1-1-2021

Modeling RNA:DNA Hybrids with Formal Grammars

Nataša Jonoska
University of South Florida, Tampa

Nida Obatake
Texas A&M University

Svetlana Poznanović
Clemson University

Candice Price
Smith College, cprice@smith.edu

Manda Riehl
Rose Hulman Institute Technology

See next page for additional authors

Modeling RNA:DNA Hybrids with Formal Grammars

Nataša Jonoska, Nida Obatake, Svetlana Poznanović, Candice Price, Manda Riehl, and Mariel Vazquez

Abstract R-loops are nucleic acid structures consisting of a DNA:RNA hybrid and a DNA single strand. They form naturally during transcription when the nascent RNA hybridizes to the template DNA, forcing the coding DNA strand to wrap around the RNA:DNA duplex. Although formation of R-loops can have deleterious effects on genome integrity, there is evidence of their role as potential regulators of gene expression and DNA repair. Here we initiate an abstract model based on formal grammars to describe RNA:DNA interactions and the formation of R-loops. Separately we use a sliding window approach that accounts for properties of the DNA nucleotide sequence, such as C-richness and CG-skew, to identify segments favoring R-loops. We evaluate these properties on two DNA plasmids that are known to form R-loops and compare results with a recent energetics model from the Chédin Lab. Our abstract approach for R-loops is an initial step toward a more sophisticated

N. Jonoska (✉)
University of South Florida, Tampa, FL, USA
e-mail: jonoska@math.usf.edu

N. Obatake
Texas A&M University, College Station, TX, USA
e-mail: nida@math.tamu.edu

S. Poznanović
Clemson University, Clemson, SC, USA
e-mail: spoznan@clemson.edu

C. Price
Smith College, Northampton, MA, USA
e-mail: cprice@smith.edu

M. Riehl
Rose-Hulman Institute of Technology, Terre Haute, IN, USA
e-mail: riehl@rose-hulman.edu

M. Vazquez
University of California, Davis, CA, USA
e-mail: mariel@math.ucdavis.edu

R. Segal et al. (eds.), *Using Mathematics to Understand Biological Complexity*,
Association for Women in Mathematics Series 22,
https://doi.org/10.1007/978-3-030-57129-0_3

framework which can take into account the effect of DNA topology on R-loop formation.

1 Introduction

RNA can have significant regulatory roles in biological processes such as gene expression, gene inhibition and others (reviewed in the special issue [19]). Recently, some interest has shifted towards the regulatory role of the transcript RNA, often assumed to be just an intermediate towards the protein coding mRNA. In particular, formation of R-loops is seen as a major factor in the RNA transcript involvement in DNA repair [23].

R-loops are three-stranded hybrid structures consisting of an RNA:DNA duplex, and a displaced single strand of DNA (illustrated in Fig. 1). Experimental results indicate the prevalence of R-loops in vastly different genomes. In particular R-loops have been shown to occur with surprising regularity at highly conserved hotspots throughout mammalian genomes [2]. A high throughput sequencing method that can provide genome-wide profiling of R-loops showed that up to 5% of the human genome has the potential of forming R-loops [18]. While R-loops seem to be the most abundant non-B DNA structures found to date (reviewed in [2]), little is known about their function, their mechanism of formation, or their geometry and topology. Most R-loop locations detected in [18] coincided with genes, and there is evidence that R-loops form concurrently with transcription. In a process that is yet to be

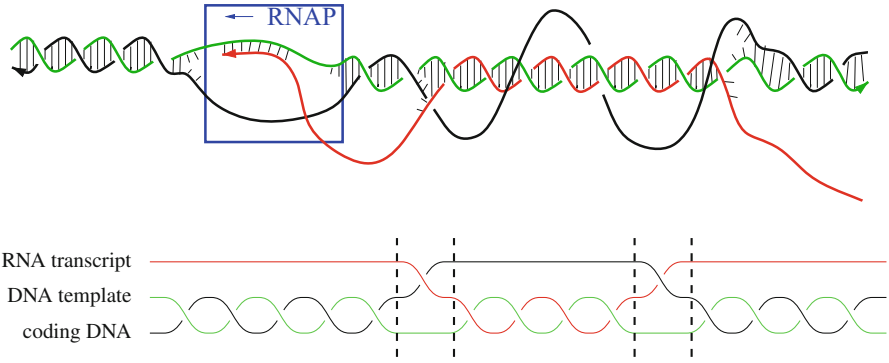

Fig. 1 A schematic depiction of an R-loop. Top: The DNA duplex is formed by two DNA strands in black and green. The black strand represents the coding DNA strand, while the green represents the DNA template (non-coding). The red strand represents the RNA transcript. The 3′-ends are indicated with an arrowhead. The blue box assumes the polymerase reading of the template DNA, and synthesizing the RNA. Bottom: Simplified depiction of the R-loop. We assume that this diagram is read from left to right, with the polymerase on the left (outside the image). The region between the two leftmost vertical dashed lines indicates the location where the RNA transcript invades the DNA duplex, thus initiating the R-loop. Likewise, the two vertical dashed lines on the right indicate the R-loop termination region

understood, the RNA transcript occasionally hybridizes with the DNA template and the second (coding) DNA strand 'entangles' with the RNA:DNA duplex causing the formation of a *co-transcriptional R-loop* [2].

Transcription and the effect of DNA topology on R-loop formation. Transcription is a molecular process that converts a gene encoded in a double stranded DNA molecule into RNA transcript, which is eventually translated into a protein. The double-stranded DNA consists of two sugar-phosphate backbones lined up by complementary sequences of nucleotides (A,T,C,G). One of the strands is the coding strand (i.e. it carries genetic code) and is indicated in black in Fig. 1. The other strand, complementary to the coding strand, is the template strand (indicated in green in Fig. 1). The DNA template is transcribed into RNA by the RNA polymerase. The coding and template DNA strands form a double helix held together by hydrogen bonds. Therefore the transcription machinery (blue box in Figs. 1 and 2a) must break the bonds and open the helix before the RNA polymerase can use the template DNA to produce the complementary RNA transcript. Because DNA is right-handed, the opening of the helix induces an accumulation of torsional stress due to over twisting.

Over twisting is promptly converted into positive supercoiling ahead of the RNA polymerase and compensatory negative supercoiling behind (see twin supercoiling domain, Fig. 2). Note that the local accumulation of supercoiling during transcription, added to the presence of any ambient supercoiling of the DNA, increases torsional stress on the DNA duplex. These factors play a role during the branch migration involving DNA template dissociation from its complementary DNA strand (the coding DNA) and hybridizing with the newly formed RNA transcript, forming and stabilizing an R-loop. To learn more about DNA topology and the effect of transcription in this context, the reader is referred to [1]. The field of DNA topology studies the topology (e.g., knotting and spatial embeddings) and geometry (e.g., twisting, supercoiling) of circular or topologically constrained DNA molecules.

An energy-based statistical mechanical model of R-loop formation was proposed in [21]. This model and its computer implementation (R-looper) incorporates contributions from both sequence and DNA topology to predict the most favorable locations of R-loop formation assuming that the system is in equilibrium. R-looper aims to identify factors that contribute to changes in energy during R-loop formation, and to predict genetic locations favorable for R-loops. The strong role of DNA supercoiling in R-loop formation predicted by the simulation was experimentally supported, as was the strong role of the intrinsic properties of the template DNA sequence (specifically, its *C-richness* and *CG-skew*) [6, 9, 15, 16, 24]. Based on these results a study of R-loop formation must include a discussion of both topology and sequence contributions [21].

