

Exploring the Use of Stem-loop Characteristics for Pinpointing Structural RNA Genes

Kirt Noël and Kay C. Wiese

Simon Fraser University

School of Interactive Arts and Technology

2400 Central City, 10153 King George Highway, Surrey, BC, Canada, V3T 2W1

{kirtnoel, wiese}@sfu.ca

Abstract

Our goal is to determine whether a stem-loop metric or combination of stem-loop metrics can be used to identify structural RNAs along genomic sequences across the entire GC content spectrum. Herein we describe the first step towards this goal by comparing selected stem-loop metric valuations between structural RNAs and their genomic counterparts.

1. Introduction

To date an effective and efficient structural RNA gene-finder has been elusive. The difficulty largely results from a lack of sequence conservation in structural RNAs [3]. This work aims to evaluate a stem-loop focused approach to identify structural RNA genes along genomic sequences.

Previous attempts to develop a structural RNA gene-finder are largely dependant on measuring the Free Energy of secondary structures formed by segments or windows along a genomic sequence [1, 2, 4]. Yet, the size of these segments carries no biological relevance. Instead, the chosen segment size is a parameter related to the RNA folding algorithm. The reasoning suggests that structural RNA regions will display unusually low Free Energy values. This is not strongly supported [5]. Lastly, the polynomial computational complexity associated with RNA folding algorithms is not enticing.

The stem-loop is an RNA secondary structure which occurs when complementary regions lying in close proximity leads an RNA transcript to fold back upon itself. Our first task is to determine whether stem-loop structures occur more frequently in structural RNAs than in their genomic counterparts - namely coding sequence (CDS) and non-coding DNA (NC). Our second task is to determine whether the average length of a stem-loop found in structural RNAs

is longer than those found in their genomic counterparts. We hypothesize that if such stem-loop based distinctions are proven they in turn may help to improve the platform upon which structural RNA gene-finders are built.

2. Methods

At the core of this project is a search algorithm designed to identify stem-loops which are typically found in structural RNAs. A cursory analysis of RNA secondary structures provides some basic insight into their general structure. Typically, the stem-loop has between 4 to 15 nucleotides residing in the loop atop the stem. These stem-loops typically have a GC base pair content of 30% or more. Where interruptions such as bulges and internal loops occur along the stem they are usually populated by less than 8 unpaired nucleotides. Such observations help to set the initial parameters for our search algorithm.

Ribosomal RNAs (rRNA) were selected as models for structural RNAs since they are found in bacterial sequences across the GC content spectrum. The genomic sequences were obtained from the NCBI website (<http://www.ncbi.nlm.nih.gov>).

The stem-loop search algorithm is implemented in C++. The average computational time complexity is $O(n)$ and the space complexity is $O(n)$ where n is the number of nucleotides in the sequence. The computer used to run this search algorithm is equipped with an Intel©Pentium 4 Processor and 1.5 gigabytes of RAM.

This exploratory work applies the Central Limit Theorem to identify suspected structural RNA genes along a genomic sequence.

3. Results & Discussion

Our hypothesis appears partially true inasmuch as stem-loops are more frequent in ribosomal RNA genes than in

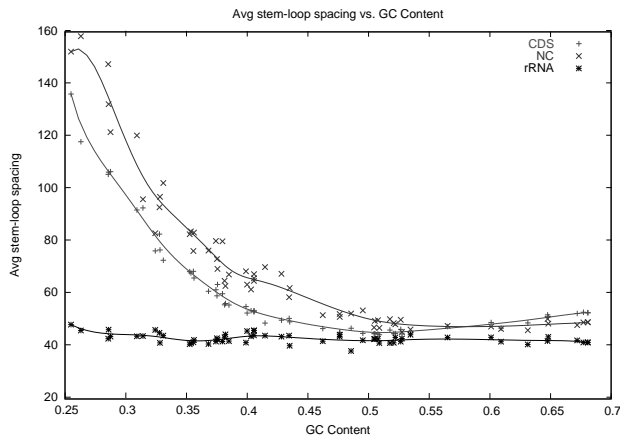


Figure 1. Average Stem-loop Spacing vs. GC Content

their genomic counterparts throughout the GC content spectrum. See Figure 1. However, as GC content levels approach 50% this difference fades. The stem-length measured as the number of base pairs comprising stem-loops is not on-average greater in ribosomal RNAs than in its counterparts across the entire GC spectrum.

The stem-loop spacing metric is particularly interesting since it has a strikingly low standard deviation, see Table 1. The explanation for the stability observed in the spacing metric may be twofold. In the face of fledgling sequence conservation, structural RNA genes (such as the ribosomal RNA genes tested) must continue to conserve vital information. This notion appears to manifest itself in this rather simple stem-loop metric. Secondly, the low standard deviation may indicate that structural RNAs aim for an equilibrium state. This equilibrium, it is postulated, fosters an environment conducive to attaining the final intended structure. RNA transcripts with an over abundance of stem-loops increase the likelihood that an obstructive amount of disassembly may be necessary to reach the final intended RNA structure. Conversely, RNA transcripts which resist folding may hinder distal segments from coming into close proximity. Presumably, an environment between these two extremes may be most conducive to efficient RNA structure folding.

The ability of the stem-loop spacing metric to identify ribosomal RNA genes in an AT rich organism is depicted in Figure 2. Numerous rRNA genes are located in 2 sections located upstream of position 50,000. The segments shown in black and denoted as “structural RNA” coincide with the actual rRNA genes along the sequence, see Figure 2.

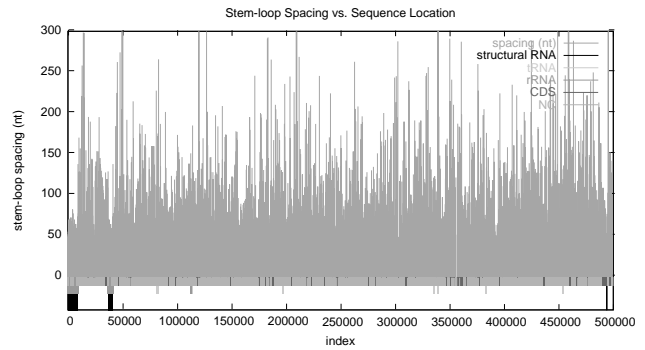


Figure 2. Ribosomal RNA genes located as indicated by the segments highlighted in black. Accession Num. NC 005791. GC Content 33%.

	Mean Spacing	Std. Deviation
Coding Sequence	60.80	20.53
Noncoding DNA	70.60	28.14
Ribosomal RNA	43.23	1.82

Table 1. Stem-loop spacing metric statistics.

4. Conclusion

Further goals aim to develop additional stem-loop metrics in hopes of more adequately distinguishing structural RNAs from their genomic counterparts throughout the GC content spectrum.

References

- [1] R. Carter, I. Dubchak, and S. Holbrook. A computational approach to identify genes for structural RNAs in genomic sequences. *Nucleic Acids Research*, 29:3928–3938, 2001.
- [2] J.-H. Chen, S.-Y. Le, B. Shapiro, K. Currey, and J. Maizel. A computational procedure for assessing the significance of RNA secondary structure. *Computer Applications in the Biosciences*, 6:7–18, 1990.
- [3] S. R. Eddy. Computational genomics of noncoding RNA genes. *Cell*, 109(2):137–140, 2002. Review.
- [4] S.-Y. Le, J.-H. Chen, K. Currey, and J. Maizel. A program for predicting significant RNA secondary structures. *Computer Applications in the Biosciences*, 4:153–159, 1988.
- [5] E. Rivas and S. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583–605, 2000.