ contains no lollipop loops.*

If n is even, then $\mathcal{G}_n$ contains two standard loops $\{A^n, T^n\}$ and $\{C^n, G^n\}$, and $\mathcal{G}_n$ contains $4^{n/2}$ lollipop loops $\{W\}$, where W ranges over all the self-complementary strings of length n (which are parametrized by the strings of length $\frac{1}{2}n$).

We close this Section by stating the two Propositions which describe the structure of $\mathcal{G}_n$ at its vertices. Proposition 2 covers the case of n odd, while Proposition 3 covers the case of n even.

Before stating Proposition 2, which describes the structure of our graph $\mathcal{G}_n$ for odd n, we illustrate the possibilities stated therein by considering the case of $\mathcal{G}_3$. The vertices are the 2-mers $\{V, V^c\}$ where V ranges over the strings of length 2. The edges in $\mathcal{G}_3$ incident to $\{V, V^c\}$ are the 3-mers $\{XV, V^cX^c\}$ and $\{VX, X^cV^c\}$ as X ranges over the letters A, C, G, T, and so the non-loop edge 3-mers at each vertex can be admissibly paired. In order to understand the structure of the graph $\mathcal{G}_3$ at its vertices, we need to determine which of these 3-mers are standard loops, and whether any two of them are equal. Since 3 is odd, there are no lollipop loops, as there are no self-complementary 3-mers.

Suppose that V is self-complementary; for example take $V = AT$. The 3-mers obtained by adding a letter to the beginning of V are $\{AAT, ATT\}$, $\{CAT, ATG\}$, $\{GAT, ATC\}$, $\{TAT, ATA\}$; the 3-mers obtained by adding a letter to the end of V are $\{ATA, TAT\}$, $\{ATC, GAT\}$, $\{ATG, CAT\}$, $\{ATT, AAT\}$. Note that the four 3-mers obtained by adding a letter to the beginning of V are the same four 3-mers that we obtain by adding a letter to the end of V. Also, the four 3-mers obtained by adding a letter to the beginning of V are distinct, and none of them is a standard loop. So the vertex $\{V, V^c\}$ is incident to exactly 4 edges, namely $\{AAT, ATT\}$, $\{CAT, ATG\}$, $\{GAT, ATC\}$, $\{TAT, ATA\}$. These can be admissably paired at the vertex $\{AT, AT\}$ by, for example, pairing $\{ATA, TAT\}$ with $\{ATG, CAT\}$ and by pairing $\{CAT, ATG\}$ with $\{GAT, ATC\}$. Another possible pairing, which we call the *standard pairing*, is to pair an edge of the form $\{XV, V^cX^c\}$ with the edge of the form $\{VX, X^cV^c\}$, where X is one of the letters A, C, G, T.

A similar argument applies to the other self-complementary vertex 2-mers $\{TA\}$, $\{CG\}$, and $\{GC\}$.

Suppose next that V contains only a single repeated letter; for example take $V = CC$. The 3-mers obtained by adding a letter to the beginning of V are $\{ACC, GGT\}$, $\{CCC, GGG\}$, $\{GCC, GGC\}$, and $\{TCC, GGA\}$, while the 3-mers obtained by adding a letter to the end of V are $\{CCA, TGG\}$, $\{CCC, GGG\}$, $\{CCG, CGG\}$, and $\{CCT, AGG\}$. One of the 3-mers in the former list coincides with one in the latter list, namely the standard loop $\{CCC, GGG\}$. The remaining six 3-mers $\{ACC, GGT\}$, $\{GCC, GGC\}$, $\{TCC, GGA\}$, $\{CCA, TGG\}$, $\{CCG, CGG\}$, $\{CCT, AGG\}$ are distinct, and

none of them is a standard loop. The non-loop edge 3-mers incident to the vertex $\{CC, GG\}$ may be admissably paired by pairing an edge of the form $\{XCC, GGX^c\}$ with the edge $\{CCX, X^cGG\}$ as X ranges over the letters A, G, T; this is the standard pairing. The standard loop is not itself paired with anything, but this will not affect the construction of an admissable Eulerian path.

A similar argument applies to the other vertex 2-mer $\{AA, TT\}$ that consists of powers of a single letter.

The remaining case is that V is neither self-complementary nor a power of a single letter; for example, take $V = AC$. The 3-mers obtained by adding a letter to the beginning of V are $\{AAC, GTT\}$, $\{CAC, GTG\}$, $\{GAC, GTC\}$, and $\{TAC, GTA\}$, while the 3-mers obtained by adding a letter to the end of V are $\{ACA, TGT\}$, $\{ACC, GGT\}$, $\{ACG, CGT\}$, and $\{ACT, AGT\}$. These eight 3-mers are distinct and none of them is a standard loop. Each edge of the form $\{XAC, GTX^c\}$ may be admissably paired with the edge $\{ACX, X^cGT\}$ where X is one of the letters A, C, G, T.

These are the only possibilities that occur in the general case of n odd.

Proposition 2. *Suppose that n is odd and $n \geq 3$. There are no lollipop loops in $\mathcal{G}_n$.*

- *Suppose the vertex $(n-1)$-mer $\{V, V^c\}$ is self-complementary. Then, its valency is 4. There is no standard loop incident to $\{V\}$. The edges incident to $\{V\}$ are the n-mers $\{AV, VT\}$, $\{CV, VG\}$, $\{GV, VC\}$, $\{TV, VA\}$ and these may be admissably paired.*
- *Suppose the vertex $(n-1)$-mer $\{V, V^c\}$ is not self-complementary. Then, its valency is 8.*
  - *If $V = X^{n-1}$ for some letter X, then the edges incident to $\{V, V^c\}$ are the n-mer $\{X^n, (X^c)^n\}$, which is a standard loop, and the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X. The number of n-mers of the form $\{ZV, V^cZ^c\}$ is equal to the number of n-mers of the form $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X, and these may be admissably paired.*
  - *If $V \neq X^{n-1}$ for any letter X, then there is no standard loop incident to $\{V, V^c\}$. The edges incident to $\{V, V^c\}$ are the n-mers $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, $\{TV, V^cA\}$, $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$. The number of n-mers of the form $\{ZV, V^cZ^c\}$ is equal to the number of n-mers of the form $\{VZ, Z^cV^c\}$, where Z ranges over the letters A, C, G, T, and these may be admissably paired.*

In particular, the edges incident to each vertex of $\mathcal{G}_n$ which are not standard loops may be admissably paired at that vertex since at each vertex $(n-1)$-mer $\{V, V^c\}$, the number of non-loop edge n-mers ending with the string V (respectively V^c) is equal to the number of edge n-mers beginning with the string V (respectively V^c).

The proof of Proposition 2 is given in Section 7.

We will see in Section 5 that this information on the admissable pairings of edge n-mers at each vertex of $\mathcal{G}_n$ for n odd is sufficient to allow the construction of minimal universal footprinting template for n-mers of length naive(n).

Before stating Proposition 3, which describes the structure of our graph $\mathcal{G}_n$ for even n, we illustrate the possibilities stated therein by considering the case of $\mathcal{G}_4$. The vertices are the 3-mers $\{V, V^c\}$, where V ranges over the strings of length 3. The edges in $\mathcal{G}_4$ incident to $\{V, V^c\}$ are the 4-mers $\{XV, V^cX^c\}$ and $\{VX, X^cV^c\}$ as X ranges over the letters A, C, G, T. In order to understand the structure of the graph $\mathcal{G}_4$ at its vertices, we need to determine which of these 4-mers are standard loops; which are lollipop loops; and whether any two of them are equal. As we saw above, each self-complementary 4-mer corresponds to a lollipop loop in $\mathcal{G}_4$, and there are at most two lollipop loops incident to any vertex 3-mer.

Suppose that V is contained in two self-complementary strings of length 4; for example, take $V = ATA$. There are two lollipop loops incident to $\{V, V^c\}$, namely $\{ATAT\}$, obtained by adding T to the end of V, and $\{TATA\}$, obtained by adding T to the beginning of V. We can therefore admissably pair these loops. The remaining six 4-mers incident to $\{V, V^c\}$ are $\{AATA, TATT\}$, $\{CATA, TATG\}$, $\{GATA, TATC\}$ (obtained by adding a letter to the beginning of V) and $\{ATAA, TTAT\}$, $\{ATAC, GTAT\}$, $\{ATAG, CTAT\}$ (obtained by adding a letter to the end of V). These six 4-mers are distinct, and none of them is a standard loop or a lollipop loop. So these non-loop edges may be admissably paired by pairing an edge of the form $\{XATA, TATX^c\}$ with the edge $\{ATAX, X^cTAT\}$ where X is one of the letters A, C, G. Each lollipop loop contributes 2 to the valency of $\mathcal{G}_4$ at $\{V, V^c\}$, while each of the six edge 4-mers contributes 1, and so the valency of $\mathcal{G}_4$ at $\{V, V^c\}$ is 10.

Suppose that V is contained in exactly one self-complementary string of length 4; for example, take $V = CAT$. There is one lollipop loop incident to $\{V, V^c\}$, namely $\{CATG\}$, obtained by adding G to the end of V. The remaining seven 4-mers incident to $\{V, V^c\}$ are $\{ACAT, ATGT\}$, $\{CCAT, ATGG\}$, $\{GCAT, ATGC\}$, $\{TCAT, ATGA\}$ (obtained by adding a letter to the beginning of V) and $\{CATA, TATG\}$, $\{CATC, GATG\}$, $\{CATT, AATG\}$ (obtained by adding a letter to the end of V). These seven 4-mers are distinct, and none of them is a standard loop or a lollipop loop. The lollipop loop contributes 2 to the valency of $\mathcal{G}_4$ at $\{V, V^c\}$, while each of the seven edge 4-mers contributes 1, and so the valency of $\mathcal{G}_4$ at $\{V, V^c\}$ is 9. It follows that we cannot pair, let alone admissably pair, the edges incident to the vertex $\{CAT, ATG\}$.

Suppose that V is a power of a single letter; for example, take $V = AAA$. There is one standard loop incident to $\{V, V^c\}$, namely $\{AAAA, TTTT\}$, which can be obtained by adding A to either the beginning or the end of V. The remaining six 4-mers incident to $\{V, V^c\}$ are $\{CAAA, TTTG\}$, $\{GAAA, TTTC\}$, $\{TAAA, TTTA\}$ (obtained by adding a letter to the be-

ginning of V) and $\{AAAC, GTTT\}$, $\{AAAG, CTTT\}$, $\{AAAT, ATTT\}$ (obtained by adding a letter to the end of V). These six 4-mers are distinct, and none of them is a lollipop loop; they may, as before, be admissably paired, using for instance the standard pairing. The standard loop contributes 2 to the valency of $\mathcal{G}_4$ at $\{V, V^c\}$, while each of the six edge 4-mers contributes 1, and so the valency of $\mathcal{G}_4$ at $\{V, V^c\}$ is 8.