In Sect. 2, we introduce a model for R-loops based on a formal grammar with the goal of building a framework for describing the structure of the R-loops and the spatial molecular embedding. This model sets the stage for a future, more sophisticated grammar that could take into account topology and geometry of R-loops (see discussion in Sect. 5). In formal language theory, a grammar is a set of

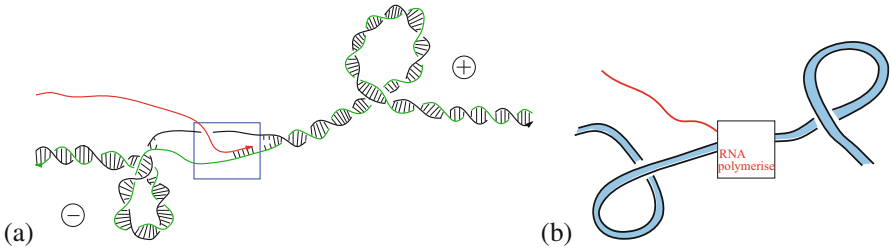

Fig. 2 Local changes in DNA topology during transcription: the twin supercoil domain. The term DNA topology is used by biologists to refer to both topology and geometry of DNA. A supercoil corresponds to a crossing of the axis of the DNA double-helix over itself. When the axis is assigned an orientation, the supercoils are positive or negative depending on the sign of the crossing (as indicated in the image). (**a**) As the DNA template (in green) is transcribed into RNA (in red), positive (+) supercoils accumulate ahead of the polymerase and compensatory negative (−) supercoils accumulate behind. The arrow on the (red) RNA strand indicates the direction of the polymerase. (**b**) The duplex DNA is represented as a ribbon, omitting the helical twists, showing only the positive and the negative supercoils

production rules that generate strings in a formal language. Applications of formal grammars can be found in a wide range of areas from theoretical computer science, to theoretical linguistics, to molecular biology. In molecular biology, applications include modeling regulation of gene expression [3], gene structure prediction [4], and RNA secondary structure prediction [17].

The formal grammar model for R-loops presented here focuses on the structure of an R-loop as described by the braiding of the strands as illustrated in Fig. 1 (bottom), and is informed by sequence contributions. More precisely, the proposed grammar rules depend on the relative nucleotide sequence favorability for R-loop formation. Several experimental results indicate that the presence of a G-rich RNA transcript provides relatively higher thermodynamic stability of an RNA:DNA duplex over a DNA:DNA duplex [10, 13, 14, 22]. This may lead to the breaking of the hydrogen bonds within a topologically strained DNA:DNA duplex. Breakage can then trigger a RNA:DNA branch migration, and the affinity for hybridization of a G-rich RNA transcript with its DNA template may yield favorable regions for R-loop formation. We propose a test for the sequence dependency for R-loop formation and compare our approach with the results from R-looper whose results have been experimentally tested with two plasmids [21].

This chapter is organized as follows. Section 2 gives the necessary background on formal grammars. Section 3 defines the grammar that describes a language for R-loops. Section 4 discusses incorporating the nucleotide sequence dependency into the mathematical framework. In particular, our model takes into account sequence contributions from C-richness and CG-skew. We conclude with an outline of future steps, including a discussion on the contributions of DNA topology and other entanglement considerations in Sect. 5.

2 Formal Grammars: An Overview

In this section we give a short background on formal grammars needed for this work. Good introductions to the different types grammars and languages as well as their properties can be found in [8] and [20].

A finite set Σ is called an *alphabet* and its elements are called *symbols*. For an alphabet Σ, let Σ^* denote the set of all finite sequences of symbols called *strings* or *words* formed by the symbols of Σ. The *empty word* is a word with no symbols and is denoted ϵ. We set $\Sigma^+ = \Sigma^* \setminus \{\epsilon\}$. For example, if $\Sigma = \{a, b\}$, then the set of words over Σ is $\Sigma^* = \{\epsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, \dots\}$ while $\Sigma^+ = \{a, b, aa, ab, ba, bb, aaa, aab, aba, \dots\}$. We use lower case letters at the beginning of the Roman alphabet (e.g., $a, b, c, \dots$) to indicate symbols and lower case letters at the end of the Roman alphabet to indicate words (e.g., $u, v, w, \dots$).

For a word $u = a_1 a_2 \cdots a_n$, where $a_i \in \Sigma$, we say that the *length* of u is $|u| = n$. The length of the empty string ϵ is 0. The number of symbols in u equal to a is denoted $|u|_a$, for example $|aab|_a = 2$ and $|aab|_b = 1$. For any $1 \leq i < j \leq n$, the substring $a_i a_{i+1} \cdots a_j$ of u is denoted by $u_{[i,j]}$.

Definition 1 A *grammar* Γ is a 4-tuple (S, N, Σ, P), where

- N is an alphabet whose symbols are called *nonterminals*,
- Σ is an alphabet whose symbols are called *terminals*,
- P is a finite set of *production rules* of the form $w \to w'$ for some words $w, w' \in (\Sigma \cup N)^*$ provided that at least one symbol in w is nonterminal, and
- $S \in N$ is a designated nonterminal called the *start symbol*.

Here we adopt the standard convention where upper case Roman characters (e.g. A, B, S) are used to denote the nonterminals, and lower case Roman characters (e.g. a, b, c) to denote the terminals. Let u be a word in $(\Sigma \cup N)^*$ and $r : w \to w'$ be a rule in P. Applying the production rule r to the word (or string) u means finding a substring w in u and replacing it with w', while keeping the rest unchanged. For example, if $x, y, w \in \Sigma^*$, applying rule $w \to w'$ to the word $u = xwy$ produces $v = xw'y$. We write $u \overset{r}{\Rightarrow} v$.

We say that a word $w \in \Sigma^*$ can be *derived*, and denote it as $S \overset{*}{\Rightarrow} w$, if there is a sequence of production rules $r_1, r_2, \dots, r_n$ and a sequence of words in $w_1, \dots, w_n \in (\Sigma \cup N)^*$, where $w_n = w$, such that

$$S \overset{r_1}{\Rightarrow} w_1 \overset{r_2}{\Rightarrow} w_2 \overset{r_3}{\Rightarrow} \cdots \overset{r_n}{\Rightarrow} w_n = w.$$

Such a sequence of applications of rules $r_1, r_2, ..., r_n$ is called a *derivation* of w. The *language* described by the grammar Γ is the set of all words with only terminal symbols that can be derived, i.e.,

$$L(\Gamma) := \{w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w\}.$$

Example 1 Consider the grammar $\Gamma = (S, N, \Sigma, P)$ with

$$\begin{aligned} N =& \{S, A, B\} \qquad \Sigma = \{a, b\} \\ P =& \{r_1 : S \to aA,\ r_2 : S \to bB,\ r_3 : A \to aA, \\ & r_4 : B \to bB,\ r_5 : A \to \epsilon,\ r_6 : B \to \epsilon,\ r_7 : S \to \epsilon\}. \end{aligned}$$

The word $aaaa$ is in $L(G)$ because it can be derived in the following way:

$$S \overset{r_1}{\Rightarrow} aA \overset{r_3}{\Rightarrow} aaA \overset{r_3}{\Rightarrow} aaaA \overset{r_3}{\Rightarrow} aaaaA \overset{r_5}{\Rightarrow} aaaa.$$

Based on its production rules, the language $L(\Gamma)$ described by this grammar Γ consists of words with a single symbol, that is

$$L(\Gamma) = a^* \cup b^*.$$

We note that a common abuse of notation when the alphabet is just a singleton is to replace $\{a\}$ with a, and $\{a\}^*$ with a^*.

