Methods 125 (2017) 36–44

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

# The effects of structure on pre-mRNA processing and stability

Rachel Soemedi [a,b], Kamil J. Cygan [a,b], Christy L. Rhine [a], David T. Glidden [a,b], Allison J. Taggart [a], Chien-Ling Lin [a], Alger M. Fredericks [a], William G. Fairbrother [a,b,*]

[a] Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, 70 Ship Street, Providence, RI 02903, USA
[b] Center for Computational Molecular Biology, Brown University, 115 Waterman Street, Providence, RI 02912, USA

ARTICLE INFO

*Article history:*
Received 19 February 2017
Received in revised form 31 May 2017
Accepted 1 June 2017
Available online 6 June 2017

*Keywords:*
Splicing
RNA processing
Secondary structure
G quadruplex
Splice site
ESE
ESS
Disease
Zebrafish
Simple repeats

ABSTRACT

Pre-mRNA molecules can form a variety of structures, and both secondary and tertiary structures have important effects on processing, function and stability of these molecules. The prediction of RNA secondary structure is a challenging problem and various algorithms that use minimum free energy, maximum expected accuracy and comparative evolutionary based methods have been developed to predict secondary structures. However, these tools are not perfect, and this remains an active area of research. The secondary structure of pre-mRNA molecules can have an enhancing or inhibitory effect on pre-mRNA splicing. An example of enhancing structure can be found in a novel class of introns in zebrafish. About 10% of zebrafish genes contain a structured intron that forms a bridging hairpin that enforces correct splice site pairing. Negative examples of splicing include local structures around splice sites that decrease splicing efficiency and potentially cause mis-splicing leading to disease. Splicing mutations are a frequent cause of hereditary disease. The transcripts of disease genes are significantly more structured around the splice sites, and point mutations that increase the local structure often cause splicing disruptions. Post-splicing, RNA secondary structure can also affect the stability of the spliced intron and regulatory RNA interference pathway intermediates, such as pre-microRNAs. Additionally, RNA secondary structure has important roles in the innate immune defense against viruses. Finally, tertiary structure can also play a large role in pre-mRNA splicing. One example is the G-quadruplex structure, which, similar to secondary structure, can either enhance or inhibit splicing through mechanisms such as creating or obscuring RNA binding protein sites.



**Contents**



*Abbreviations:* MFE, minimum free energy; DP, dynamic programming; MEA, maximum expected accuracy; RBP, RNA-binding protein; U2AF2, U2 Small Nuclear RNA Auxiliary Factor 2; HLA, Human Leukocyte Antigen; SNP, single nucleotide polymorphism; HGMD, Human Gene Mutation Database; ESM, exonic splicing mutations; ssRNA, single stranded RNA; dsRNA, double stranded RNA; PAMP, pathogen associated molecular pattern; PPR, pathogen-recognition receptor; RNAi, RNA interference; TLR, toll-like receptor; IFN, interferon; SG, stress granule; LAT, latency associate transcript; HSV, Herpes Simplex Viruses; G4, G-quadruplexes.

* Corresponding author at: Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, 70 Ship Street, Providence, RI 02903, USA.
*E-mail addresses:* rachel_soemedi@brown.edu (R. Soemedi), kamil_cygan@brown.edu (K.J. Cygan), christy_rhine@brown.edu (C.L. Rhine), david_glidden@brown.edu (D.T. Glidden), allison_taggart@brown.edu (A.J. Taggart), chien-ling_lin@brown.edu (C.-L. Lin), alger_fredericks@brown.edu (A.M. Fredericks), william_fairbrother@brown.edu (W. G. Fairbrother).

http://dx.doi.org/10.1016/j.ymeth.2017.06.001
1046-2023/

## 1. Introduction

### 1.1. Determining the structure of an RNA molecule

RNAs play a crucial role in many biological processes including regulation of gene expression and catalysis of cellular processes. RNA molecules can form a variety of structures. There is a correlation between the tertiary structure of an RNA molecule and its interactive and functional capacity [1]. It is however very difficult to predict 3D structure of an RNA molecule with high accuracy and precision. Multiple reviews previously covered the experimental tools used to solve RNA secondary structures [2–4]. This review section will focus on the utilization of computational methods for secondary structure predictions.

Many algorithms have been developed to study RNA secondary structure. Generally speaking, folding algorithms can be divided into two categories: single sequence methods and comparative methods, each having their own advantages and disadvantages.