The remaining case is that V is neither contained in a self-complementary string nor a power of a single letter; for example, take $V = ACC$. The 4-mers obtained by adding a letter to the beginning of V are $\{AACC, GGTT\}$, $\{CACC, GGTG\}$, $\{GACC, GGTC\}$, and $\{TACC, GGTA\}$, while the 4-mers obtained by adding a letter to the end of V are $\{ACCA, TGGT\}$, $\{ACCC, GGGT\}$, $\{ACCG, CGGT\}$, and $\{ACCT, AGGT\}$. These eight 4-mers are distinct and none of them is a standard loop or a lollipop loop. Once more we may use the standard pairing to admissably pair the edges at this vertex. Each edge contributes 1 to the valency of $\mathcal{G}_4$ at $\{V, V^c\}$, and so the valency of $\mathcal{G}_4$ at $\{V, V^c\}$ is 8.

These are the only possibilities that occur in the general case of n even (and $n \geq 4$).

Proposition 3. *Suppose that n is even and $n \geq 4$. Let $m = \frac{1}{2}(n-2)$.*

- *Suppose there is a self-complementary edge incident to the vertex $(n-1)$-mer $\{V, V^c\}$. Then, $\{V, V^c\}$ has the form $\{XUU^c, UU^cX^c\}$, where X is one of the letters A, C, G, T, and U is a string of length m.*
  - *If $U = (X^cX)^{m/2}$, then there are two lollipop loops incident to $\{V, V^c\}$, namely the n-mers $\{X(X^cX)^mX^c\}$ and $\{X^cX(X^cX)^m\}$. There is no standard loop incident to $\{V, V^c\}$. If $V = XUU^c$, the remaining edges incident to $\{V, V^c\}$ are the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X^c. (If instead $V = UU^cX^c$, the remaining edges incident to $\{V, V^c\}$ are the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X.) The number of n-mers of the form $\{ZV, V^cZ^c\}$ is equal to the number of n-mers of the form $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X. The valency of $\mathcal{G}_n$ at $\{V, V^c\}$ is 10. The lollipop loops can be admissably paired with one another and the other edges may be admissably paired.*
  - *If $U \neq (X^cX)^{m/2}$, then there is one lollipop loop incident to $\{V, V^c\}$, namely the n-mer $\{XUU^cX^c\}$. There is no standard loop incident to $\{V, V^c\}$. If $V = XUU^c$, the remaining edges incident to $\{V, V^c\}$ are the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X^c, together with the n-mer $\{X^cV, V^cX\}$. (If instead $V = UU^cX^c$, the remaining edges incident to $\{V, V^c\}$ are the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X, together with the n-mer $\{XV, V^cX^c\}$.) The valency of $\mathcal{G}_n$ at $\{V, V^c\}$ is 9. There is no admissable pairing at these vertices.*

- *Suppose there is no self-complementary edge incident to* $\{V, V^c\}$*. Then, there is no lollipop loop incident to* $\{V, V^c\}$*. The valency of* $\mathcal{G}_n$ *at* $\{V, V^c\}$ *is* 8 *and the edges may be admissably paired.*
  - *If* $V = X^{n-1}$*, then the edges incident to* $\{V, V^c\}$ *are the standard loop* $\{X^n, (X^c)^n\}$ *and the six* n*-mers* $\{ZV, V^cZ^c\}$ *and* $\{VZ, Z^cV^c\}$*, where* Z *ranges over the letters not equal to* X*. The number of* n*-mers of the form* $\{ZV, V^cZ^c\}$ *is equal to the number of* n*-mers of the form* $\{VZ, Z^cV^c\}$*, where* Z *ranges over the letters not equal to* X*.*
  - *If* $V \neq X^{n-1}$ *for any letter* X*, then there is no standard loop incident to* $\{V, V^c\}$*. The edges incident to* $\{V, V^c\}$ *are the eight* n*-mers* $\{AV, V^cT\}$*,* $\{CV, V^cG\}$*,* $\{GV, V^cC\}$*,* $\{TV, V^cA\}$*,* $\{VA, TV^c\}$*,* $\{VC, GV^c\}$*,* $\{VG, CV^c\}$*,* $\{VT, AV^c\}$*. The number of* n*-mers of the form* $\{ZV, V^cZ^c\}$ *is equal to the number of* n*-mers of the form* $\{VZ, Z^cV^c\}$*, where* Z *ranges over the letters* A*,* C*,* G*,* T*.*

The proof of Proposition 3 is given in Section 8.

Notice that in particular, since there are more than two vertices of odd valency, there will not even exist an Eulerian path in $\mathcal{G}_n$, let alone an admissable Eulerian path, and so there cannot exist a minimal universal footprinting template for n-mers of length naive(n). Hence, in order to implement the algorithm for finding an Eulerian path, we need to modify the algorithm slightly. A full discussion is given in Section 6. Roughly, we begin by constructing a collection of short circular paths involving the vertex $(n-1)$-mers incident to self-complementary edge n-mers. The information on the pairing of edge n-mers at each vertex of $\mathcal{G}_n$ for n even is sufficient to allow the construction of a complete collection of circular paths in $\mathcal{G}_n$ with these short circular paths marked as used, and hence the construction of a universal footprinting template for n-mers, together with upper and lower bounds on its length.

There are two special cases that we consider separately, in which the graphs are slightly different from the general cases. The first is the case of $n = 1$. In this case, the graph theoretic approach does not work; there are no edges in the graph $\mathcal{G}_1$, as there are no 0-mers. However, as we saw in Section 1, we have a complete solution in this case.

The other is the case $n = 2$. In this case, unlike the general case of n even, the graph $\mathcal{G}_2$ has no vertices of odd valency, as there are no vertices incident to a single lollipop loop; rather, both vertices of $\mathcal{G}_2$ are incident to two lollipop loops. (This is why we need to exclude the case of $n = 2$ from consideration in Proposition 3.) In this case, by referring back to Figure 1, it is easy to see that it is possible to pair the non-loop edges. In fact, the standard pairing, of pairing $\{XA, TX^c\}$ with $\{AX, X^cT\}$ for $X = C$ or $X = G$, yields an admissible pairing, as does the pairing of $\{XT, AX^c\}$ with $\{AX, X^cT\}$ for $X = C$ or $X = G$. In fact, this is a complete list of admissible pairings.

5. The algorithm for n odd

We are now ready to describe the algorithm in detail. In this Section, we describe the algorithm in detail for n odd, and work through all of the details for the case $n = 3$. This includes the proof of Theorem 1.

In Section 2, we saw that a graph in which there are zero or two vertices of odd valency always possesses an Eulerian path, and we described the algorithm for finding such a path. In order to construct an admissable Eulerian path in $\mathcal{G}_n$ for n odd, we use the description of the vertex $(n-1)$-mers and edge n-mers of $\mathcal{G}_n$ given in Proposition 2. The information about the edge n-mers is important, given the discussion of the relationship between strings and paths in $\mathcal{G}_n$ given in Section 3.

The algorithm for constructing an admissable Eulerian path in $\mathcal{G}_n$ is a mild adaptation of the standard algorithm for finding an Eulerian path in a graph discussed in Section 2. For n odd, Proposition 2 yields that every vertex of $\mathcal{G}_n$ has even valency and that the edge n-mers at every vertex $(n-1)$-mer can be admissibly paired. This yields all the information that is needed to generate a minimal universal footprinting template for n-mers. We are now in a position to describe the algorithm.

Step 1: pair the non-loop edge n-mers at each vertex $(n-1)$-mer:

As described in Proposition 2, the number of non-loop edge n-mers at each vertex $(n-1)$-mer $\{V, V^c\}$ is even, and the number of edge n-mers of the form $\{XV, V^cX^c\}$ (obtained by adding a letter to the beginning of V) is equal to the number of edge n-mers of the form $\{VX, X^cV^c\}$ (obtained by adding a letter to the end of V). To summarize:

- If $\{V, V^c\}$ is self-complementary, so that $V = V^c$, there are four edge n-mers incident to $\{V, V^c\}$, namely $\{AV, VT\}$, $\{CV, VG\}$, $\{GV, VC\}$, and $\{TV, VA\}$;
- if $V = X^{n-1}$, where X is one of the letters A, C, G, T, then there is a standard loop incident to $\{V, V^c\}$, and there are six other edge n-mers incident to $\{V, V^c\}$, three of which have the form $\{ZV, V^cZ^c\}$ and three of which have the form $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X;
- in all other cases, there are eight edge n-mers incident to $\{V, V^c\}$, four of which have the form $\{XV, V^cX^c\}$ (obtained by adding a letter to the beginning of V) and four of which have the form $\{VX, X^cV^c\}$ (obtained by adding a letter to the end of V);
- the only loops which occur in $\mathcal{G}_n$ are the two standard loops $\{A^n, T^n\}$ and $\{C^n, G^n\}$.

Thus, it is always possible to pair the non-loop edge n-mers at each vertex $(n-1)$-mer, so that each non-loop edge n-mer of the form $\{XV, V^cX^c\}$ is paired with a non-loop edge n-mer of the form $\{VX, X^cV^c\}$. Any path constructed using these pairings is an admissible path, and so will give rise to a string. Moreover, since every vertex in $\mathcal{G}_n$ has even valency, the path constructed will be an Eulerian path, and so the associated string will be a

minimal universal footprinting template for n-mers. A non-standard pairing is given below:

vertex	pairing of incident edges
$\{A^{n-1}, T^{n-1}\}$	$\{CA^{n-1}, T^{n-1}G\} \leftrightarrow \{A^{n-1}G, CT^{n-1}\}$ $\{GA^{n-1}, T^{n-1}C\} \leftrightarrow \{A^{n-1}T, AT^{n-1}\}$ $\{TA^{n-1}, T^{n-1}A\} \leftrightarrow \{A^{n-1}C, GT^{n-1}\}$
$\{C^{n-1}, G^{n-1}\}$	$\{AC^{n-1}, G^{n-1}T\} \leftrightarrow \{C^{n-1}G, CG^{n-1}\}$ $\{GC^{n-1}, G^{n-1}C\} \leftrightarrow \{C^{n-1}T, AG^{n-1}\}$ $\{TC^{n-1}, G^{n-1}A\} \leftrightarrow \{C^{n-1}A, TG^{n-1}\}$
$\{V, V^c\}$ where $V = V^c$	$\{AV, VT\} \leftrightarrow \{CV, VG\}$ $\{VC, GV\} \leftrightarrow \{VA, TV\}$
$\{V, V^c\}$, otherwise	$\{AV, V^cT\} \leftrightarrow \{VG, CV^c\}$ $\{CV, V^cG\} \leftrightarrow \{VT, AV^c\}$ $\{GV, V^cC\} \leftrightarrow \{VA, TV^c\}$ $\{TV, V^cA\} \leftrightarrow \{VC, GV^c\}$

It is this pairing that prevents difficulties arising in navigating the graph. Note that there is a large number of possible pairings: at a vertex of valency 4, there are three possible pairings; at a vertex of valency 8, there are 24 possible pairings; and the pairings at each vertex $(n-1)$-mer are independent of the pairings at the other vertex $(n-1)$-mers.