We often remove the superscript over the arrows when rules are easily identifiable and write $u \to v$ instead of $u \overset{r}{\Rightarrow} v$. We use the convention to shorten the description of the rules that have the same left hand-side by using vertical bars. For example, the expression $w \to w_1 \mid w_2 \mid w_3$ means that the set of rules P contains $w \to w_1$, $w \to w_2$, and $w \to w_3$. With these conventions, a grammar is completely determined by the list of production rules.

Example 2 Consider the grammar defined with rules:

$$S \to aSb \mid \epsilon.$$

In this grammar there are two rules, $S \to aSb$ and $S \to \epsilon$. The symbol S is the only nonterminal, and the set of terminals is $\{a, b\}$. Then, every derivation is of the form:

$$S \Rightarrow aSb \Rightarrow aaSbb \cdots \Rightarrow aa \cdots aSbb \cdots b \Rightarrow aa \cdots abb \cdots b.$$

Because with the application of the rules the number of a's remains equal to the number of b's, the language defined by this grammar is $L = \{a^n b^n \mid n \geq 0\}$.

In formal language theory, the Chomsky hierarchy refers to the containment hierarchy of four levels of languages: regular, context-free, context-sensitive and computably enumerable languages [20].

A grammar is said to be *regular* if all production rules are of the type: $A \to a$, $A \to aB$, or $A \to \epsilon$, where $A, B \in N$ and $a \in \Sigma$. This means that with each rule, either a nonterminal is replaced by a terminal, or it is replaced by a terminal followed by another nonterminal, or it is erased. *Regular languages* are those that

can be generated by a regular grammar. The grammar in Example 1 is regular and, consequently it describes a regular language $a^* \cup b^*$.

In *context-free* grammars, the rules are of the type $A \rightarrow x$, where $A \in N$ and x is a string (possibly empty) of terminals and nonterminals. In *context-sensitive* grammars, the rules are of the type $xAy \rightarrow xzy$, where x, y, z are strings of terminals and nonterminals. Context-free (respectively, context-sensitive) languages are the ones that can be generated by context-free (respectively, context-sensitive) grammars. The grammar in Example 2 and the language it generates are context-free. One can show that this language is not regular, i.e., there is no regular grammar that defines this language [8, 20]. *Computably enumerable languages* are defined by grammars without constraints.

3 R-Loop Grammars

An R-loop is a structure consisting of a RNA:DNA hybrid and a displaced DNA single strand (Fig. 1). While here we focus on co-transcriptional R-loops, the mathematical framework can be applied to any R-loop or other nucleic acid triplex. First we summarize the transcription process which infers the construction of the grammar.

During transcription, the RNA polymerase complex binds to the promoter region of a gene in a double-stranded DNA molecule, unwinds the DNA double helix, and transcribes the template strand into a single-stranded RNA molecule (the *RNA transcript*) one nucleotide at a time in the $3' \rightarrow 5'$ direction. The nucleotide sequence of the RNA transcript is complementary to that of the DNA template, and is identical to the sequence of the coding DNA after replacing each T with a U. As transcription proceeds along the template (as the polymerase moves) the RNA transcript exits the 'bubble' formed by the polymerase complex and the unwound DNA duplex. Simultaneously, the DNA double helix reforms behind the complex. At the end of the process the RNA transcript is released. For reasons that are still unclear, occasionally the RNA transcript hybridizes with the DNA template thus giving rise to a co-transcriptional R-loop.

We represent the formation of an R-loop as a string (word) over the alphabet $\Sigma = \{\sigma, \hat{\sigma}, \tau, \hat{\tau}, \alpha, \omega\}$. Each symbol in the alphabet can be described as a 3-stranded local structure corresponding to the length of one half turn of B-form DNA, approximately 5 nucleotides (see Fig. 3). The symbols τ and $\hat{\tau}$ represent a RNA:DNA hybrid, σ and $\hat{\sigma}$ represent a DNA:DNA duplex, and α and ω represent a structure where all three strands interact. Note that in τ and $\hat{\tau}$, and in σ and $\hat{\sigma}$, the third strand is assumed to not interact (via hydrogen bonds) with the duplex.

Presence of $\hat{}$ on top of the symbols, such as $\hat{\sigma}$ or $\hat{\tau}$, indicates that the corresponding duplex is in a stable configuration. Less stable half-turn configurations are denoted by σ and τ, without the $\hat{}$, are more likely to transition into one of the 3-stranded hybrids α or ω via strand branch migration. The production rules will be guided by the stability of the half-turns σ and τ (DNA:DNA and RNA:DNA,

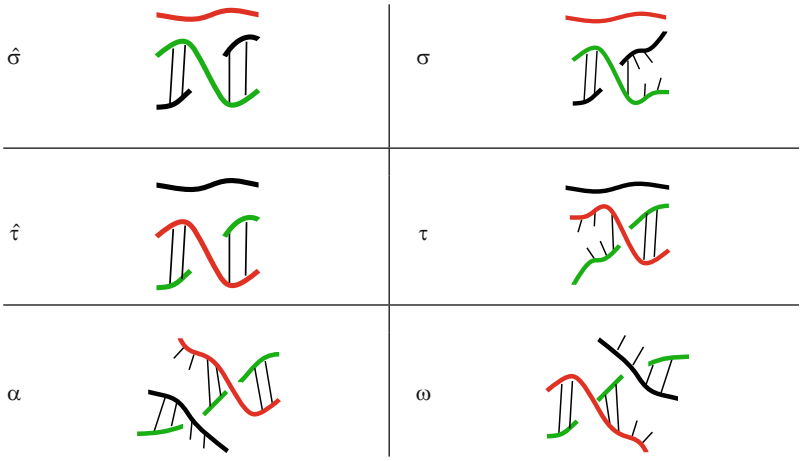

Fig. 3 Interpretation of each of the symbols used in the grammar. The black strand represents the coding DNA strand, the green strand represents the template DNA strand, and the red strand represents the RNA transcript. Here, σ and $\hat{\sigma}$ are DNA:DNA hybrids, τ and $\hat{\tau}$ are RNA:DNA hybrids, and α and ω are transitions between the two. The ' $\hat{}$ ' indicate more stable configurations. Less stable configurations are depicted with some breakage in the hydrogen bonds to suggest that there is more prevalent 'breathing' of the duplex in that region. The breakage in σ and τ is indicated only by symbols in Fig. 4

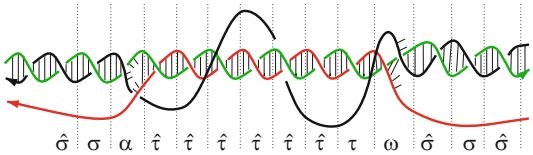

Fig. 4 An example R-loop associated with the word $\cdots\ \hat{\sigma}\sigma\alpha\hat{\tau}\hat{\tau}\hat{\tau}\hat{\tau}\hat{\tau}\hat{\tau}\tau\omega\hat{\sigma}\sigma\ \cdots$. Note that if the sequence stability weakens within an R-loop then a τ may follow after an initial string of one or more $\hat{\tau}$'s, and this may lead into an R-loop termination region, indicated by ω. The three strand sections corresponding to α and ω indicate the branch migration when RNA 'invades' the DNA duplex (α) and leaves the duplex (ω). Observe that there may be other words that correspond to the same R-loop, because the sequence stability may vary both within and outside the R-loop

respectively). The remaining two symbols of the alphabet, α and ω, are used to represent R-loop initiation and termination regions. The start of the R-loop, denoted by α, is the structure formation at the moment when RNA 'invades' the DNA:DNA duplex. The end of the R-loop, denoted by ω, is the structure obtained when the RNA dissociates from the RNA:DNA duplex and the DNA returns to its native state. Figure 4 illustrates how an R-loop can be represented by a string over Σ.

