Journal of Theoretical Biology 267 (2010) 495–501

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

# A folding "framework structure" of *Tetrahymena* group I intron

Xi Zhang [a,b], Chen Guo [a,1], Wen Zhang [a,c], Huai Cao [a], Huazhen Xie [a], Kan Wang [a], Ciquan Liu [a,d,e,*]

[a] Modern Biological Research Center, Yunnan University, Kunming 650091, China
[b] Department of Food Science and Nutrition, Zhejiang University, Hangzhou 310029, China
[c] Department of Cell Biology and Genetics, Kunming Medical College, Kunming 650031, China
[d] Kunming Institute of Zoology, Chinese Academy of Science, Kunming 650223, China
[e] Center of Theoretical Biology, Peking University, Beijing 100871, China

ARTICLE INFO

*Article history:*
Received 12 March 2010
Received in revised form
16 August 2010
Accepted 1 September 2010
Available online 19 September 2010

*Keywords:*
Ribozyme
Splice site
Dynamic extended folding
RNA secondary structure

ABSTRACT

We have published the dynamic extended folding (DEF) method, which is a RNA secondary structure prediction approach—to simulate the *in vivo* RNA co-transcriptional folding process. In order to verify the reliability of the method, we selected the X-ray-determined *Tetrahymena* group I intron as a sample to construct the framework of its folding secondary structure. Our prediction coincides well with the secondary structure predicted by T.R. Cech and the X-ray diffraction crystal structure determined by Lehnert V. Our results show that the DEF framework structure of *Tetrahymena* group I intron reflects its function sites in a concise and straightforward manner, and the scope of the simulation was expanded.


## 1. Introduction

The 26S pre-ribosomal RNA intron of *Tetrahymena thermophilia* is a group I intron ribozyme (Kruger et al., 1982). Its self-splicing process is a typical two-stepped RNA splicing reaction via phosphate transferring. The process of self-splicing does not require energy, but it needs some cofactors, such as $Mg^{2+}$, exogenous guanosine (G′) or its phosphorylated derivatives (GMP, GDP, and GTP) (Grosshans and Cech, 1989; Piccirilli et al., 1993; Cech, 1990). In the first step, an exogenous G′ attacks the 5′ splice site and forms 3′,5′-phosphodiester bond. Then the 3′-OH of the 5′-exon attacks the 3′ splice site, leading to the intron release and exon ligation.

The self-splicing reaction of *Tetrahymena* group I intron depends on its secondary structure; the structural information is necessary for the reactions. The secondary structure of *Tetrahymena* group I intron is highly conserved (Cech, 1988; Lee and Gutell, 2004; Lescoute et al., 2006), and only the correct natural structure has the activity (Schneider et al., 2004). Since the special function of *Tetrahymena* group I intron was revealed by Thomas Cech et al., algorithms and experimental methods have been used to analyze its secondary structure. As early as 1983, the free energy minimization and phylogenetics were applied to predict and publish the possible secondary structure of *Tetrahymena* group I intron (Michel et al., 1983; Waring et al., 1983; Cech et al., 1983). Then the structure was determined for several times (Ye et al., 2008; Cate et al.,1996; Golden et al., 1998; Guo et al., 2004). The newest X-ray crystallography structure was determined in 2008 with a resolution of 1.95 Å (Ye et al., 2008).

The natural protein structures are framed (Kim and Baldwin, 1982; Luo and Li, 2000). For RNAs, we have treated 335 human pre-mRNA samples and spliced mature mRNAs using the described method and process of dynamic extended folding and framing. The resulting "framework structure" of the sample RNA molecules can be found in our website: http://www.mbrc.ynu.edu.cn/Human_mRna_db.html. The large-sample comparison and analysis showed that the DEF approach can efficiently simulate RNA folding process in the cell and can construct simple and general framework structure. We hope to upgrade the prediction by improving the method. Based on the previous two mRNA samples, HBZ and HBB (Cao et al., 2009), the present "framework" of *Tetrahymena* group I intron, which is rRNA, probably be more attractive.

## 2. The sample

The *Tetrahymena* group I intron sequence came from the Nucleotide Database of National Center for Biotechnology Information (NCBI), and the accession is V01416, which is 517 nt, with 413 nt of intron and 104 nt of flanking exons.

* Corresponding author at: Modern Biological Research Center, Yunnan University, Kunming 650091, China
*E-mail address:* liucq@ynu.edu.cn (C. Liu).
[1] These authors contributed equally to this work.

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.09.001

We took 427 nt as the sample, which contains 413 nt of intron, and 8 nt and 6 nt sequences of exons segments at the 5′ and 3′ ends, respectively. The selected region is comparable to the published secondary structures (Michel et al., 1983; Cech et al., 1983).

## 3. Dynamic extended folding method

In order to simulate the co-transcriptional RNA folding, we chose 30 nt as the base unit and 30 nt as the increment, until the full length was reached (1–30, 1–60, 1–90, …, 1–end nt) and $n$ segments were made. The preparatory work has been published (Zhou et al., 2005; Zhang et al., 2005; Cao et al., 2009). The 14 elongate folding units (EFUs) of the sample (427 nt) was subjected to RNAstructure 4.4 to predict the secondary structures. The common parts were searched as in Fig. 1. The theoretical variables are as the following.

