Combinatorics and Knot Theory for RNA-DNA Complexes

Margherita Maria Ferrari [1]; Chris Soteros [2] Matthew Schmirler [2] Jean Li [2]; Mykyta Shvets [2]; Mingze Sun [3]

[1]University of Manitoba [2]University of Saskatchewan [3]University of Alberta

Pacific Institute *for the* Mathematical Sciences

VXML

Project Overview

Markov chains are used to explore models of 3-stranded RNA-DNA complexes known as R-loops. Two first-step models are explored: one for R-loop formation and the other for R-loop geometry.

R-loop formation: Ferrari and coworkers are developing a DNA-sequence-dependent method for predicting the formation of R-loops. Their approach is based on a formal grammar model and is informed by experimental data [1, 2]. Here, we use a simpler Markov chain model derived from a formal grammar model [3] to explore whether this simpler random model can capture any of the DNA-sequence-dependent features of R-loop formation.

R-loop geometry: Soteros and coworkers have developed lattice models to study the geometry and topology of 2-stranded (polygons) polymers under tubular confinement [4, 5]. Here, the simplest such 2-stranded model is used to model R-loop geometry. Markov chain theory and simulations are used to explore the effect of strand flexibility on R-loop geometry.

What is an R-loop?

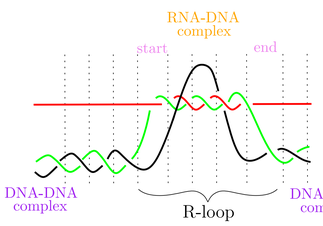
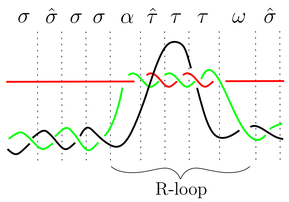


Figure 1. Schematics of an R-loop based on Fig. 4 in [3].

During cellular processes, modifications of the geometry and topology of DNA and RNA can yield multi-stranded structures such as R-loops. An R-loop is a 3-stranded structure composed of an RNA-DNA complex and another single strand of DNA. Experimental studies indicate that R-loops can play either destructive or regulatory roles in cellular processes [6, 7, 8].

A Formal Grammar Model for R-loops



- Symbols correspond to about 4 or 5 nucleotides in the DNA sequence; symbol set $= \{\sigma, \hat{\sigma}, \tau, \hat{\tau}, \alpha, \omega\}$
- DNA-DNA: $\hat{\sigma}$ (stable), σ (unstable)
- RNA-DNA & R-loop: α (start), $\hat{\tau}$ (stable), τ (unstable), ω (end)
- Symbol stability depends on the CG-content of the corresponding DNA sequence [3]

Simplified Markov Chain Model of R-loop Formation

$$
\begin{aligned}
S &\to \sigma\, S\ (71\%) \mid \hat{\sigma}\, S\ (21\%) \mid \sigma\, Q_1\ (8\%)\\
Q_1 &\to \alpha\, Q_2\ (100\%)\\
Q_2 &\to \hat{\tau}\, Q_3\ (100\%)\\
Q_3 &\to \tau\, Q_3\ (35\%) \mid \hat{\tau}\, Q_3\ (52\%) \mid \tau\, Q_4\ (13\%)\\
Q_4 &\to \omega\, Q_5\ (100\%)\\
Q_5 &\to \hat{\sigma}\, Q_6\ (100\%)\\
Q_6 &\to \sigma\, Q_6\ (24\%) \mid \hat{\sigma}\, Q_6\ (67\%) \mid \epsilon\ (9\%)
\end{aligned}
$$

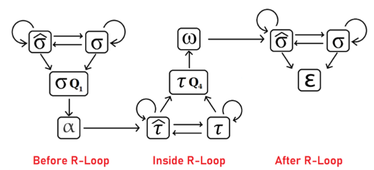


Figure 2. Left: Formal grammar from [3], with probabilities determined by using experimental data (ongoing work with Ferrari and coworkers [1, 2]). Right: Markov chain model corresponding to the formal grammar on the left.

Methods:

- Markov chain theory analysis via SageMath in CoCalc (Jean Li)
- Markov chain simulations via Python and Scala (Mingze Sun and Mykyta Shvets)

Results:

Average Word length

Exact: 36.3034188034188 letters *Simulations:* 36.296 +/- 0.606 letters (600 time-steps, 10000 reps)

Average R-loop length

Exact: 10.6923076923077 letters *Simulations:* 10.568 +/- 0.461 letters (600 time-steps, 10000 reps)

Position-dependent probability of being in an R-loop

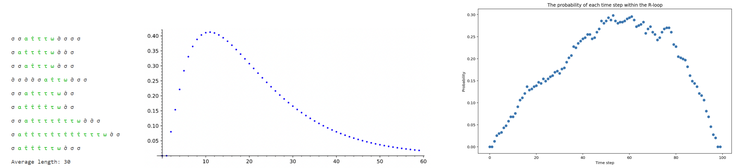


Figure 3. Left: Sample output words from M. Shvets' Python Code 1 with R-loops colored in green. Middle: The probability of the chain being in an R-loop state at time-step n, i.e. $P(X_n \in \{\alpha, \tau, \hat{\tau}, \omega\} | X_0 = \hat{\sigma}S), n = 0, 1, 2, ...$, generated from J. Li's SageMath Code 1. Right: Probability of being in an R-loop at time-step $n, n = 0, 1, ..., 100$ in a word of fixed length 100, generated from M. Sun's Python Code 1 with 10^6 replications.