How the different symbols are assigned to the specific genomic region will be guided by the biology. Some nucleotide sequences that are prone to R-loop formation have been identified experimentally and models have been proposed to predict them. As a first approximation, in the next section we use the preliminary

data used in a recent energetics model to assign stable vs unstable half-turn segments in both a RNA:DNA duplex and in a DNA:DNA duplex [21]. We assume that a DNA sequence of nucleotides is favorable for R-loop formation (i.e. the RNA:DNA duplex is more stable) if it is C-rich and CG-skewed (see Sect. 4 for definitions of these terms).

We assume that an R-loop necessarily starts with a short nucleotide sequence that has an "unstable" DNA:DNA duplex, indicated by σ. The three strand formation when the RNA 'invades' the DNA duplex through branch migration is indicated by α and is followed by a half-turn of a stable RNA:DNA duplex indicated by $\hat{\tau}$. Hence the word contains the subword $\sigma\alpha\hat{\tau}$. Similarly, the end of the R-loop is obtained from a sequence starting with at least one unstable RNA:DNA half-turn denoted by τ, followed by the three strand formation ω where the RNA dissociates from the DNA and by a stable DNA duplex ($\hat{\sigma}$). Hence the R-loop word must also contain the subword $\tau\omega\hat{\sigma}$.

The following grammar Γ generates the words associated with R-loops. Recall from Sect. 2 that by defining the production rules we uniquely define a grammar.

	$S \to \hat{\sigma} D \mid \sigma D$	rules:start
	$\hat{\sigma} D \to \hat{\sigma}\hat{\sigma} D \mid \hat{\sigma}\sigma D$	rules:s − D − duplex
the grammar Γ :	$\sigma D \to \sigma\hat{\sigma} D \mid \sigma\sigma D \mid \sigma\alpha\hat{\tau} R$	rules:us − D − duplex
	$\hat{\tau} R \to \hat{\tau}\hat{\tau} R \mid \hat{\tau}\tau R$	rules:s − R − duplex
	$\tau R \to \tau\hat{\tau} R \mid \tau\tau R \mid \tau\omega\hat{\sigma} D'$	rules:us − R − duplex
	$D' \to \hat{\sigma} D' \mid \sigma D' \mid \epsilon$	rules:end

The set of nonterminals in Γ is $N = \{S, D, R, D'\}$, where each of the nonterminals is associated with one of the hybrid structures. The nonterminals D and D' are used to generate DNA duplexes before and after the R-loop, respectively. The nonterminal R is used to generate the symbols that correspond to the RNA:DNA duplex within the R-loop. The symbol S is the starting symbol of the grammar. The grammar Γ uses six types of rules as indicated above. Rules start are the starting rules that generate either σD or $\hat{\sigma} D$, a half-turn DNA duplex. If the DNA duplex represented by $\hat{\sigma}$ is stable, then according to rules s-D-duplex (stable DNA duplex) the next half-turn must be another DNA duplex, which could be stable ($\hat{\sigma} D$), or not (σD). If the DNA duplex represented by σ is not stable (i.e. σD), then according to the rules us-D-duplex (the unstable duplex rules) it can also be followed by a three strand formation α and a stable RNA:DNA duplex $\hat{\tau}$ (e.g. $\sigma D \to \sigma\alpha\hat{\tau} R$). Rules s-R-duplex (stable RNA:DNA duplex) and rules us-R-duplex (unstable RNA:DNA duplex) are analogous to the rules s-D-duplex and us-D-duplex, except that they generate the string corresponding to the R-loop. The first two rules in end are analogous to rules start, with the addition of the last rule $D' \to \epsilon$ which is used to stop the word derivation.

The grammar Γ is context-sensitive, meaning that the rules with nonterminals D and R on the left hand side depend on the preceding symbol. Recall that the language derived from the grammar Γ is defined as the set of words with only terminal symbols which can be derived. Describing the set of words generated by the grammar is straightforward. Every word derivation in Γ starts with S and generates σ or $\hat{\sigma}$ followed by a nonterminal D. Depending on whether σ or $\hat{\sigma}$ precedes D, the next symbols that are generated are again σ or $\hat{\sigma}$ (rules s-D-duplex). In addition, if σ precedes D (rules us-D-duplex), then the next symbols could be $\sigma\alpha\hat{\tau}$ which generate the word $x\sigma\alpha\hat{\tau}R$ where $x \in \{\sigma, \hat{\sigma}\}^*$. After this word one can only apply rules s-R-duplex which generate new symbols τ or $\hat{\tau}$ with a non-terminal R. Rules s-R-duplex and us-R-duplex are then applied to symbols $\hat{\tau}$ and/or τ followed by R. If at some point we use the last rule of us-R-duplex, the symbols that follow are $\tau\omega\hat{\sigma}$ followed by a nonterminal D', and the corresponding word has the form $x\sigma\alpha\hat{\tau}y\tau\omega\hat{\sigma}D'$, where $y \in \{\tau, \hat{\tau}\}^*$. Note that once D' appears in a word, the rules end are the only rules that can be applied and they generate symbols σ or $\hat{\sigma}$. The derivation stops with an application of rule $D' \to \epsilon$. In sum, the final word generated by the grammar has the form $x\sigma\alpha\hat{\tau}y\tau\omega\hat{\sigma}z$ where $z \in \{\sigma, \hat{\sigma}\}^*$. Based on this we have the following proposition. We call the formal language specified with Γ the *R-loop language*.

Proposition 1 *The R-loop language described by Γ is*

$$L(\Gamma) = \left\{x\sigma\alpha\hat{\tau}y\tau\omega\hat{\sigma}z \mid x, z \in \{\sigma, \hat{\sigma}\}^*, y \in \{\tau, \hat{\tau}\}^*\right\}$$

$$\textit{or equivalently,} \quad L(\Gamma) = (\sigma \cup \hat{\sigma})^* \, \sigma \, \alpha \, \hat{\tau} \, (\tau \cup \hat{\tau})^* \, \tau \, \omega \, \hat{\sigma} \, (\sigma \cup \hat{\sigma})^*.$$

As a consequence of Proposition 1, the language of the grammar Γ is regular, that is, it can be described by a regular expression [20]. Therefore, there is a regular grammar $\hat{\Gamma}$ with the same alphabet Σ that is equivalent to Γ. $\hat{\Gamma}$ defines the same R-loop language with the following production rules:

the grammar $\hat{\Gamma}$:

$$\begin{aligned}
S &\to \sigma S \mid \hat{\sigma} S \mid \sigma Q_1 \\
Q_1 &\to \alpha Q_2 \\
Q_2 &\to \hat{\tau} Q_3 \\
Q_3 &\to \tau Q_3 \mid \hat{\tau} Q_3 \mid \tau Q_4 \\
Q_4 &\to \omega Q_5 \\
Q_5 &\to \hat{\sigma} Q_6 \\
Q_6 &\to \sigma Q_6 \mid \hat{\sigma} Q_6 \mid \epsilon
\end{aligned}$$

Although $\hat{\Gamma}$ is regular, i.e., it is a different grammar from the initial context-sensitive grammar Γ, the meaning of the symbols $\sigma, \hat{\sigma}, \tau, \hat{\tau}$ that indicate nucleotide stability remains unchanged.