The single sequence category of tools relies on finding the structure with minimum free energy (MFE) for each sequence. Dynamic programming (DP) has been used extensively in such tools. In essence, the algorithm computes the sum of all possible base pairs and their energy parameters or to put it simply, it determines base pairing probability from the partition function calculation. The partition function is used to count the particles of a system that are distributed over the available energy levels (i.e. the particles follow the Boltzmann distribution). In the RNA structure prediction problem, instead of counting particles, possible RNA structures conformations were counted. In accordance with the Boltzmann distribution, the fraction that a secondary structure conformation occupies in the ensemble of structures amounts to $e^{\frac{-}{\Delta}G(S)}RT$, where ΔG is the Gibb's free energy change of RNA secondary structure conformation $S$, $R$ is the gas constant, and $T$ is the absolute temperature of the system. Partition function $Q$ is therefore a summation over all possible energy levels of all possible secondary structure conformations ($Q = \sum_S e^{\frac{-\Delta G(S)}{RT}}$). If the partition function is seen as a way to calculate the spread over the energy levels of RNA secondary structures, its reciprocal gives an easy way to calculate the probability of any given secondary structure conformation: $P(\textit{secondary structure}) = \frac{e^{\frac{-\Delta G(\textit{secondary structure})}{RT}}}{Q}$. From this equation it follows that to calculate the base pairing probability between nucleotides $i$ and $j$ ($P_{i,j}$), we take the sum over the probabilities of all secondary structures $k$ that contain the pair: $P_{i,j} = \sum_k \frac{e^{\frac{-\Delta G(k)}{RT}}}{Q}$ [5]. Examples of folding algorithms in this category include mfold [6] and RNAfold [7]. Multiple additional assumptions have been used to limit the search space for MFE structures, including limiting base paring to Watson-Crick and wobble GU, as well as exclusion of pseudoknots (features of RNA structure in which base pairings occur between the unpaired bases of a loop and some other bases outside the loop).

Furthermore, the free energy landscape is constantly changing depending on multiple environmental variables like pH, temperature, chaperones and solvent condition. It is therefore not uncommon to consider structures that reside close to the MFE structure on the energy landscape. Programs have been developed to find the suboptimal structures for a given RNA sequence e.g. RNAsubopt [7]. Recently, several methods have been developed to predict secondary structures using maximum expected accuracy (MEA) approach, e.g. CONTRAfold [8], CentroidFold [9], and IPknot [10]. The goal of these algorithms is to find a secondary structure that maximizes the expected base-pair accuracy.

In all of the above tools, the accuracy of a prediction is inversely proportional to the sequence length. To counteract this problem, it is possible to predict stable structures using a small window based approach, where base pairs are limited to a particular distance within that window (Fig. 1). Algorithms such as Rfold [11] and RNAplfold [7] implement this strategy. However, such approach will not capture the long distance interactions between distant sequences and there is a need for further noise reduction (i.e. exclusion of pairs that exhibit a low thermodynamic probability) after calculation.

There are many applications where it may not be necessary to accurately predict the entire structure. For example, one may only

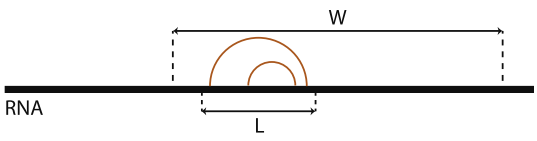


**Fig. 1.** Detecting Local Stable Structures. RNA is examined for local stable structures using a sliding window approach. Within window size of W only base pairings within a maximum distance apart (L) are considered. The orange arcs represent possible valid base pairings.

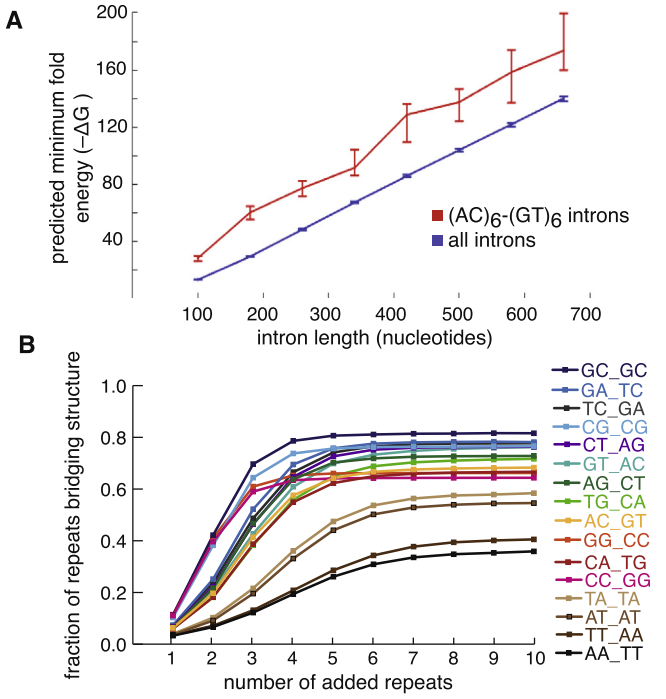


**Fig. 2.** Complementary dinucleotide repeats at the ends of Zebrafish introns drive secondary structure. A. Minimum fold energy of Zebrafish (AC)m-(GT)n introns (red) and all Zebrafish introns (blue). Introns are binned by length and minimum fold energy is calculated using RNAfold. B. All possible dinucleotide repeats are added *in silico* to Zebrafish introns 20 nucleotides downstream of the 5′ss, and the complementary dinucleotide repeat is added 20 nucleotides upstream of the 3′ss. Synthetic introns are folded using RNAfold and the number of introns with predicted repeats base-pairing to form a hairpin structure is counted.