Consider the case $n = 3$. The vertex 2-mers in $\mathcal{G}_3$ are $\{AT\}$, $\{TA\}$, $\{CG\}$, $\{GC\}$, $\{AA, TT\}$, $\{CC, GG\}$, $\{AC, GT\}$, $\{AG, CT\}$, $\{CA, TG\}$, and $\{GA, TC\}$. There are two standard loops, $\{AAA, TTT\}$ and $\{CCC, GGG\}$, and no lollipop loops. A non-standard pairing of the non-loop edge 3-mers (different from the general pairing given above, for variety) is given below:

vertex	pairing of incident edges	
$\{AC, GT\}$	$\{AAC, GTT\} \leftrightarrow \{ACG, CGT\}$ $\{GAC, GTC\} \leftrightarrow \{ACA, TGT\}$	$\{CAC, GTG\} \leftrightarrow \{ACT, AGT\}$ $\{TAC, GTA\} \leftrightarrow \{ACC, GGT\}$
$\{AG, CT\}$	$\{AAG, CTT\} \leftrightarrow \{AGT, ACT\}$ $\{GAG, CTC\} \leftrightarrow \{AGG, CCT\}$	$\{CAG, CTG\} \leftrightarrow \{AGC, GCT\}$ $\{TAG, CTA\} \leftrightarrow \{AGA, TCT\}$
$\{CA, TG\}$	$\{ACA, TGT\} \leftrightarrow \{CAC, GTG\}$ $\{GCA, TGC\} \leftrightarrow \{CAT, ATG\}$	$\{CCA, TGG\} \leftrightarrow \{CAG, CTG\}$ $\{TCA, TGA\} \leftrightarrow \{CAA, TTG\}$
$\{GA, TC\}$	$\{AGA, TCT\} \leftrightarrow \{GAT, ATC\}$ $\{GGA, TCC\} \leftrightarrow \{GAC, GTC\}$	$\{CGA, TCG\} \leftrightarrow \{GAG, CTC\}$ $\{TGA, TCA\} \leftrightarrow \{GAA, TTC\}$
$\{AA, TT\}$	$\{CAA, TTG\} \leftrightarrow \{AAG, CTT\}$ $\{TAA, TTA\} \leftrightarrow \{AAC, GTT\}$	$\{GAA, TTC\} \leftrightarrow \{AAT, ATT\}$
$\{CC, GG\}$	$\{ACC, GGT\} \leftrightarrow \{CCA, TGG\}$ $\{TCC, GGA\} \leftrightarrow \{CCG, CGG\}$	$\{GCC, GGC\} \leftrightarrow \{CCT, AGG\}$
$\{AT\}$	$\{AAT, ATT\} \leftrightarrow \{CAT, ATG\}$	$\{ATA, TAT\} \leftrightarrow \{ATC, GAT\}$
$\{TA\}$	$\{ATA, TAT\} \leftrightarrow \{GTA, TAC\}$	$\{CTA, TAG\} \leftrightarrow \{TTA, TAA\}$
$\{CG\}$	$\{ACG, CGT\} \leftrightarrow \{CCG, CGG\}$	$\{GCG, CGC\} \leftrightarrow \{TCG, CGA\}$
$\{GC\}$	$\{AGC, GCT\} \leftrightarrow \{TGC, GCA\}$	$\{CGC, GCG\} \leftrightarrow \{GGC, GCC\}$

We do not include the two standard loops in this pairing operation, as we will make a convention in Step 2 that deals with the standard loops.

Step 2: construct a complete set of maximal circular paths in $\mathcal{G}_n$:

The procedure for constructing maximal circular paths for general n follows the procedure given in Section 2. Namely: choose a vertex $(n-1)$-mer v_0 and an edge n-mer e_1 incident to v_0; the choices of the vertex $(n-1)$-mer v_0 and the incident edge n-mer e_1 force which string contained in e_1 we use. Mark the edge n-mer e_1 as used. At the other vertex $(n-1)$-mer v_1 incident to the edge n-mer e_1, we use the pairing of the edge n-mers at v_1 to find the next edge n-mer e_2 in the path. Mark the edge n-mer e_2 as used. We continue until we can go no further, meaning that we find ourselves at a vertex $(n-1)$-mer at which the pairing yields no edge n-mer that is admissible. Because of the chosen pairing of edges incident to this vertex, we must be at our starting vertex, arriving via the admissable pair with our initial edge. It follows that we have constructed an admissable circular path. We then begin a second path by choosing another vertex $(n-1)$-mer and an incident unmarked edge n-mer, to produce another admissable circular path and continuing until all the edge n-mers have been used.

We make the convention that the first time we visit a vertex of the form $\{X^{n-1}, (X^c)^{n-1}\}$, we insert the appropriate standard loop $\{X^n, (X^c)^n\}$. Also, in constructing the paths, we organize the two strings in each n-mer in each path so that we use the first string in each n-mer, so as to more easily keep track of the fact that we are constructing an admissible path.

The properties of $\mathcal{G}_n$ as described in Proposition 2, namely the pairing of the edge n-mers at each vertex $(n-1)$-mer and the fact that the valency of every vertex $(n-1)$-mer is even, imply that there is a maximal collection of circular paths, as constructed in this Step, so that each edge n-mer of $\mathcal{G}_n$ occurs once and only once in exactly one of the paths.

Consider the case $n = 3$. Begin at the vertex 2-mer $\{AC, GT\}$ and the incident edge 3-mer $\{ACC, GGT\}$. Note that there is only one way to choose a string from the vertex 2-mer $\{AC, GT\}$ and a string from the edge 3-mer $\{ACC, GGT\}$ so that a string from the vertex 2-mer is the initial substring of the string from the edge 3-mer. Namely, we choose the string AC from $\{AC, GT\}$ and the string ACC from $\{ACC, GGT\}$.

Construct a path beginning with $\{ACC, GGT\}$, using the pairing as decided in Step 1. Since we are forced to choose the string ACC in the edge 3-mer $\{ACC, GGT\}$, we come to the vertex 2-mer $\{CC, GG\}$. Since this is our first visit to the vertex 2-mer $\{CC, GG\}$, we insert the standard loop $\{CCC, GGG\}$, according to the convention described above. At the vertex $\{CC, GG\}$, the edge 3-mer $\{ACC, GGT\}$ is paired with the edge 3-mer $\{CCA, TGG\}$, and so after the standard loop $\{CCC, GGG\}$, we insert the edge 3-mer $\{CCA, TGG\}$. This brings us to the vertex 2-mer $\{CA, TG\}$. At the vertex 2-mer $\{CA, TG\}$, the edge 3-mer $\{CCA, TGG\}$ is paired with the edge 3-mer $\{CAG, CTG\}$. Continuing yields the following circular path:

$$\begin{aligned}\mathcal{P}_1 = \{\ &\{ACC, GGT\}, \{CCC, GGG\}, \{CCA, TGG\}, \{CAG, CTG\},\\ &\{AGC, GCT\}, \{GCA, TGC\}, \{CAT, ATG\}, \{ATT, AAT\},\\ &\{TTT, AAA\}, \{TTC, GAA\}, \{TCA, TGA\}, \{CAA, TTG\},\\ &\{AAG, CTT\}, \{AGT, ACT\}, \{GTG, CAC\}, \{TGT, ACA\},\end{aligned}$$

$$\{GTC, GAC\}, \{TCC, GGA\}, \{CCG, CGG\}, \{CGT, ACG\},$$
$$\{GTT, AAC\}, \{TTA, TAA\}, \{TAG, CTA\}, \{AGA, TCT\},$$
$$\{GAT, ATC\}, \{ATA, TAT\}, \{TAC, GTA\}\}.$$

The path $\mathcal{P}_1$ is a circular path beginning and ending at the vertex 2-mer $\{AC, GT\}$. At this point, we can go no further, as the pairing at the vertex 2-mer pairs the edge 3-mer $\{TAC, GTA\}$ with the edge 3-mer $\{ACC, GGT\}$ that begins the path $\mathcal{P}_1$. Also, there are no more unused edge 3-mers at the vertex 2-mer $\{AC, GT\}$. However, not all of the edge 3-mers in $\mathcal{G}_3$ have been used. So, we choose another vertex 2-mer, say $\{GA, TC\}$, and an incident edge 3-mer that has not been used, say $\{GAG, CTC\}$, and construct a second circular path:

$$\mathcal{P}_2 = \{\{GAG, CTC\}, \{AGG, CCT\}, \{GGC, GCC\},$$
$$\{GCG, CGC\}, \{CGA, TCG\}\}.$$

At this point, we have used all of the edge 3-mers of $\mathcal{G}_3$, and we have used each edge 3-mer exactly once.

Step 3: plumb the maximal circular paths together to get an Eulerian path in $\mathcal{G}_n$:

Let $\mathcal{P}_1, \ldots, \mathcal{P}_p$ be the collection of maximal circular paths that come from the construction in Step 2. Since the graph $\mathcal{G}_n$ is connected, there exists a pair of the paths $\mathcal{P}_1, \ldots, \mathcal{P}_p$ that pass through the same vertex $(n-1)$-mer. For the sake of notational convenience, reindex the list of paths so that $\mathcal{P}_1$ and $\mathcal{P}_2$ pass through the same vertex $(n-1)$-mer $\{V, V^c\}$.

Cyclically reorder the edge n-mers in the paths $\mathcal{P}_1$ and $\mathcal{P}_2$ so that they both begin (and hence end) at the common vertex $(n-1)$-mer $\{V, V^c\}$ that they both pass through. (This is not an essential change to the path; the cyclically reordered path contains the same edge n-mers in the same cyclic order.) (If there are several vertex $(n-1)$-mers of $\mathcal{G}_n$ that both $\mathcal{P}_1$ and $\mathcal{P}_2$ pass through, choose one.) There is one subtlety here, which involves the strings used in the n-mers when the paths pass through the vertex $(n-1)$-mer $\{V, V^c\}$. If $\mathcal{P}_1$ and $\mathcal{P}_2$ pass through $\{V, V^c\}$ using the same string, then plumb them together to form a path $\mathcal{Q}_1$ by first following $\mathcal{P}_1$ and then following $\mathcal{P}_2$.