4 A Discrete Model to Estimate R-Loop Favorability

In order for the grammar Γ presented in Sect. 3 to be useful, it requires information about how favorable or unfavorable a particular stretch of DNA is for R-loop formation. The grammar as constructed in the previous section contains no such information. We are interested in incorporating into the grammar Γ information on nucleotide sequence contributions combined with the topological changes, to detect and predict R-loop formation. We initiate this line of work in this publication. In this section we start with a simple discrete model for estimating R-loop favorability based only on properties of the nucleotide sequence. More specifically, we focus on two different measures, C-richness (cytosine-rich sequence) and CG-skew (cytosine-guanine ratio).

Intuitively, one would expect R-loops to occur infrequently. However, they account for up to 5% of the human genome, and represent the most common non-B DNA structures quantified to date (reviewed in [2]). General investigations of nucleic acid hybrid stability have measured the relative free energy of a nucleic acid duplex (denoted by $\Delta\Delta G^0$ in kcal/mole) as the free energy of the given duplex minus the free energy of the most stable duplex [13, 14]. The results of these studies are summarized in Fig. 5 and show that a hybrid consisting of a purine-rich RNA and pyrimidine-rich DNA (denoted r(GA).d(CT)) is significantly more stable than the corresponding DNA:DNA hybrid (denoted d(GA).d(CT)). The stabilities of the other two possible duplexes (denoted r(GU).d(CA) and d(GT).d(CA)) are comparable to each other. This suggests that the strand migration initiated by an RNA strand as it invades a DNA:DNA duplex is more likely to occur when the template DNA is C-rich (or equivalently, when the transcript RNA is G-rich). On the other hand, the d(GA).r(CU) stacking is significantly less stable than the d(GA).d(CT) stacking, while the stabilities of r(CA).d(GT) and d(CA).d(GT) (not shown in Fig. 5) are comparable to each other. So, a G-rich DNA template would result in a thermodynamically unfavorable RNA:DNA duplex. Based on this, we expect that the R-loops are more likely to form in regions where the template DNA is C-rich and CG-skew. In the discrete model below we focus on those two quantities.

Experimental observations suggest that C-rich regions that are also CG-skewed are correlated with R-loop occurrence [6]. However, not all CG-skewed areas are associated with R-loops [15]. In [21] the authors proposed a statistical mechanics model of R-loop energetics that provides predictions of R-loop favorability for a given nucleotide sequence. This model takes into account contributions of both the sequence and the supercoiling. The theoretical predictions were tested experimentally on two plasmids [2, 21].

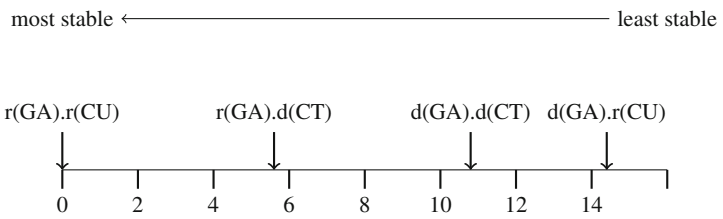

Fig. 5 Plot of the relative stability of duplexes; from the most stable duplex (r(GA).r(CU)) at the left, to the least stable duplex (d(GA).r(CU)) at the right. Stability is determined by computing the relative free energy ($\Delta\Delta G^0$) of the duplexes in kilocalories per mole [13, 14]. The scale indicates the relative free energy of a duplex with respect to the most stable duplex shown as reference at 0. Other duplexes, such as r(CA).d(GT) and d(CA).d(GT) that are above 8 are not indicated in order to keep the figure readable, and because they are not significant in our sequence analysis of R-loops

Before we proceed we introduce several definitions. The DNA alphabet is $\Sigma_{DNA} = \{A, G, C, T\}$, and the binary alphabet is $\Sigma_B = \{0, 1\}$. A DNA sequence is a word $w \in \Sigma^*_{DNA}$. Recall that the number of symbols in w equal to a symbol a is denoted $|w|_a$.

Definition 2 ([7]) The *CG-skew* of a DNA sequence w is a function $C_{sk} : \Sigma^+_{DNA} \to [-1, 1]$ where

$$C_{sk}(w) = \begin{cases} \dfrac{|w|_C - |w|_G}{|w|_C + |w|_G} & \text{for } |w|_C + |w|_G > 0 \\ \\ 0 & \text{otherwise .} \end{cases}$$

The CG-skew of a DNA sequence measures the dominance in occurrences of cytosine with respect to guanine. If $C_{sk}(w) = 1$, then the DNA segment w is a sequence that contains cytosines but no guanines. Conversely, $C_{sk}(w) = -1$ indicates a sequence that contains guanines but no cytosines. Further, a 2:1 ratio of cytosine to guanine in a strand would result in a CG-skew value of 0.33.

Definition 3 The *C-richness* of a DNA sequence w is a function $C_r : \Sigma^+_{DNA} \to [0, 1]$ where

$$C_r(w) = \frac{|w|_C}{|w|}.$$

In order to explore the contributions of C-richness and CG-skew as predictors for R-loops, we consider the plasmids (given as template strands in the 5'-3' direction) of the experimental analysis from the Chédin lab [21].

We compute CG-skew and C-richness using a sliding window approach. Given thresholds t_1 and t_2, and a sequence of nucleotides, we associate a binary score for

CG-skew and C-richness to each subword of length ℓ; a score of 1 if the threshold is met, and a score of 0 if the threshold is not met. We employ the following definitions.

Let $f : \Sigma^+_{DNA} \to \mathbb{R}$ be a function defines on the set of DNA sequences (i.e. Σ^+_{DNA}). The function can be measuring properties of the sequence, such as C-richness or CG-skew. For each function f let t be a real number indicating the threshold. A *ℓ-window t-threshold for f* is the function $T^t_{\ell,f} : \Sigma^+_{DNA} \to \Sigma_B$ defined with $T^t_{\ell,f}(w) = b_1 b_2 \cdots b_{|w|}$ where

$$\text{for } 1 \le i \le |w| - \ell \qquad b_i = \begin{cases} 1 \text{ if } f(w_{[i,i+\ell-1]}) \ge t \\ 0 \text{ otherwise} \end{cases}$$

$$\text{for } |w| - \ell < i \le |w| \qquad b_i = \begin{cases} 1 \text{ if } f(w_{[i,|w|]}) \ge t \\ 0 \text{ otherwise.} \end{cases}$$

Note that as we reach the end of the string, if there are fewer than ℓ nucleotides left, we consider the same thresholds but on the remaining, shorter, substrings. Thus it becomes more difficult to meet the predetermined thresholds towards the end of the string.

Of interest to the study of R-loops are two functions: the ℓ-window t_1-threshold for CG-skew, $T^{t_1}_{\ell,C_{sk}}$; and the ℓ-window t_2-threshold for C-richness, $T^{t_2}_{\ell,C_r}$. Starting from the first nucleotide in the template DNA we consider a substring of length ℓ and compute the values of $T^{t_1}_{\ell,C_{sk}}$ and $T^{t_2}_{\ell,C_r}$ moving from left to right one nucleotide at each step. If a substring is found to be both CG-skewed and C-rich, then we say it is *double-C-rich*. The double-C-rich string corresponding to w is $w_B = T^{t_1}_{\ell,C_{sk}}(w) \wedge T^{t_2}_{\ell,C_r}(w)$. For two binary strings $u = b_1 \cdots b_k$ and $u' = b'_1 \cdots b'_k$ we define $v = u \wedge u' = c_1 \cdots c_k$ such that for all $i = 1, \ldots, k$ we have $c_i = 1$ if and only if $b_i = b'_i = 1$.