Let $m$ be the number of certain hairpin or chain segment in all of the EFU predictions, and $n$ be the number of the EFUs, and obviously $m/n \leq 1$. When $m/n > 0.5$, the hairpin or chain segment could be taken as the common one. We also considered the reliability of the ratio using a set of mathematical formulas to define common hairpin and common chain segment.

The estimator of the stability of a specific hairpin or chain segment was defined as

$$\hat{P} = \frac{m}{n} \tag{1}$$

But the above formula requires $n \geq 5$. The estimated standard deviation of $\hat{P}$ was defined as

$$\hat{\sigma}(\hat{P}) = \sqrt{\frac{\hat{P}(1-\hat{P})}{n-1}} \tag{2}$$

When $n < 5$, the $m/n$ value was not creditable. So we introduced a new standard mixture-reverse-confidence (MRC, short for $c$ value). From formulas (1) and (2) the $c$ was defined as (Zhou et al., 2005)

$$c = \frac{\hat{\sigma}(\hat{P})}{\hat{P}} \tag{3}$$

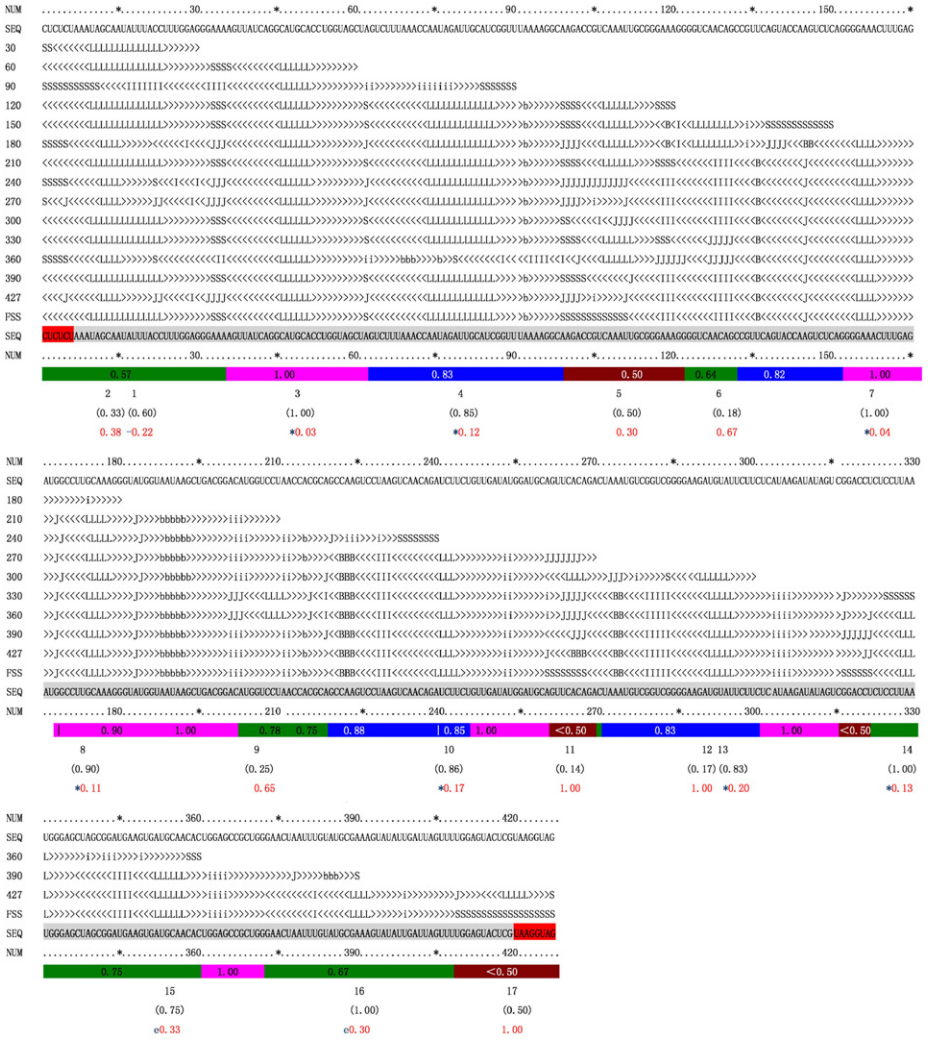

**Fig. 1.** The symbol string illustration of the dynamic extended folding simulation. The first line is the number of the nucleotides. The second line is the nucleotide sequence. The exons are in red and the introns in grey. In the following lines, 30, 60, 90,…,390, 427 shows the lengths of all the elongate folding unit (EFUs) in their symbol strings of secondary structures. In the symbol strings, L denotes terminal loop, S denotes single strand, and J denotes multi-branched loop. Because the nucleotides are from 5′ to 3′, helical stems, inner loops, and bulge loops are denoted as $<$, $>$, I, i, B, and b, respectively. For example, $<<<<<$LLLLL$>>>>>$ is a hairpin. In the last seven lines of the table, the first line of framework structure folding sequence is followed by two lines of the nucleotide sequence and numbers. And the last four lines mark the frequency of the segment conformation, the hairpin number, and its frequency and $c$ value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.) over-CH CH sub-CH FH.