If $\mathcal{P}_1$ and $\mathcal{P}_2$ pass through $\{V, V^c\}$ using complementary strings (so that one uses V and the other uses V^c), then first take the complement $\mathcal{P}_2^c$ of $\mathcal{P}_2$, and then plumb $\mathcal{P}_1$ and $\mathcal{P}_2^c$ together to form the path $\mathcal{Q}_1$ by first following $\mathcal{P}_1$ and then following $\mathcal{P}_2^c$. By the *complement of the path* $\mathcal{P} = \{e_1, \ldots, e_n\}$, we mean the path $\mathcal{P}^c = \{e_n^c, \ldots, e_1^c\}$ formed by reversing the order of the n-mers occurring in $\mathcal{P}$. For each edge n-mer e_j appearing in $\mathcal{P}$, we also reverse the order of the two strings comprising the n-mer e_j, which is the meaning of the notation e_j^c; this is necessary, as we have adopted the convention that we use the first string in each n-mer, and so the order in which we write the strings in each n-mer is important. The complement $\mathcal{P}^c$ is the path obtained

from $\mathcal{P}$ by walking along $\mathcal{P}$ backwards, and the string corresponding to $\mathcal{P}^c$ is the complement of the string corresponding to $\mathcal{P}$.

We continue. One of the remaining paths $\mathcal{P}_3, \ldots, \mathcal{P}_p$ must share a vertex $(n-1)$-mer with $\mathcal{Q}_1$; again for notational convenience, reindex the list so that $\mathcal{P}_3$ shares a vertex $(n-1)$-mer with $\mathcal{Q}_1$. Plumb together $\mathcal{Q}_1$ and $\mathcal{P}_3$, taking the complement of $\mathcal{P}_3$ if necessary, to get a new path $\mathcal{Q}_2$ that passes along the same edges of $\mathcal{G}_n$ as $\mathcal{P}_3$ and $\mathcal{Q}_1$. Continue this process until all the paths $\mathcal{P}_1, \ldots, \mathcal{P}_p$ have been plumbed together to obtain a single circular path $\mathcal{Q}$. Since the original paths together pass along every edge of $\mathcal{G}_n$ once and only once, the path $\mathcal{Q}$ we construct by plumbing them together also passes along each edge of $\mathcal{G}_n$ once and only once. Hence, $\mathcal{Q}$ is an admissible Eulerian path in $\mathcal{G}_n$, and so the string $S_{\mathcal{Q}}$ arising from $\mathcal{Q}$ is a minimal universal footprinting template for n-mers.

Consider the case $n = 3$. In order to plumb together these two circular paths, we first need to find a vertex 2-mer through which both of these circular paths pass. For this particular example, both circular paths pass through the vertex 2-mer $\{CC, GG\}$. We first need to write both paths to begin and end at this vertex 2-mer.

In fact, $\{CC, GG\}$ is the second vertex 2-mer that $\mathcal{P}_1$ passes through, and so we can rewrite $\mathcal{P}_1$ as

$$\begin{aligned}\mathcal{P}_1 = \{\ &\{CCA, TGG\}, \{CAG, CTG\}, \{AGC, GCT\}, \{GCA, TGC\},\\ &\{CAT, ATG\}, \{ATT, AAT\}, \{TTT, AAA\}, \{TTC, GAA\},\\ &\{TCA, TGA\}, \{CAA, TTG\}, \{AAG, CTT\}, \{AGT, ACT\},\\ &\{GTG, CAC\}, \{TGT, ACA\}, \{GTC, GAC\}, \{TCC, GGA\},\\ &\{CCG, CGG\}, \{CGT, ACG\}, \{GTT, AAC\}, \{TTA, TAA\},\\ &\{TAG, CTA\}, \{AGA, TCT\}, \{GAT, ATC\}, \{ATA, TAT\},\\ &\{TAC, GTA\}, \{ACC, GGT\}, \{CCC, GGG\}\}.\end{aligned}$$

(Since both paths named $\mathcal{P}_1$ pass along the same edge 3-mers in the same order, we use the same name $\mathcal{P}_1$ for both.)

The other path $\mathcal{P}_2$ also passes through $\{CC, GG\}$, and so we can rewrite $\mathcal{P}_2$ as

$$\begin{aligned}\mathcal{P}_2 = \{&\{GGC, GCC\}, \{GCG, CGC\}, \{CGA, TCG\},\\ &\{GAG, CTC\}, \{AGG, CCT\}\}.\end{aligned}$$

Note though that $\mathcal{P}_1$ passes through the vertex 2-mer $\{CC, GG\}$ using the string CC, while $\mathcal{P}_2$ passes through the vertex 2-mer $\{CC, GG\}$ using the string GG. Hence, we first need to take the complement $\mathcal{P}_2^c$ of $\mathcal{P}_2$, as described above:

$$\begin{aligned}\mathcal{P}_2^c = \{&\{CCT, AGG\}, \{CTC, GAG\}, \{TCG, CGA\},\\ &\{CGC, GCG\}, \{GCC, GGC\}\}.\end{aligned}$$

The complemented path $\mathcal{P}_2^c$ now passes through the vertex $\{CC, GG\}$ using the string CC, as needed.

To plumb the two paths together, we follow the edges in the first path $\mathcal{P}_1$, and then follow the edges in the second path $\mathcal{P}_2^c$. The resulting path $\mathcal{P}$ starts and ends at the common vertex 2-mer and passes along every edge of $\mathcal{G}_3$ once and only once.

$$\begin{aligned}\mathcal{P} = \{\ &\{CCA, TGG\}, \{CAG, CTG\}, \{AGC, GCT\}, \{GCA, TGC\},\\ &\{CAT, ATG\}, \{ATT, AAT\}, \{TTT, AAA\}, \{TTC, GAA\},\\ &\{TCA, TGA\}, \{CAA, TTG\}, \{AAG, CTT\}, \{AGT, ACT\},\\ &\{GTG, CAC\}, \{TGT, ACA\}, \{GTC, GAC\}, \{TCC, GGA\},\\ &\{CCG, CGG\}, \{CGT, ACG\}, \{GTT, AAC\}, \{TTA, TAA\},\\ &\{TAG, CTA\}, \{AGA, TCT\}, \{GAT, ATC\}, \{ATA, TAT\},\\ &\{TAC, GTA\}, \{ACC, GGT\}, \{CCC, GGG\}, \{CCT, AGG\},\\ &\{CTC, GAG\}, \{TCG, CGA\}, \{CGC, GCG\}, \{GCC, GGC\}\}.\end{aligned}$$

This path $\mathcal{P}$ is an admissible Eulerian path in $\mathcal{G}_3$.

Step 4: read the minimal n-mer string from the Eulerian path:

Once we have constructed the Eulerian path $\mathcal{P}$ in $\mathcal{G}_n$ by plumbing together, as described in Step 3, all of the circular paths constructed in Step 2, we can construct the corresponding string $S_{\mathcal{P}}$, using the first string in every n-mer in $\mathcal{P}$. Since $\mathcal{P}$ is an Eulerian path in $\mathcal{G}_3$, the string $S_{\mathcal{P}}$ has length naive(n) and so is a minimal universal footprinting template for n-mers. This completes the proof of Theorem 1.

Consider the case $n = 3$. Since the path $\mathcal{P}$ constructed in Step 3 passes along each edge of $\mathcal{G}_3$ once and only once, we can read from this a minimal universal footprinting template for 3-mers by taking the first string in each edge 3-mer appearing in $\mathcal{P}$. The path $\mathcal{P}$ constructed in Step 3 gives rise to the string

$$S_{\mathcal{P}} = CCAGCATTTCAAGTGTCCGTTAGATACCCTCGCC.$$

This string of length 34 is a minimal universal footprinting template for 3-mers.

6. The algorithm for n even

The algorithm for the construction of a universal footprinting template for n-mers in the case of n even uses the same steps as the algorithm for the case of n odd, as discussed in Section 5. The main difference is that, as noted in Proposition 3, for $n \geq 4$ there are multiple vertex $(n-1)$-mers in $\mathcal{G}_n$ of odd valency, and so it is no longer possible to construct an Eulerian path in $\mathcal{G}_n$. Therefore, we augment the algorithm as discussed for n odd by inserting a Step 0 at the beginning: we first construct some short circular paths in $\mathcal{G}_n$ that cover some edge n-mers in $\mathcal{G}_n$ twice. After these multiply-used edge n-mers are removed from consideration, the vertex $(n-1)$-mers will all have even valency and the remaining edge n-mers at each vertex $(n-1)$-mer can

be admissibly paired, and so an Eulerian path using these remaining edge n-mers can be constructed using the algorithm as described in Section 5. The universal footprinting template for n-mers is then constructed using the same plumbing operation as discussed in Section 5, Step 3.

The case $n = 2$ behaves differently from the general case of n even in that the graph $\mathcal{G}_2$ does not contain any vertices of odd valency, though it does contain lollipop loops. Together with the fact that the non-loop edge 2-mers at each vertex 1-mer in $\mathcal{G}_2$ can easily be seen to be admissibly paired (as discussed at the end of Section 4), we see that the case of $n = 2$ is distinctly simpler than the general case of n even. We discuss the case of $n = 2$ at the end of this Section.

We use Proposition 3. The main technical difficulty arises at the vertex $(n-1)$-mers incident to self-complementary n-mers. There are two vertex $(n-1)$-mers incident to two self-complementary edge n-mers (lollipop loops), and there are $4^{n/2} - 4$ vertex $(n-1)$-mers incident to a single self-complementary edge n-mer. The main technical issue here is that passing along a lollipop loop n-mer changes the string used in the corresponding vertex $(n-1)$-mer. We first deal with this issue at the vertex $(n-1)$-mers incident to two lollipop loops, and then consider the vertex $(n-1)$-mers of odd valency, which are the vertex $(n-1)$-mers that are incident to exactly one lollipop loop.

Let $m = \frac{1}{2}(n-2)$. There are two vertex $(n-1)$-mers that are each incident to two lollipop loops: the vertex $(n-1)$-mer $\{A(TA)^m, (TA)^m T\}$ is incident to the lollipop loops $\{A(TA)^m T\}$ and $\{TA(TA)^m\}$, while the vertex $(n-1)$-mer $\{C(GC)^m, (GC)^m G\}$ is incident to the lollipop loops $\{C(GC)^m G\}$ and $\{GC(GC)^m\}$. We impose the convention, as in the case $n = 4$, that the first time we visit one of these vertex $(n-1)$-mers during the construction of a path, we follow one lollipop loop and then immediately follow the other before leaving the vertex $(n-1)$-mer. This yields two short circular paths $\{\{A(TA)^m T\}, \{TA(TA)^m\}\}$ and $\{\{C(GC)^m G\}, \{GC(GC)^m\}\}$ of length 2. Note that each of these two paths uses the same string of length $n-1$ reentering the vertex $(n-1)$-mer as it uses leaving it; for example, the first path yields the string $A(TA)^m TA$, which begins and ends with the string $A(TA)^m$ of length $n-1$.