We create a binary string of double-C-richness for the entire nucleotide string (called the *double-C-rich string*). In Examples 3–4, we illustrate these definitions for a hypothetical string of nucleotides.

Example 3 Consider an example string of length 30 in Σ^+_{DNA} given by $w =$AGAGCCCGATCCAGACCCCGACGTTACGAA and a window size $\ell = 10$. Suppose the CG-skew threshold is $t_1 = 0.3$ and the threshold for C-richness is $t_2 = 0.5$.

In the first ten nucleotides, there are 3 C's and 3 G's, so the CG-skew $C_{sk}(w_{[1,10]})$ is $\frac{0}{6}$. For nucleotides 2–11 the CG-skew is $C_{sk}(w_{[2,11]}) = \frac{1}{7}$ and for nucleotides 3–12 we have $C_{sk}(w_{[3,12]}) = \frac{3}{7}$. Since $\frac{3}{7} \ge 0.3$, the first three symbols in the ℓ-window t_1-threshold for $T^{t_1}_{\ell,C_{sk}}(w)$ are 001. The entire string is

$$T^{0.3}_{10,C_{sk}}(w) = 001111111111101111000000000000.$$

If the threshold changes to $t_1 = 0.35$, the string becomes

$$T^{0.35}_{10,C_{sk}}(w) = 00111000111110110000000000000.$$

In this example the threshold for C-richness is $t_2 = 0.5$, and there are only 3 C's in the first ten nucleotides, then the C-richness is $C_r(w_{[1,10]}) = \frac{3}{10} < 0.5$. Therefore the first ten nucleotides are not C-rich and the first symbol of $T^{t_2}_{\ell,C_r}(w)$ is 0. Nucleotides 3–12 have 5 C's, so this meets the threshold of 0.5 and is considered a C-rich region. The resulting string with threshold 0.5 is $T^{0.5}_{10,C_r}(w) = 00111000111111110000000000000$. If we use a threshold of 0.6, the string becomes $T^{0.6}_{10,C_r}(w) = 00000000011000000000000000000$.

Only when the window of size ℓ is both CG-skewed and C-rich do we say that it is double-C-rich, and the corresponding entry in the string w_B receives a value of 1. With thresholds 0.3 and 0.5 (for CG-skew and C-richness, respectively), $w_B = T^{0.3}_{10,C_{sk}}(w) \wedge T^{0.5}_{10,C_r}(w) = 00111000111110110000000000000$. Observe that the string w_D is sensitive to the thresholds chosen. With thresholds $t_1 = 0.3$ and $t_2 = 0.6$, for the same nucleotide sequence the string becomes $w_B = T^{0.3}_{10,C_{sk}}(w) \wedge T^{0.6}_{10,C_r}(w) = 00000000011000000000000000000$.

We consider that an isolated occurrence of double-C-richness does not correspond to likelihood of R-loop formation. This agrees with an in vitro analysis that showed that some accumulation of double-C-richness provides an optimal situation for R-loop formation [15]. Therefore here we consider an *accumulation string with window* j to be a function $Acc^j : \Sigma_B^+ \to \{1, \ldots, j\}^*$ defined by $Acc^j(w_B) = w_A = a_1 a_2 \cdots a_k$ where

$$a_i = \begin{cases} |w_{B[i,i+j-1]}|_1 & \text{for } 1 \leq i \leq |w_B| - j \\ \\ |w_{B[i,|w_B|]}|_1 & \text{for } |w_B| - j < i \leq |w_B|. \end{cases}$$

The string $w_A = Acc^j(w_B)$ is obtained from the binary string $w_B = b_1 b_2 \cdots b_k$ such that a_i gives the number of 1's within a window of size j in the substring $b_i b_{i+1} \cdots b_{i+j-1}$. For any fixed pair of thresholds t_1 and t_2 of C-richness and CG-skew, the values of the symbols in the accumulation string can be interpreted as an indication of R-loop likelihood.

Example 4 Consider the sequence w from Example 3. If we record occurrences of double-C-richness within a window size $j = 5$, then the double-C-rich binary string $w_B = 00111000111110110000000000000$ has an accumulation string of $Acc^5(w_B) = w_A = 33322234544432210000000000000$. Note that $a_1 = 3$, since the substring $b_1 \cdots b_5 = 00111$ has three 1's.

Remark 1 The accumulation strings give a sequence-based estimate for favorable sites for R-loop formation. Such portions of the accumulation strings with high values help in deciding which rule of Γ applies for each half turn of DNA. The accumulation string in Example 4 has length 30 and corresponds to six half-turns of DNA (five nucleotides per half-turn) and therefore it can be represented with a word

containing six σ,τ-symbols. If, for example, we set a threshold of 4 as a minimum requirement in the accumulation string for R-loop favorability, then the length 5 substrings in the accumulation string that contain a 4 (or larger) could correspond to symbols σ, $\alpha\hat{\tau}$, or $\hat{\tau}$, while the length 5 substrings with no values larger than 3 could correspond to symbols τ, $\omega\hat{\sigma}$, or $\hat{\sigma}$. We note that symbols α and ω indicate the location of the 3-strand branch migration. For simplicity we assume that they correspond to transitions between a DNA:DNA hybrid and a RNA:DNA hybrid, and do not assign a number of nucleotides to them.

In Example 4 the segment $a_1a_2\ldots a_5 = 33322$ of the string w_A does not correspond with a favorable region. In this case the first symbol of a word in the R-loop language corresponding to w would be $\hat{\sigma}$. Since $a_8 = 4$, $a_9 = 5$ and $a_{10} = 4$, the second symbol of the word (corresponding to $a_6 \cdots a_{10} = 23454$) could be σ. The string $a_{11} \cdots a_{15} = 44322$ continues to be favorable for R-loop formation, and could in fact indicate the start of an R-loop. Continuing in a similar manner, one possibility for a word from the grammar corresponding to our accumulation string would be $w_R = \hat{\sigma}\sigma\alpha\hat{\tau}\tau\omega\hat{\sigma}\hat{\sigma}$. The use of a threshold of 4 for R-loop favorability in this example is only for illustration purposes.

Hence, starting from a DNA string w, we apply ℓ-window threshold functions for C_{sk} and C_r to obtain the corresponding double-C-rich binary string w_B, to which we associate a j-window accumulation string w_A. Using an accumulation threshold we then produce a word w_R in the R-loop language providing structural information about the DNA:DNA and RNA:DNA hybrids and branch migration.

In order to compare the likelihood of R-loop formation for a given DNA string w using its accumulation string w_A to the existing energetics model [21], we obtained the template sequences (written in the 5' to 3' direction) for the experimentally tested plasmids w = pFC53 of length 3906 nucleotides, and w' = pFC8 of length 3669 nucleotides (personal communication with the authors). The window value for $T^{t_1}_{\ell,C_{sk}}$ and $T^{t_2}_{\ell,C_r}$ for both plasmids was set to $\ell = 10$. We examined the sequences for both plasmids, and experimented with different ℓ-window threshold values t_1 and t_2 of CG-skew and C-richness. For each pair (t_1, t_2), we created an accumulation string w_A by recording the number of occurrences of double-C-richness in each window of length j. We tested values of j between 5 and 100 and found that when j is too small, the accumulation strings are not sensitive enough, and when j is too large, the values in w_A and w'_A miss fluctuations within the double-C-rich regions. We used $j = 50$ (this roughly corresponds to 5 full turns of DNA) for the strings w and w' associated with the plasmids pFC53 and pFC8. We considered all possible pairs of threshold values t_1, t_2 between 0 and 1 with step size 0.05 for the functions C_{sk} and C_r (a total of 400 accumulation strings, 20 for each threshold), and computed the accumulation strings w_A and w'_A. We then compared accumulation string w_A and w'_A to the R-looper output probability string (both indexed by nucleotide position number), and computed Spearman's rank correlation coefficient to find the accumulation string with optimal threshold values.

