Comparing RNA models for something something

***Abstract*—results indicate the entropy-ordered enumeration is a promising general-purpose technique to accelerate goal-directed top-down search for program synthesis.**

I. Introduction

RNA molecules can be grouped into families based on shared characteristics such as structural patterns, sequence similarities, and functional roles. This allows researchers to study them in a more organized and meaningful way. To capture these similarities, a technique known as Multiple Read Alignment (MRA) is often employed, in which several RNA sequences from the same family are aligned together. From this alignment, a model can be constructed that encodes the common features of the family. This model can then be used as a reference to evaluate new RNA reads, providing a measure of how likely it is that a given sequence belongs to the same RNA family.

The chosen model for representing RNA families has often been a Stochastic Markov Model (SMM), which is capable of describing many RNA molecules in terms of both their raw symbolic strings and certain structural configurations that the molecules can take. This model has the same expressive power as a Stochastic Regular Grammar (SRG), meaning it can capture a wide range of sequence and structural patterns. However, it is limited in that it cannot represent more complex, long-range dependencies such as knots. These more intricate configurations are referred to as Secondary Structures (SS), and they require more powerful modeling techniques to be fully captured. To address this issue, prior research has utilized a Stochastic Context Free Grammar (SCFG), sometimes also called a probabilistic context free grammar, in order to capture more complex details about RNA Secondary Structures (SS).

We hypothesize that RNA families themselves can be examined for similarity. And further, that this similarity can be determined by comparing the similarity of the models used to represent them. The true similarity between RNA families might be characterized in several different ways. We have chosen to measure similarity by two criteria: the amount of overlap in the languages represented by the models, as quantified by similar RNA/probability pairs, and as quantified by shared secondary structure/likelihood pairs. In both cases, what matters is the overlap in the languages that the models are capable of generating. **Bryan:** Alternatively we could merely classify the difference purely by the similarity of the grammars, but this gives us no way to align our hypothesis with a real world definition of "similar families"

As we are primarily concerned with secondary structural properties, we choose to use the latter technique when comparing different RNA families. However, comparing two context free grammars (CFGs) directly is challenging, if not impossible, in the general case. **Bryan:** CM's might be more restrictive and bring this back into the realm of possibility. Because of the inherent difficulty of performing direct CFG comparisons, in this paper we instead measure the grammatical edit distance Grammatical Edit Distance (GED) between the CFGs. We then use libraries of RNA molecules to empirically evaluate the efficacy of this edit distance as a proxy for family similarity. **Bryan:** MM's might have a more direct comparison means. If so, we might be able to use this as another empirical ground truth comparator, but it may be challenging to determine if variance is explained by the true difference between the families or because the CFG can capture the knots. Much less if the MM's cover the same families as the SCFG models.