There are $4^{n/2} - 4$ vertex $(n-1)$-mers of odd valency in $\mathcal{G}_n$, each of which is incident to exactly one lollipop loop. We now pair lollipop loops as follows: let $p = \frac{1}{2}n$. We construct a circular path joining the lollipop loop

$$W = X_1 \dots X_p X_p^c \dots X_1^c,$$

which is incident to the vertex $(n-1)$-mer

$$\{X_1 \dots X_p X_p^c \dots X_2^c, X_2 \dots X_p X_p^c \dots X_1^c\},$$

with the lollipop loop

$$W' = X_p^c \dots X_1^c X_1 \dots X_p,$$

which is incident to the vertex $(n-1)$-mer

$$\{X_p^c \dots X_1^c X_1 \dots X_{p-1}, X_{p-1}^c \dots X_1^c X_1 \dots X_p\}.$$

The path comes from cyclically permuting the letters, as follows:

$$\begin{aligned}
&W = X_1 \dots X_p X_p^c \dots X_1^c \\
&W_1 = X_2 \dots X_p X_p^c \dots X_1^c X_1 \\
&W_2 = X_3 \dots X_p X_p^c \dots X_1^c X_1 X_2 \\
&\dots \\
&W_{p-1} = X_p X_p^c \dots X_1^c X_1 X_2 \dots X_{p-1} \\
&W' = X_p^c \dots X_1^c X_1 \dots X_p \\
&W_{p-1}^c = X_{p-1}^c \dots X_1^c X_1 \dots X_p X_p^c \\
&\dots \\
&W_1^c = X_1^c X_1 \dots X_p X_p^c \dots X_2^c
\end{aligned}$$

This path $\{\{W, W^c\}, \{W_1, W_1^c\}, \dots, \{W', (W')^c\}, \dots, \{W_1^c, W_1\}\}$ passes along each of the non-loop edge n-mers $\{W_1, W_1^c\}, \dots, \{W_{p-1}, W_{p-1}^c\}$ it contains twice, and over both lollipop loops $\{W\}$, $\{W'\}$ once. As with the short circular paths constructed from pairs of lollipop loops incident to the same vertex $(n-1)$-mer, this short circular path uses the same string of length $n-1$ reentering the vertex $(n-1)$-mer $\{X_1 \dots X_p X_p^c \dots X_2^c, X_2 \dots X_p X_p^c \dots X_1^c\}$ as it uses leaving it, namely $X_1 \dots X_p X_p^c \dots X_2^c$.

Now, we count. There are $p - 1 = \frac{1}{2}(n-2)$ extra edge n-mers used in the construction of this short circular path, coming from the second copy of the path from one of the vertex $(n-1)$-mers to the other. Hence, if we add $\frac{1}{2}(n-2)$ edge n-mers for each of the $\frac{1}{2}(4^{n/2} - 4)$ pairs of lollipop loops (paired as above) then we can use the algorithm already described to construct a circular path in $\mathcal{G}_n$ that passes along each edge n-mer at least once. This path will have length $\frac{1}{2}(4^n + 4^{n/2}) + n - 1 + \frac{1}{4}(4^{n/2} - 4)(n-2)$. We can begin this circular path at one of the duplicated edge n-mers and then remove the beginning edge n-mer of the path, to obtain a non-circular path of length $\max(n) = \frac{1}{2}(4^n + 4^{n/2}) + n - 2 + \frac{1}{4}(4^{n/2} - 4)(n-2)$. This simplifies to $\max(n) = 2^{2n-1} + n\,2^{n-2}$ and is in most cases an overestimate.

To get the lower bound $\min(n)$ of the length of a minimal universal footprinting template for n-mers, note that we need to add at least one edge n-mer for each pair of lollipop loops. In this case, we get a circular path of length $\frac{1}{2}(4^n + 4^{n/2}) + n - 1 + \frac{1}{2}(4^{n/2} - 4) = \frac{1}{2}4^n + 4^{n/2} + n - 3$. As before, we can begin this circular path at one of the duplicated edge n-mers and then remove the beginning edge n-mer of the path, to obtain a non-circular path of length $\min(n) = \frac{1}{2}4^n + 4^{n/2} + n - 4$. This is in most cases an underestimate, though it is sharp in the case of $n = 4$. We can achieve this lower bound precisely when the vertices incident to a single lollipop loop occur as incident vertices in $\mathcal{G}_n$, and so each pair are separated by a single edge; in this case, the short circular paths have length 4, comprising

of the two lollipop loops and the single edge joining them included twice. This occurs in the case $n = 4$. This completes the proof of Theorem 2.

Consider the case of $n = 4$, and construct the graph $\mathcal{G}_4$ as before. There are $4^2 = 16$ self-complementary strings of length 4, namely $ATAT$, $TATA$, $CGCG$, $GCGC$, $AATT$, $TTAA$, $CCGG$, $GGCC$, $ACGT$, $TCGA$, $AGCT$, $TGCA$, $CATG$, $GATC$, $CTAG$, and $GTAC$. Each of these yields a lollipop loop at a vertex 3-mer, as follows:

edge 4-mer	vertex 3-mer	edge 4-mer	vertex 3-mer
$ATAT$	$\{ATA, TAT\}$	$TATA$	$\{ATA, TAT\}$
$CGCG$	$\{CGC, GCG\}$	$GCGC$	$\{CGC, GCG\}$
$AATT$	$\{AAT, ATT\}$	$TTAA$	$\{TTA, TAA\}$
$CCGG$	$\{CCG, CGG\}$	$GGCC$	$\{GGC, GCC\}$
$ACGT$	$\{ACG, CGT\}$	$TCGA$	$\{TCG, CGA\}$
$AGCT$	$\{AGC, GCT\}$	$TGCA$	$\{TGC, GCA\}$
$CATG$	$\{CAT, ATG\}$	$GATC$	$\{GAT, ATC\}$
$CTAG$	$\{CTA, TAG\}$	$GTAC$	$\{GTA, TAC\}$

Consider the two lollipop loops $\{ATAT\}$ and $\{TATA\}$ incident to the vertex 3-mer $\{ATA, TAT\}$. (A similar argument holds for the two lollipop loops $\{CGCG\}$ and $\{GCGC\}$ incident to the vertex 3-mer $\{CGC, GCG\}$.) These two lollipop loops yield the path $\{\{ATAT\}, \{TATA\}\}$.

Other than the two lollipop loops, there are six edge 4-mers incident to $\{ATA, TAT\}$, namely $\{AATA, TATT\}$, $\{CATA, TATG\}$, $\{GATA, TATC\}$, $\{ATAA, TTAT\}$, $\{ATAC, GTAT\}$, $\{ATAG, CTAT\}$. These six edges can be admissibly paired, as there are three of the form $\{ZATA, TATZ^c\}$ and three of the form $\{ATAZ, Z^cTAT\}$, where Z ranges over the letters A, C, G.

Recall that by Proposition 3, there are 12 vertices of odd valency in $\mathcal{G}_4$, namely the vertex 3-mers that are incident to a single lollipop loop, and there are six edges that pair these twelve vertices, as follows:

edge 4-mer	vertex 3-mers joined by edge 4-mer
$\{TAAT, ATTA\}$	$\{AAT, ATT\}$, $\{TAA, TTA\}$
$\{TACG, CGTA\}$	$\{ACG, CGT\}$, $\{TAC, GTA\}$
$\{TAGC, GCTA\}$	$\{AGC, GCT\}$, $\{TAG, CTA\}$
$\{GCAT, ATGC\}$	$\{CAT, ATG\}$, $\{GCA, TGC\}$
$\{GCCG, CGGC\}$	$\{CCG, CGG\}$, $\{GCC, GGC\}$
$\{ATCG, CGAT\}$	$\{TCG, CGA\}$, $\{ATC, GAT\}$

We work with the vertex 3-mers $\{AAT, ATT\}$ and $\{TAA, TTA\}$; the other five pairs are handled similarly.

Construct a short circular path as follows: start at one of the two vertex 3-mers, say $\{ATT, AAT\}$, pass along the edge 4-mer $\{ATTA, TAAT\}$ that joins $\{ATT, AAT\}$ to $\{TAA, TTA\}$, then along the lollipop loop $\{TTAA\}$ at $\{TAA, TTA\}$, then back along the edge 4-mer $\{TAAT, ATTA\}$ from $\{TAA, TTA\}$ to $\{ATT, AAT\}$, and then along the lollipop loop $\{AATT\}$ at $\{ATT, AAT\}$. This yields the short circular path

$$\{\{ATTA, TAAT\}, \{TTAA\}, \{TAAT, ATTA\}, \{AATT\}\}.$$

We do the same for the other five pairs of vertex 3-mers of odd valency: corresponding to each pair, we construct a short circular path that passes twice along the edge 4-mer joining them, as well as once along the lollipop loops at each of the two vertex 3-mers in the pair.

To complete the construction, note that at the remaining vertex 3-mers, namely those that are not incident to a self-complementary edge 4-mer, the edge 4-mers can be paired, as in the case of n odd. At the edge 4-mers incident to a self-complementary edge 4-mer, there is the lollipop loop (or loops, for two of these vertex 4-mers), the edge n-mer that is to be passed along twice, and the remaining six edges, which can be paired as in the case of n odd. So, as before, we can admissibly pair these edge 4-mers at each vertex 3-mer, and run the algorithm as described in the case of n odd.

Begin with the two short circular paths of length 2 that comes from the pairs of lollipop loops that are incident to the same vertex 3-mer and the six circular paths of length 4 just constructed, and mark the edge 4-mers that occur in these short circular paths as used. We then proceed with the construction as before. Since the remaining edge 4-mers can be admissibly paired, we obtain some number of circular paths that each of the remaining of edge 4-mers of $\mathcal{G}_4$ once and only once. We can then plumb these circular paths together with the eight short circular paths constructed above. This yields an admissible circular path in $\mathcal{G}_4$ that covers every edge once, and covers six edges twice. (The short circular paths of length 2 coming from the pairs of lollipop loops incident to the same vertex 3-mer do not entail passing along an edge 4-mer more than once.) From this circular path, we obtain a string S of length 145 with the property that for most 4-mers, exactly one of the strings from each 4-mer is contained exactly once; however, the six 4-mers coming from the six twice-used edge 4-mers all have two representatives in S. This is why S is longer than the hoped for length, $\text{naive}(4) = 139$.

By rewriting this circular path to begin at one with one of the edge 4-mers used twice and then deleting this edge 4-mer from the path, we convert the circular path describing S to a non-circular path $\mathcal{P}$ describing a string of length 144 which still contains every 4-mer at least once, and which contains five of them twice. To show that this is a minimal universal footprinting template for 4-mers, we need to know that it has minimal length over all universal footprinting template for 4-mers.