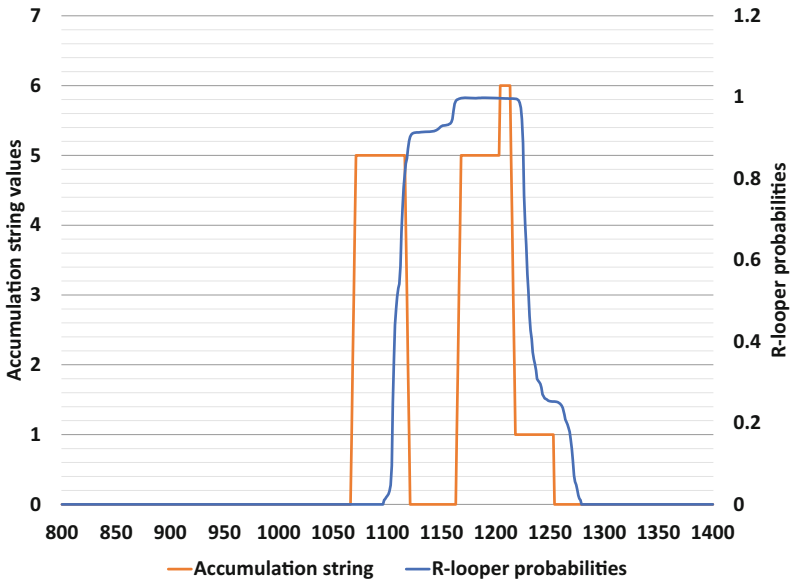

Fig. 6 Comparison between R-looper and the accumulation string analysis for the region of plasmid pFC53 from nucleotide position 800 to 1400. The R-looper probabilities output is indicated by the blue curve. The accumulation string values for pFC53 with $t_1 = 0.1$ and $t_2 = 0.8$ are plotted in orange. Each entry of the accumulation string counts the number of 1's in the corresponding double-C-rich string in the succeeding 50 nucleotides. The region of the plasmid not included in the figure has an R-loop probability zero, and the corresponding accumulation string consists entirely of zeros

For plasmid pFC53 (3906 nucleotides), the correlation coefficient was maximized (at 0.799) when $t_1 = 0.1$ and $t_2 = 0.8$. This result is somewhat surprising since it suggests that C-richness is the determining factor in approximating the R-looper output. The result is shown in Fig. 6. In the graph the maximum accumulation string value was normalized to the same height as a probability of 1 from R-looper.

For plasmid pFC8 (3669 nucleotides) we found optimal thresholds of $t_1 = 0.1$ and $t_2 = 0.6$, with Spearman coefficient of 0.658. Again we see that C-richness is driving the correlation with the R-looper output, as depicted in Fig. 7. The code used for these computations is available publicly at https://github.com/mandariehl/rloopsplusstats.

Comparing the results for these two plasmids, one can observe that the optimal C_r thresholds t_2 disagree. This suggests that C-richness is a larger driving force in R-loop formation than CG-skew, at least when taking R-looper results as a reference. Because the CG-skew threshold is not 0, we believe it is important to continue to include this parameter as structure and topology is incorporated with this model.

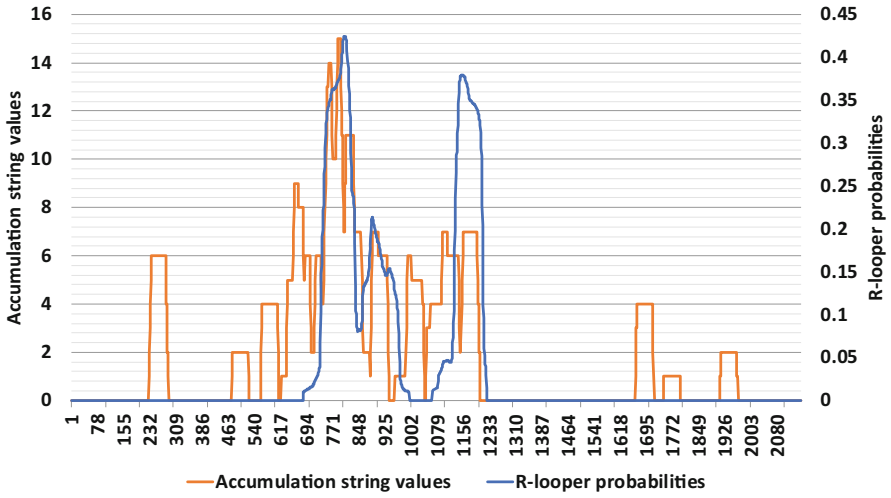

Fig. 7 Comparison between R-looper and the accumulation string analysis for the region of plasmid pFC8 from nucleotide position 1 to 2170. The R-looper probabilities output is indicated by the blue curve. The accumulation string values with $t_1 = 0.1$ and $t_2 = 0.6$ are plotted in orange. Each entry of the accumulation string counts the number of 1's in the corresponding double-C-rich string in the succeeding 50 nucleotides. The plasmid region not included in the figure has R-loop probability zero, and its accumulation string consists entirely of zeros

5 Discussion

The study of R-loops has gained visibility in recent years due to their prevalence and importance for the well-being of the cell. Understanding their mechanism of formation as well as their geometry and topology is key to establishing their biological role. In this paper we propose an abstract framework to model R-loops based on formal grammars, as well as a simple method to assess probability of R-loop formation based on sequence contributions.

It is of interest to obtain a stochastic grammar and a probabilistic model for R-loop formation by attaching probabilities to the production rules in the R-loop grammar Γ proposed in Sect. 3.

Proposition 1 shows that the R-loop language is regular. The class of regular languages is the class of languages described by regular grammars, and is equivalent to the class of languages accepted by finite state automata [20]. Finite state automata equipped with probability measure associated with their state transitions are Markov chain discrete dynamical systems. The framework proposed in Sect. 4 sets the base for determining appropriate probabilities associated with each transition rule in $\hat{\Gamma}$, which is the subject of a future study. Such a model could supplement the predictions based on energy minimization given by R-looper.

A potential advantage of a probabilistic modeling approach is that it can be readily extended to include future sources of statistical information of R-loop

formation. An example of a successful probabilistic model is Pfold [11, 12], which combines a stochastic context-free grammar with evolutionary tree information for a consensus secondary structure prediction of homologous RNA sequences.

The design of a probabilistic grammar affects the prediction of R-loop formation. One advantage of 'lightweight' grammars such as the one proposed in this work is the practicality in their implementation. Moreover, the simplicity of the grammar does not necessarily imply poor predictive power, as can be seen in the case of RNA secondary structure prediction [5].

Since our grammar $\hat{\Gamma}$ is regular and each R-loop has exactly one derivation, the probabilities can be obtained by simple counting: one needs to determine the frequency with which each production rule is used for a set of R-loops. On the other hand, since in Γ the symbols σ and τ correspond to a half-turn pairwise interaction between the two DNA strands, for the template DNA and the RNA strand, the derived strings from such a grammar can be used to infer the three dimensional structure of the molecule when the R-loop initiates.

