Formal grammar modeling three-stranded DNA:RNA braids

Margherita Maria Ferrari

University of South Florida

`mmferrari@usf.edu`

Nataša Jonoska

University of South Florida

`jonoska@math.usf.edu`

Svetlana Poznanović

Clemson University

`spoznan@clemson.edu`

Manda Riehl

Rose-Hulman Institute of Technology

`riehl@rose-hulman.edu`

Mariel Vazquez

University of California, Davis

`mariel@math.ucdavis.edu`

A formal grammar is a system to generate words; it consists of a set of symbols, partitioned into terminals and non-terminals, and a set of production rules. The production rules specify how to rewrite non-terminal symbols, so that successive applications of those rules yield words formed by only terminals. Adding probabilities to the production rules defines stochastic grammars, which can be used for biological sequence analysis. In this talk, we focus on a "braid grammar" to model R-loops, that are three-stranded structures formed by a DNA:RNA hybrid plus a single strand of DNA, often appearing during transcription. R-loops are described as strings of terminal symbols representing the braiding of the strands in the structure, where each symbol corresponds to a different state of the braided structure. We discuss approaches to develop a stochastic grammar and a probabilistic model for R-loop prediction, as well as refinements of the model by incorporating the effect of DNA topology on R-loop formation.