So, suppose T is any universal footprinting template for 4-mers. This string gives rise to a path $\mathcal{P}_T$ in the graph $\mathcal{G}_4$, which may or may not give rise to a circular path. Since this string is a universal footprinting template for 4-mers, it must pass along every edge 4-mer of $\mathcal{G}_4$ at least once. We have already seen that it must pass along some edge 4-mers twice. Since $\mathcal{G}_4$ has 136 edges, a string covering each edge 4-mer at least once and which covers k edge 4-mers twice has length $139 + k$. (We do not worry about the case that some edge 4-mers are covered three or more times, since the argument given below would then show that the string has length greater than 144.)

There are twelve vertex 3-mers of odd valency in $\mathcal{G}_4$. Even if we allow for the non-circular path coming from T to begin and to end at a vertex

3-mer of odd valency, this still leaves ten vertex 3-mers of odd valency. Since we are constructing a path that arises from a string, this implies that at a minimum, if these ten vertex 3-mers are joined by edge 4-mers in pairs and if we pass along each of these edge 4-mers twice (as we do in the construction given above), this leads to $k = 5$ edge 4-mers being traversed twice, which then implies that the string has length 144. If other edge 4-mers are covered twice, then the number of edge 4-mers covered twice is greater than 5, since at each of the ten vertex 3-mers of odd valency not at the beginning or end of the path, we need to have at least one edge 4-mer covered twice. So, the construction given above results in a string of minimal length, and so any minimal universal footprinting template for 4-mers has length 144.

For the case $n = 2$, since there are no vertices of $\mathcal{G}_2$ that are incident to only a single lollipop loop, we need only the first part of Step 0, at which we construct short circular loops that arise from traversing the two lollipop loops incident to each vertex. As we saw at the end of Section 4, the non-loop edges of $\mathcal{G}_2$ can be admissibly paired, and therefore we can run the algorithm. Since there are no vertices of odd valency, it is possible (as in the case of n odd) to construct an Eulerian path in $\mathcal{G}_2$, and hence a minimal universal footprinting template of length naive(2) = 11, as we saw in Section 1.

7. Analysis of the vertices of $\mathcal{G}_n$ for n odd (and $n \geq 3$)

In this Section, we prove Proposition 2, which gives a detailed description of the structure of the graph $\mathcal{G}_n$ at its vertices, for n odd.

Proposition 2 *Suppose that n is odd and $n \geq 3$. There are no lollipop loops in $\mathcal{G}_n$.*

- *Suppose the vertex $(n-1)$-mer $\{V, V^c\}$ is self-complementary. Then, its valency is 4. There is no standard loop incident to $\{V\}$. The edges incident to $\{V\}$ are the n-mers $\{AV, VT\}$, $\{CV, VG\}$, $\{GV, VC\}$, $\{TV, VA\}$ and these may be admissably paired.*
- *Suppose the vertex $(n-1)$-mer $\{V, V^c\}$ is not self-complementary. Then, its valency is 8.*
  - *If $V = X^{n-1}$ for some letter X, then the edges incident to $\{V, V^c\}$ are the n-mer $\{X^n, (X^c)^n\}$, which is a standard loop, and the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X. The number of n-mers of the form $\{ZV, V^cZ^c\}$ is equal to the number of n-mers of the form $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X, and these may be admissably paired.*
  - *If $V \neq X^{n-1}$ for any letter X, then there is no standard loop incident to $\{V, V^c\}$. The edges incident to $\{V, V^c\}$ are the n-mers $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, $\{TV, V^cA\}$, $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$. The number of n-mers of the form $\{ZV, V^cZ^c\}$ is equal to the number of n-mers of the form $\{VZ, Z^cV^c\}$, where Z*

ranges over the letters A, C, G, T, and these may be admissably paired.

In particular, the edges incident to each vertex of $\mathcal{G}_n$ which are not standard loops may be admissably paired at that vertex since at each vertex $(n-1)$-mer $\{V, V^c\}$, the number of non-loop edge n-mers ending with the string V (respectively V^c) is equal to the number of edge n-mers beginning with the string V (respectively V^c).

Proof. We prove Proposition 2 by examining each type of vertex $(n-1)$-mer in turn. A vertex of $\mathcal{G}_n$ is an $(n-1)$-mer $\{V, V^c\}$, where V is a string of even length. Since the strings in the edge n-mers have odd length, namely n, no string contained in an edge n-mer can be self-complementary, and so there are no lollipop loops. There are two cases to consider, namely that V is self-complementary and that V is not self-complementary.

Suppose that V is self-complementary, so that $V = V^c$. Since V is self-complementary, it cannot be equal to X^{n-1} for any letter X. The edges incident to the vertex $\{V\}$ are the n-mers $\{AV, VT\}$, $\{CV, VG\}$, $\{GV, VC\}$, $\{TV, VA\}$ (obtained by adding a letter to the beginning of V), and $\{VA, TV\}$, $\{VC, GV\}$, $\{VG, CV\}$, $\{VT, AV\}$ (obtained by adding a letter to the end of V). Note that each of the four n-mers obtained by adding a letter to the end of V is equal to one of the n-mers obtained by adding a latter to the beginning of V. So, there are at most four edges incident to the vertex $\{V\}$, namely $\{AV, VT\}$, $\{CV, VG\}$, $\{GV, VC\}$, $\{TV, VA\}$.

It remains to show that the four n-mers obtained by adding a letter to the beginning of V are distinct. We show that $\{AV, VT\}$ is distinct from all of $\{CV, VG\}$, $\{GV, VC\}$, and $\{TV, VA\}$; the other cases are handled similarly. If $\{AV, VT\} = \{CV, VG\}$, then either $AV = CV$ or $AV = VG$. The former case cannot occur, since the first letters of the two strings are different. The latter case cannot occur, as can be seen by counting the number of occurances of the letter A at the beginning of the two strings (and by recalling that, since V is not a power of a single letter, there exists a first letter in V that is not A): that is, if V has an initial string of the form A^k but not of the form A^{k+1} then AV has an initial string of the form A^{k+1} at its beginning, while VG does not, and so they cannot be equal. A similar argument shows that $\{AV, VT\}$ is distinct from $\{GV, VC\}$ and $\{TV, VA\}$.

Since V is not equal to X^{n-1} for any letter X, there is no standard loop incident to $\{V\}$. Hence, there are four edges incident to $\{V\}$, namely the four n-mers $\{AV, VT\}$, $\{CV, VG\}$, $\{GV, VC\}$, and $\{TV, VA\}$. As there are no loops incident to $\{V\}$, each edge contributes 1 to the valency of $\{V\}$. So, the valency of $\mathcal{G}_n$ at $\{V\}$ is 4.

Suppose that V is not self-complementary, so that $V \neq V^c$. The edges incident to the vertex $\{V, V^c\}$ are $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, $\{TV, V^cA\}$, which are obtained by adding a letter to the beginning of V, and $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$, which are obtained by adding a letter to the end of V.

The four n-mers $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, $\{TV, V^cA\}$ obtained by adding a letter to the beginning of V are distinct. We show that $\{AV, V^cT\}$ is distinct from all of $\{CV, V^cG\}$, $\{GV, V^cC\}$, and $\{TV, V^cA\}$; the other cases are handled similarly. If $\{AV, V^cT\} = \{CV, V^cG\}$, then either $AV = CV$ or $AV = V^cG$. The former case cannot occur, since the first letters of the two strings are different. The latter case cannot occur, since otherwise V would end in G, and so V^c and hence V^cG would begin in C; the two strings AV and V^cG would then begin with different letters. A similar argument shows that $\{AV, V^cT\}$ is distinct from $\{GV, V^cC\}$ and $\{TV, V^cA\}$. A similar argument shows that the four n-mers $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$ obtained by adding a letter to the end of V are distinct.

Suppose that $V = X^{n-1}$ for some letter X. Since $V = X^{n-1}$, we have that $\{XV, V^cX^c\} = \{VX, X^cV^c\}$, since both are equal to $\{X^n, (X^c)^n\}$. We need to show that the other six edges incident to $\{X^{n-1}, (X^c)^{n-1}\}$, namely the six n-mers $\{ZX^{n-1}, (X^c)^{n-1}Z^c\}$ and $\{X^{n-1}Z, Z^c(X^c)^{n-1}\}$, where Z ranges over the letters not equal to X, are all distinct, and are all distinct from $\{X^n, (X^c)^n\}$. From the arguments already given, we know that the three n-mers $\{ZX^{n-1}, (X^c)^{n-1}Z^c\}$ are distinct, as are the three n-mers $\{X^{n-1}Z, Z^c(X^c)^{n-1}\}$. It is easy to see that these six n-mers are all distinct from $\{X^n, (X^c)^n\}$, by counting the number of occurrences of the letters X or X^c in each string in each of the n-mers.

Hence, it only remains to show that each of the $\{ZX^{n-1}, (X^c)^{n-1}Z^c\}$ is distinct from each of the three n-mers $\{X^{n-1}Z, Z^c(X^c)^{n-1}\}$, where Z ranges over the letters not equal to X. So, suppose there are letters Z and Y, both not equal to X, so that $\{ZX^{n-1}, (X^c)^{n-1}Z^c\} = \{X^{n-1}Y, Y^c(X^c)^{n-1}\}$. Then, either $ZX^{n-1} = X^{n-1}Y$ or $ZX^{n-1} = Y^c(X^c)^{n-1}$. The former case cannot occur, since the two strings begin with different letters (as $Z \neq X$). The latter case cannot occur, since the two strings end in different letters (as $X \neq X^c$).

So, if $V = X^{n-1}$ for some letter X, then there is one standard loop $\{X^n, (X^c)^n\}$ incident to $\{V, V^c\}$, and there are six edges incident to $\{V, V^c\}$, namely the n-mers $\{ZX^{n-1}, (X^c)^{n-1}Z^c\}$ and $\{X^{n-1}Z, Z^c(X^c)^{n-1}\}$, where Z ranges over the three choices of the letters A, C, G, T not equal to X. The standard loop contributes 2 to the valency of $\{V, V^c\}$ and each of the other six edges contributes 1. So, the valency of $\mathcal{G}_n$ at $\{V, V^c\}$ is 8.

Suppose now that $V \neq X^{n-1}$ for any letter X. In this case, we show that the eight edge n-mers $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, $\{TV, V^cA\}$, $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$ are distinct. From the arguments already given, we know that $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, and $\{TV, V^cA\}$ are distinct, as are $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, and $\{VT, AV^c\}$. We give the details to show that $\{AV, V^cT\}$ is distinct from each of $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$; the other cases are handled similarly. If $\{AV, V^cT\} = \{VA, TV^c\}$, then either $AV = VA$ or $AV = TV^c$. The former case cannot occur, by counting the number of

occurances of the letter A at the beginning of the two strings (and using that $V \neq A^{n-1}$, so there must be some letter in V other than A). The latter case cannot occur, since the first letters of the two strings are different. The cases of comparing $\{AV, V^cT\}$ to $\{VC, GV^c\}$ and $\{VG, CV^c\}$ are handled similarly. To see that $\{AV, V^cT\}$ and $\{VT, AV^c\}$ are distinct, note that $AV \neq VT$, again by counting the number of occurances of the letter A at the beginning of the two strings, and that $AV \neq AV^c$, as otherwise V would equal V^c (by deleting the A from the beginning of each of the two strings), and we are working in the case that $V \neq V^c$.