In Sect. 4, we used plasmids pFC53 and pFC8 [21] to test a simple measure of R-loop favorability based on C-richness and CG-skew. In this way, a given sequence is assigned optimal thresholds t_1, t_2 for R-loop formation. The optimal thresholds obtained for the two plasmids turned out to be different. Averaging them results in a drop of the Spearman correlation coefficient to 0.634 for pFC53 and 0.345 for pFC8. This suggests that C_r and C_{sk} values are not sufficient to predict R-loops. Indeed, it has been shown that R-loops are very sensitive to changes in the topology of the DNA template and that negative superhelicity is required for R-loop stability [21]. The parameters of our discrete model are sensitive to the nucleotide sequence and provide useful information about R-loop favorability in some regions and to inform on the eventual probability assignments to grammar rules. In a refinement of the model, one should consider supercoiling and other measures of topological entanglement.

A discussion on entanglement and geometry must include both the topology of the double stranded DNA before and after the formation of the R-loop, as well as a detailed description of the wrapping of the single stranded DNA around the RNA:DNA hybrid. Important considerations that have not been incorporated in the current model include a description of the wrapping of the nontemplate DNA around the RNA:DNA hybrid, and the supercoiling of DNA prior to R-loop formation. It has been observed that superhelicity can have a dramatic effect on R-loop formation [21]. The twin supercoiling domain model (see Fig. 2) predicts that transcription induces positive supercoiling ahead of the transcription complex, and negative supercoiling behind it. When added to the ambient supercoiling of the DNA template, the negative supercoiling behind the polymerase increases the energetic favorability for R-loop formation [21]. It is of interest to use experimental data to expand Γ with probability parameters, as well as to include topological and geometric constraints within the production rules.

Acknowledgments This research was initiated at the Collaborative Workshop for Women in Mathematical Biology at IPAM in 2019. The authors express their gratitude to the AWM and

SIAM for funding, and to IPAM for fostering an exceptional work environment. The authors acknowledge partial support from NSF grants DMS-1752672 (NO), DMS-1815832 (SP), DMS-1716987 and DMS-1817156 (MV). This work was also partially supported by the grants NSF DMS-1800443/1764366 and the Southeast Center for Mathematics and Biology, an NSF-Simons Research Center for Mathematics of Complex Biological Systems, under National Science Foundation Grant No. DMS-1764406 and Simons Foundation Grant No. 594594 (NJ). The authors thank Eric Reyes and Margherita Ferrari for helpful discussions, and Robert Stolz and the Chédin lab for their help with R-looper and with obtaining the plasmid sequences from their publications.

References

1. Andrew D Bates and Anthony Maxwell. *DNA Topology* Oxford University Press, 2005. ISBN: 978-0198506553.
2. Frédéric Chédin. "Nascent Connections: R-Loops and Chromatin Patterning". In: *Trends in genetics : TIG* 32.12 (Dec. 2016), pp. 828–838. DOI: 10.1016/j.tig.2016.10.002 URL: https://www.ncbi.nlm.nih.gov/pubmed/27793359.
3. Julio Collado-Vides. "Grammatical model of the regulation of gene expression." In: *Proceedings of the National Academy of Sciences* 89.20 (1992), pp. 9405–9409.
4. Shan Dong and David B Searls. "Gene structure prediction by linguistic methods". In: *Genomics* 23.3 (1994), pp. 540–551.
5. Robin D Dowell and Sean R Eddy. "Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction". In: *BMC bioinformatics* 5.1 (2004), p. 71.
6. P. A. Ginno et al. "R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters". In: *Molecular Cell* 45.6 (2012), pp. 814–825. DOI: 10.1016/j.molcel.2012.01.017
7. A. Grigoriev "Analyzing genomes with cumulative skew diagrams". In: *Nucleic Acids Research* 26.10 (Jan. 1998), pp. 2286–2290. DOI: 10.1093/nar/26.10.2286.
8. John E. Hopcroft and Jeffrey D. Ullman. *Introduction to automata theory languages, and computation* Addison-Wesley Series in Computer Science. Addison-Wesley Publishing Co., Reading, Mass., 1979, pp. x+418.
9. F.-T. Huang et al. "Downstream boundary of chromosomal R-loops at murine switch regions: Implications for the mechanism of class switch recombination". In: *Proceedings of the National Academy of Sciences* 103.13 (2006), pp. 5030–5035. DOI: 10.1073/pnas.0506548103.
10. Julian L. Huppert. "Thermodynamic prediction of RNA–DNA duplex-forming regions in the human genome". In: *Molecular BioSystems* 4.6 (2008), p. 686. DOI: 10.1039/b800354h.
11. Bjarne Knudsen and Jotun Hein. "Pfold: RNA secondary structure prediction using stochastic context-free grammars". In: *Nucleic acids research* 31.13 (2003), pp. 3423–3428.
12. Bjarne Knudsen and Jotun Hein. "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history." In: *Bioinformatics (Oxford, England)* 15.6 (1999), pp. 446–454.
13. Lynda Ratmeyer et al. "Sequence Specific Thermodynamic and Structural Properties for DNA.RNA Duplexes". In: *Biochemistry* 33.17 (1994), pp. 5298–5304. DOI: 10.1021/bi00183a037.
14. RW Roberts and DM Crothers. "Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition". In: *Science* 258.5087 (1992), pp. 1463–1466. ISSN: 0036-8075. DOI: 10.1126/science.1279808. eprint: https://science.sciencemag.org/content/258/5087/1463.full.pdf. URL: https://science.sciencemag.org/content/258/5087/1463.
15. D. Roy and M. R. Lieber. "G Clustering Is Important for the Initiation of Transcription-Induced R-Loops In Vitro, whereas High G Density with- out Clustering Is Sufficient Thereafter". In: *Molecular and Cellular Biology* 29.11 (2009), pp. 3124–3133. DOI: 10.1128/mcb.00139-09

16. D. Roy, K. Yu, and M. R. Lieber. “Mechanism of R-Loop Formation at Immunoglobulin Class Switch Sequences”. In: *Molecular and Cellular Biology* 28.1 (2007), pp. 50–60. DOI: 10.1128/mcb.01251-07
17. Yasubumi Sakakibara et al. “Stochastic context-free grammers for tRNA modeling”. In: *Nucleic acids research* 22.23 (1994), pp. 5112–5120.
18. Lionel A. Sanz et al. “Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals”. In: *Molecular Cell* 63.1 (2016), pp. 167–178. ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2016.05.032. URL: http://www.sciencedirect.com/science/article/pii/S1097276516301964.
19. Science. *Special issue on signals in RNA*. Vol. 352(6292). AAAS, June 2016.
20. Michael Sipser. *Introduction to the Theory of Computation* Vol. 2. Thomson Course Technology Boston, 2006.
21. Robert Stolz et al. “Interplay between DNA sequence and negative superhelicity drives R-loop structures”. In: *Proceedings of the National Academy of Sciences* 116.13 (2019), pp. 6260–6269. ISSN: 0027-8424. DOI: 10.1073/pnas.1819476116. eprint: https://www.pnas.org/content/116/13/6260.full.pdf. URL: https://www.pnas.org/content/116/13/6260.
22. Naoki Sugimoto et al. “Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes”. In: *Biochemistry* 34.35 (May 1995), pp. 11211–11216. DOI: 10.1021/bi00035a029.
23. Takaaki Yasuhara et al. “Human Rad52 Promotes XPG-Mediated R-loop Processing to Initiate Transcription-Associated Homologous Recombination Repair”. In: *Cell* 175.2 (2018), 558–570.e11. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2018.08.056. URL: http://www.sciencedirect.com/science/article/pii/S0092867418311176.
24. Kefei Yu et al. “R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells”. In: *Nature Immunology* 4.5 (July 2003), pp. 442–451. DOI: 10.1038/ni919.