A Lattice Model for R-loop Geometry/Topology

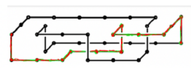


Figure 4. Example of an R-loop modeled as a polygon in the $(2, 1)$-tube. The red/green walk represents the RNA-DNA complex, while the black walk represents the free DNA strand.

- DNA and RNA can be thought as polymers.
- In the standard lattice model, a polymer is represented by a set of vertices in $\mathbb{Z}^3$ joined by unit edges.
- (L, M)-tube: the vertex coordinates are such that $x > 0, 0 < y \leq L, 0 < z \leq M$.

Simplified Lattice Markov Chain Model of R-loop Geometry

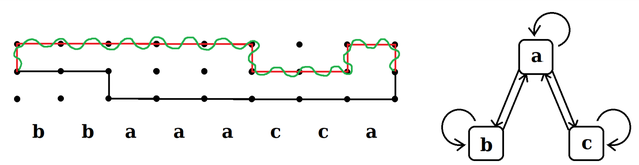


Figure 5. Left: An example of a polygon representing an R-loop embedded in the $(2, 0)$-tube. The top (red/green) walk represents the RNA-DNA complex and the bottom walk represents the free DNA strand. The letters underneath each half-integer plane denote the column state for that position in the R-loop polygon. Right: The state transition diagram used in the R-loop geometry Markov chain model.

Methods:

- Markov chain simulations via Python (Mingze Sun and Mykyta Shvets)
- Transition probability matrix:

$$
P_{p_r,p_b} = \begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{cccc} a & & b & c \end{array} \\ \begin{bmatrix} 1-p_b-p_r & p_b & p_r \\ p_b & 1-p_b & 0 \\ p_r & 0 & 1-p_r \end{bmatrix} \end{array}
$$

where p_r (respectively, p_b) is the probability of bending the top 'red' (respectively, bottom 'black') walk.

- We assume that $p_r < p_b$ since the top walk represents a less flexible RNA-DNA complex.

Results:

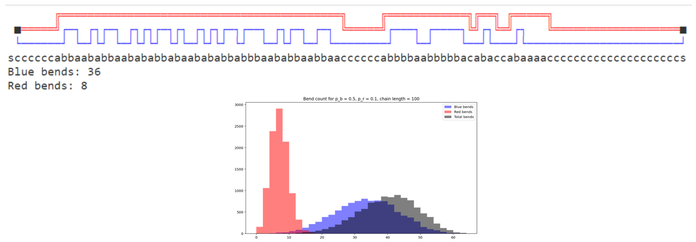


Figure 6. Top: Sample output of M. Shvets' Python Code 2 for span 100. The RNA-DNA complex is in red and the free DNA strand is in blue. Bottom: Histograms of bend counts in R-loops for the RNA-DNA complex (red), free DNA (blue), and total bends (grey). Bending parameters are $p_r = 0.1$, $p_b = 0.5$.

Conclusions

- Exact and Monte Carlo analyses of the two Markov models for R-loops provide consistent results.
- R-loop formation: a purely random model is not sufficient to capture features of the experimental data.
- R-loop geometry: a simplified model can take into account differing flexibility for the R-loop components.
- Future directions: visualize lattice tube configurations in the 3-dimensional space.

References

[1] M. M. Ferrari et al. *In preparation* (2023).
[2] R. Stolz et al. *Proceedings of the National Academy of Sciences* **116**.13 (2019), pp. 6260–6269.
[3] N. Jonoska et al. *Using Mathematics to Understand Biological Complexity: From Cells to Populations*. Ed. by R. Segal, B. Shtylla, and S. Sindi. Cham: Springer International Publishing, 2021, pp. 35–54. DOI: 10.1007/978-3-030-57129-0_3.
[4] N. R. Beaton, J. W. Eng, and C. E. Soteros. *Journal of Physics A: Mathematical and Theoretical* **52**.14 (2019).
[5] J. W. Eng, PhD thesis, University of Saskatchewan, 2020.
[6] F. Chédin. *Trends in Genetics* **32**.12 (2016), pp. 828–838.
[7] F. Chédin and C. J. Benham. *Journal of Biological Chemistry* **295**.14 (2020), pp. 4684–4695.
[8] Y. A. Hegazy, C. M. Fernando, and E. J. Tran. *Journal of Biological Chemistry* **295**.4 (2020), pp. 905–913.