So, if $V \neq X^{n-1}$ for any letter X, then there is no standard loop at $\{V, V^c\}$. There are eight edges incident to $\{V, V^c\}$, namely the n-mers $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, $\{TV, V^cA\}$, $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$, and each edge contributes 1 to the valency of $\{V, V^c\}$. Hence, the valency of $\mathcal{G}_n$ at $\{V, V^c\}$ is 8.

This completes the analysis of the structure and valencies of the vertices of $\mathcal{G}_n$ in the case of n odd.

8. Analysis of the vertices of $\mathcal{G}_n$ for n even (and $n \geq 4$)

In this Section, we prove Proposition 3, which gives a detailed description of the structure of the graph $\mathcal{G}_n$ at its vertices, for n even.

Proposition 3 *Suppose that n is even and $n \geq 4$. Let $m = \frac{1}{2}(n-2)$.*

- *Suppose there is a self-complementary edge incident to the vertex $(n-1)$-mer $\{V, V^c\}$. Then, $\{V, V^c\}$ has the form $\{XUU^c, UU^cX^c\}$, where X is one of the letters A, C, G, T, and U is a string of length m.*
  - *If $U = (X^cX)^{m/2}$, then there are two lollipop loops incident to $\{V, V^c\}$, namely the n-mers $\{X(X^cX)^mX^c\}$ and $\{X^cX(X^cX)^m\}$. There is no standard loop incident to $\{V, V^c\}$. If $V = XUU^c$, the remaining edges incident to $\{V, V^c\}$ are the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X^c. (If instead $V = UU^cX^c$, the remaining edges incident to $\{V, V^c\}$ are the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X.) The number of n-mers of the form $\{ZV, V^cZ^c\}$ is equal to the number of n-mers of the form $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X. The valency of $\mathcal{G}_n$ at $\{V, V^c\}$ is 10. The lollipop loops can be admissably paired with one another and the other edges may be admissably paired.*
  - *If $U \neq (X^cX)^{m/2}$, then there is one lollipop loop incident to $\{V, V^c\}$, namely the n-mer $\{XUU^cX^c\}$. There is no standard loop incident to $\{V, V^c\}$. If $V = XUU^c$, the remaining edges incident to $\{V, V^c\}$ are the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X^c, together with the n-mer $\{X^cV, V^cX\}$. (If instead $V = UU^cX^c$, the remaining edges incident to $\{V, V^c\}$ are the six n-mers $\{ZV, V^cZ^c\}$ and $\{VZ, Z^cV^c\}$, where Z ranges over the letters not equal to X, together with the n-mer $\{XV, V^cX^c\}$.) The*

valency of $\mathcal{G}_n$ *at* $\{V, V^c\}$ *is* 9*. There is no admissable pairing at these vertices.*

- *Suppose there is no self-complementary edge incident to* $\{V, V^c\}$*. Then, there is no lollipop loop incident to* $\{V, V^c\}$*. The valency of* $\mathcal{G}_n$ *at* $\{V, V^c\}$ *is* 8 *and the edges may be admissably paired.*
  - *If* $V = X^{n-1}$*, then the edges incident to* $\{V, V^c\}$ *are the standard loop* $\{X^n, (X^c)^n\}$ *and the six* n*-mers* $\{ZV, V^cZ^c\}$ *and* $\{VZ, Z^cV^c\}$*, where* Z *ranges over the letters not equal to* X*. The number of* n*-mers of the form* $\{ZV, V^cZ^c\}$ *is equal to the number of* n*-mers of the form* $\{VZ, Z^cV^c\}$*, where* Z *ranges over the letters not equal to* X*.*
  - *If* $V \neq X^{n-1}$ *for any letter* X*, then there is no standard loop incident to* $\{V, V^c\}$*. The edges incident to* $\{V, V^c\}$ *are the eight* n*-mers* $\{AV, V^cT\}$*,* $\{CV, V^cG\}$*,* $\{GV, V^cC\}$*,* $\{TV, V^cA\}$*,* $\{VA, TV^c\}$*,* $\{VC, GV^c\}$*,* $\{VG, CV^c\}$*,* $\{VT, AV^c\}$*. The number of* n*-mers of the form* $\{ZV, V^cZ^c\}$ *is equal to the number of* n*-mers of the form* $\{VZ, Z^cV^c\}$*, where* Z *ranges over the letters* A*,* C*,* G*,* T*.*

Proof. We prove Proposition 3 by examining each type of vertex in turn. A vertex of $\mathcal{G}_n$ is an $(n-1)$-mer $\{V, V^c\}$, where V is a string of odd length. Since V has odd length, it cannot be self-complementary, and so $V \neq V^c$. There are two cases to consider, namely that there is a self-complementary edge incident to $\{V, V^c\}$, and that there is no self-complementary edge incident to $\{V, V^c\}$.

We begin with the following observation: the four n-mers obtained by adding a letter to the beginning of V, namely $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, and $\{TV, V^cA\}$, are distinct. We show that $\{AV, V^cT\}$ is distinct from the other three; the remaining cases are handled similarly. If $\{AV, V^cT\} = \{CV, V^cG\}$, then either $AV = CV$ or $AV = V^cG$. The former case cannot occur, since the two strings begin with different letters. The latter case cannot occur, since otherwise V would end in G, and so V^c, and hence V^cG, would begin in C; the two strings AV and V^cG would then begin with different letters. The same argument shows that $\{AV, V^cT\}$ is distinct from $\{GV, V^cC\}$ and $\{TV, V^cA\}$. Similarly, the four n-mers obtained by adding a letter to the end of V, namely $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, and $\{VT, AV^c\}$, are distinct.

Suppose there is a self-complementary n-mer $\{W, W\}$ incident to $\{V, V^c\}$. Suppose that W is obtained by adding the letter X^c to the end of V. Then, $W = VX^c$, where X is one of the letters A, C, G, T. Since $W = W^c$, we see that $VX^c = XV^c$. In particular, this means that V begins in X, and so we can write $V = XP$, where P is a string of length $n-2$. Plugging $V = XP$ back into the equation $W = W^c$, we see that $XPX^c = VX^c = W = W^c = XV^c = XP^cX^c$, and so $P = P^c$. Since P is self-complementary, we can write $P = UU^c$, where U is a string of length $m = \frac{1}{2}(n-2)$. Therefore, $V = XP = XUU^c$. (If W is obtained by adding the letter X to the beginning of V^c, then the argument just given yields that $V = UU^cX^c$.) Therefore, if there is a self-complementary edge incident to $\{V, V^c\}$, then

$\{V, V^c\}$ has the form $\{XUU^c, UU^cX^c\}$, where X is one of the letters A, C, G, T, and U is a string of length m. In particular, note that there cannot be a standard loop incident to $\{V, V^c\}$, since we cannot write $V = X^{n-1}$ for some letter X. (It is at this point that we use that $n \geq 4$, so that U is not the empty string.)

We show now that there are most two lollipop loops incident to $\{V, V^c\}$, and in the process we will determine the structure of the vertices that are incident to two lollipop loops. Suppose that there are at least two lollipop loops incident to $\{V, V^c\}$. We know from the preceeding paragraph that we can write V either as $V = XUU^c$ or as $V = UU^cX^c$, where X is one of the letters A, C, G, T, and U is a string of length m. We give full details in the case that $V = XUU^c$; the details in the case that $V = UU^cX^c$ are similar. For the sake of concreteness (and to avoid continuing proliferation of notation), suppose that $\{VA, TV^c\}$ is one of the lollipop loops incident to $\{V, V^c\}$, so that $VA = TV^c$; the other cases, of adding C, G, or T to the end of V, are handled similarly. Since $VA = TV^c$ and $V = XUU^c$, we have that $XUU^cA = TUU^cX^c$, and so $X = T$. So, $V = TUU^c$.

We first argue that none of the other n-mers obtained by adding a letter to the end of V can be lollipop loops: Since V begins in T, V^c ends in A. Therefore, we have that $VC \neq GV^c$, $VG \neq CV^c$, and $VT \neq AV^c$, since the two strings in each non-equality end with different letters. So, if there is a second lollipop loop incident to $\{V, V^c\}$, it is obtained by adding a letter to the beginning of V.

Write $U = Y_1 \dots Y_m$, so that $V = TUU^c = TY_1 \dots Y_m Y_m^c \dots Y_1^c$. We are given that there is a second lollipop loop incident to $\{V, V^c\}$, and we have just seen that this second lollipop loop must be obtained by adding a letter to the beginning of V: Call this letter Y, and note that since $\{YV, V^cY^c\}$ is a lollipop loop, it is self-complementary, and so $YV = V^cY^c$. Comparing the letters in YV and V^cY^c, together with the information above that $V = TUU^c$, yields the following relationships between the Y_k:

$$\begin{array}{cccccccccccc}
Y & T & Y_1 & Y_2 & \dots & Y_{m-2} & Y_{m-1} & Y_m & Y_m^c & \dots & Y_3^c & Y_2^c Y_1^c \\
\| & \| & \| & \| & & \| & \| & \| & \| & & \| & \| \ \| \\
Y_1 & Y_2 & Y_3 & Y_4 & \dots & Y_m & Y_m^c & Y_{m-1}^c & Y_{m-2}^c & \dots & Y_1^c & A \ Y^c
\end{array}$$

These relationships yield that all the $Y_{\text{odd}} = Y$, that all the $Y_{\text{even}} = T$, and that $Y = A$. Hence, we have that $V = TY_1 \dots Y_m Y_m^c \dots Y_1^c = T(AT)^m$. So, if there are at least two lollipop loops incident to $\{V, V^c\}$ and one of these lollipop loops is $\{VA, TV^c\}$, then $V = T(AT)^m$, and the two lollipop loops incident to $\{V, V^c\}$ are $\{(TA)^{m+1}\}$ and $\{(AT)^{m+1}\}$. By inspection, we can see that there are no further lollipop loops incident to $\{V, V^c\}$. Similarly, if there are at least two lollipop loops incident to $\{Q, Q^c\}$ and one of these lollipop loops is $\{QT, AQ^c\}$, then $Q = (AT)^m A = V^c$, and so $\{V, V^c\} = \{Q, Q^c\}$. The other vertex $(n-1)$-mer incident to two lollipop

loops is $\{(CG)^m C, G(CG)^m\}$, and the two lollipop loops are $\{(GC)^{m+1}\}$ and $\{(CG)^{m+1}\}$.

To complete this case, we need to show that the remaining six n-mers incident to $\{V, V^c\}$ are distinct. We give complete details in the case that $V = T(AT)^m$; the other case is handled similarly. The two lollipop loops are $\{(TA)^{m+1}\}$ (obtained by adding the letter A to the end of V) and $\{(AT)^{m+1}\}$ (obtained by adding the letter A to the beginning of V). The remaining edge n-mers are $\{CT(AT)^m, (AT)^m AG\}$, $\{GT(AT)^m, (AT)^m AC\}$, $\{TT(AT)^m, (AT)^m AA\}$ (obtained by adding a letter to the beginning of V), $\{T(AT)^m C, G(AT)^m A\}$, $\{T(AT)^m G, C(AT)^m A\}$, $\{T(AT)^m T, A(AT)^m A\}$ (obtained by adding a letter to the end of V). From the observation given earlier, we know that the three n-mers obtained by adding a letter to the beginning of V are distinct, as are the three n-mers obtained by adding a letter to the end of V. To see that $\{CT(AT)^m, (AT)^m AG\}$ is distinct from each of $\{T(AT)^m C, G(AT)^m A\}$, $\{T(AT)^m G, C(AT)^m A\}$, $\{T(AT)^m T, A(AT)^m A\}$, we need only observe that the first two letters of the strings involved are all distinct. The cases of $\{GT(AT)^m, (AT)^m AC\}$ and $\{TT(AT)^m, (AT)^m AA\}$ are handled similarly.

Each of the two lollipop loops contributes 2 to the valency of $\mathcal{G}_n$ at $\{V, V^c\}$, while each of the remaining six edge n-mers contributes 1. There is no standard loop incident to $\{V, V^c\}$. Hence, the valency of $\mathcal{G}_n$ at $\{V, V^c\}$ is 10.

Suppose that there is a single lollipop loop incident to $\{V, V^c\}$. From the discussion above, we know that there is a letter X and a string U of length m so that $\{V, V^c\} = \{XUU^c, UU^c X^c\}$, so that either $V = XUU^c$ or $V = UU^c X^c$. We give full details in the case that $V = XUU^c$; the details in the case that $V = UU^c X^c$ are similar. Since $V = XUU^c$, there is no standard loop incident to $\{V, V^c\}$. Since there is a single lollipop loop incident to V, we know from the preceding discussion that V is not of the form $X(X^c X)^m$. Then, the single lollipop loop at $\{V, V^c\}$ is the n-mer $\{XUU^c X^c\}$, obtained by adding X^c to the end of V. It remains only to show that the seven n-mers $\{ZV, V^c Z^c\}$ and $\{VZ, Z^c V^c\}$, where Z ranges over the letters not equal to X^c, together with $\{X^c V, V^c X\}$, are distinct. Using the observation given earlier, we know that the four n-mers of the form $\{YV, V^c Y^c\}$, obtained by adding any letter to the beginning of V, are distinct, as are the three n-mers $\{VZ, Z^c V^c\}$, obtained by adding a letter Z not equal to X^c, to the end of V.

Suppose then that there are letters Z and Y, where Z is not equal to X^c but where there is no restriction on Y, so that $\{VZ, Z^c V^c\} = \{YV, V^c Y^c\}$. Then, either $YV = VZ$ or $YV = Z^c V^c$. Since there is no standard loop incident to $\{V, V^c\}$, we see that $V \neq Y^{n-1}$; the former case cannot then occur, by counting the number of occurrences of the letter Y at the beginning of the two strings.

To see that the latter case cannot occur, write $U = Y_1 \dots Y_m$, so that

$$V = XUU^c = XY_1 \dots Y_m Y_m^c \dots Y_1^c.$$

Then, setting $YV = Z^cV^c$ and comparing the two strings letter by letter, we see that

$$\begin{array}{lccccccccc} YV = & Y & X & Y_1 & \dots Y_{m-1} & Y_m & Y_m^c & \dots & Y_2^c & Y_1^c \\ & \| & \| & \| & \| & \| & \| & & \| & \| \\ Z^cV^c = & Z^c & Y_1 & Y_2 & \dots Y_m & Y_m^c & Y_{m-1}^c & \dots & Y_1^c & X^c. \end{array}$$

This is impossible since no letter is its own complement, and in particular $Y_m \neq Y_m^c$.

The lollipop loop contributes 2 to the valency of $\mathcal{G}_n$ at $\{V, V^c\}$, while each of the remaining seven edge n-mers contributes 1. There is no standard loop incident to $\{V, V^c\}$. Hence, the valency of $\mathcal{G}_n$ at $\{V, V^c\}$ is 9.

This completes the case that there is a self-complementary edge n-mer incident to $\{V, V^c\}$.

The case that there is no self-complementary edge n-mer incident to $\{V, V^c\}$ is very similar to the argument given in the case that n is odd.

Suppose that $V = X^{n-1}$ for some letter X. Since $V = X^{n-1}$, we have that $\{XV, V^cX^c\} = \{VX, X^cV^c\}$, since both are equal to $\{X^n, (X^c)^n\}$. We need to show that the other six edges incident to $\{X^{n-1}, (X^c)^{n-1}\}$, namely the six n-mers $\{ZX^{n-1}, (X^c)^{n-1}Z^c\}$ and $\{X^{n-1}Z, Z^c(X^c)^{n-1}\}$, where Z ranges over the letters not equal to X, are all distinct, and are all distinct from $\{X^n, (X^c)^n\}$. From the observation given earlier, we know that the three n-mers $\{ZX^{n-1}, (X^c)^{n-1}Z^c\}$ are distinct, as are the three n-mers $\{X^{n-1}Z, Z^c(X^c)^{n-1}\}$. It is easy to see that these six n-mers are all distinct from $\{X^n, (X^c)^n\}$, by counting the number of occurrences of the letters X or X^c in each string in each of the n-mers.

Hence, it only remains to show that each of $\{ZX^{n-1}, (X^c)^{n-1}Z^c\}$ is distinct from each of the three n-mers $\{X^{n-1}Z, Z^c(X^c)^{n-1}\}$, where Z ranges over the letters not equal to X. So, suppose there are letters Z and Y, both not equal to X, so that $\{ZX^{n-1}, (X^c)^{n-1}Z^c\} = \{X^{n-1}Y, Y^c(X^c)^{n-1}\}$. Then, either $ZX^{n-1} = X^{n-1}Y$ or $ZX^{n-1} = Y^c(X^c)^{n-1}$. The former case cannot occur, since the two strings begin with different letters (as $Z \neq X$). The latter case cannot occur, since the two strings end in different letters (as $X \neq X^c$).

So, if $V = X^{n-1}$ for some letter X, then there is one standard loop $\{X^n, (X^c)^n\}$ incident to $\{V, V^c\}$, and there are six edges incident to $\{V, V^c\}$, namely the n-mers $\{ZX^{n-1}, (X^c)^{n-1}Z^c\}$ and $\{X^{n-1}Z, Z^c(X^c)^{n-1}\}$, where Z ranges over the three choices of the letters $A, C, G\,T$ not equal to X. There is no lollipop loop incident to $\{V, V^c\}$, since there is no self-complementary edge n-mer incident to $\{V, V^c\}$. The standard loop contributes 2 to the valency of $\{V, V^c\}$ and each of the other six edges contributes 1. So, the valency of $\mathcal{G}_n$ at $\{V, V^c\}$ is 8.

Suppose now that $V \neq X^{n-1}$ for any letter X. In this case, we show that the eight edge n-mers $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, $\{TV, V^cA\}$,

$\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$ are distinct. From the observation given earlier, we know that $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, $\{TV, V^cA\}$ are distinct, as are the four n-mers $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$. We give the details to show that $\{AV, V^cT\}$ is distinct from each of $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$; the other cases are handled similarly. If $\{AV, V^cT\} = \{VA, TV^c\}$, then either $AV = VA$ or $AV = TV^c$. The former case cannot occur, by counting the number of occurances of the letter A at the beginning of the two strings (and using that $V \neq A^{n-1}$, so there must be some letter in V other than A). The latter case cannot occur, since the first letters of the two strings are different. The cases of comparing $\{AV, V^cT\}$ to $\{VC, GV^c\}$ and $\{VG, CV^c\}$ are handled similarly. To see that $\{AV, V^cT\}$ and $\{VT, AV^c\}$ are distinct, note that $AV \neq VT$, again by counting the number of occurances of the letter A at the beginning of the two strings, and that $AV \neq AV^c$, as otherwise V would equal V^c (by deleting the A from the beginning of each of the two strings), and we are working in the case that $V \neq V^c$.

So, if $V \neq X^{n-1}$ for any letter X, then there is no standard loop at $\{V, V^c\}$. There is no lollipop loop incident to $\{V, V^c\}$, since there is no self-complementary edge n-mer incident to $\{V, V^c\}$. There are eight edges incident to $\{V, V^c\}$, namely the n-mers $\{AV, V^cT\}$, $\{CV, V^cG\}$, $\{GV, V^cC\}$, $\{TV, V^cA\}$, $\{VA, TV^c\}$, $\{VC, GV^c\}$, $\{VG, CV^c\}$, $\{VT, AV^c\}$, and each edge contributes 1 to the valency of $\{V, V^c\}$. Hence, the valency of $\mathcal{G}_n$ at $\{V, V^c\}$ is 8.

This completes the analysis of the structure and valencies of the vertices of $\mathcal{G}_n$ in the case of n even (and $n \geq 4$).

Acknowledgments: The authors wish to thank Prof. Colin Please for bringing us together on this problem, and for many helpful discussions. We also wish to thank the referees for their helpful comments.

References

1. B. Bollobas, *Graph Theory: an introductory course*, *Graduate Texts in Mathematics* **63**, Springer-Verlag, New York, 1979.
2. K. R. Fox and M. J. Waring, High Resolution footprinting studies of drug-DNA complexes using chemical and enzymic probes, *Methods in Enzymol.* **340** (2001), 412–430.
3. D. J. Galas and A. Schmitz, DNAase footprinting – Simple method for detection of protein – DNA binding specificity, *Nucleic Acids Res.* **5** (1978), 3157–3170.
4. M. J. Guille and G. Kneale, Methods for the analysis of DNA-protein interactions, *Molecular Biotechnology* **8** (1997), 35–52.
5. A. H. Y. Kwan, R. Czolij, J. P. Mackay and M. Crossley, Pentaprobe: a comprehensive sequence for the one-step detection of DNA-binding activities, *Nucleic Acids Res.* **31** (2003), e124.
6. M. Lavesa and K. R. Fox, Preferred binding sites for [N-MeCys3,N-MeCys7]TANDEM determined using a universal footprinting substrate, *Analytical Biochemistry* **293** (2001), 246–250.
7. P. Pevzner, H. Tang, and M. S. Waterman, An Eulerian path approach to DNA fragment assembly, *Proc. Nat. Acad. Sci. U.S.A.* **98** (2001), 9748–9753.