Journal of Theoretical Biology 261 (2009) 93–99

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Dynamic extended folding: Modeling the RNA secondary structures during co-transcriptional folding

Huai Cao *, Hua-Zhen Xie, Wen Zhang, Kan Wang, Wei Li, Ci-Quan Liu *

Modern Biological Research Center, Yunnan University, Kunming 650091, China

ARTICLE INFO

Article history:
Received 16 September 2008
Received in revised form
8 July 2009
Accepted 15 July 2009
Available online 28 July 2009

Keywords:
Elongate folding units
Conservation hairpins
Pre-mRNA
Mature mRNA
Simulation

ABSTRACT

For RNA secondary structure prediction, it is an important issue that how to deal with co-transcriptional folding during the RNA synthesis in the cell. On one hand, co-transcriptional folding, leads to the correct final structure of the whole RNA molecule. On the other hand, it may form the recognition sites for the progress of the transcription. Considering the hurdles in the experimental determination of RNA folding structures, we proposed a so-called "dynamic extended folding simulation" approach. We used two human pre-mRNA samples, the first functional α-gene *HBZ* and the fifth β-gene *HBB*, to "display" the co-transcriptional folding images in detail. The modeling process starts from the prediction of a 30-nucleotide (nt) sequence, then in each update 30 nts was extended, say, 1–30, 1–60, 1–90, 1–120,…, 1–1651 nts (for *HBB*, 1–1606 nts). We selected the RNAstructure program to predict the folding secondary structures of all the segments. We defined "hairpin" as the unit of the secondary structure and analyzed the states of such unit during the sequential dynamic extended folding processes. We found that some hairpins are "conserved", i.e., after its appearance, it always is there in the followed foldings. Some hairpins present partially in the folding segments, and some hairpins appear for only once or twice. This phenomenon vividly depicts the generation and adjusting of the temporal structural units during the co-transcriptional folding process. It is these "hairpins" that support the thermodynamically stable structure at the end of the RNA synthesis. They may also play a role in RNA splicing process and even in the folding structure of the synthesized protein.

1. Introduction

The first step of gene expression is the transcription of RNA from DNA. During the transcription, the RNA starts to fold whenever it emerges (Boyle et al., 1980; Kramer and Mills, 1981). The folding is possibly co-transcriptional, i.e., parallel to the transcription process, the RNA folds as it grows. After the synthesis, the RNA needs to be adjusted to the final natural folding state. Experiments on RNA synthesis revealed that transcription is a directed process with adjustable rate. The 5′ terminal of the RNA is synthesized before the 3′ terminal, so in terms of time, the hydrogen bonds at the 5′ terminal are formed earlier than that at the 3′ terminal. It is the paired bases with the hydrogen bonds and the unpaired bases that formulate the RNA secondary structure units. Obviously, the secondary structure units are instantaneous or persistent, depending on the factors such as stability, forming time and competitiveness of the selective partners. The co-transcriptional folding forms the kinetic RNA secondary structure (Meyer and Miklos, 2004). Sometimes, the co-transcriptional folding forms temporary structure motifs with biological function, e.g. the viroids or the protein recognition sites in pre-mRNA transcription (Repsilber et al., 1999; Ro-Choi and Choi, 2003). In a sense, this may guarantee the correct folding of the whole RNA molecule.

Because of the importance of the RNA secondary structure units, a battery of RNA secondary structure prediction programs have been developed. Notable methods include the minimal free energy-based programs PCfold, Mfold and the follow-up version RNAstructure (Mathews et al., 1999; Zuker, 2000, 2003; Mathews et al., 2004); intelligent algorithms such as genetic algorithm (van Batenburg et al., 1995), neural networks (Steeg, 1993); statistical partition function algorithm (McCaskill, 1990), and so on. All these algorithms target on a post-transcriptional complete RNA sequence, without taking account of the importance of the co-transcriptional features, especially those functional RNA secondary structures which may form during transcription. What these algorithms get is indeed a thermodynamic RNA structure, but theoretical research implied that the thermodynamically optimal structure may not correspond with the functional structure even for a moderate length of RNA molecule (Morgan and Higgs, 1996).

* Corresponding authors. Tel.: +86 871 5033496.
E-mail addresses: caohuai@ynu.edu.cn (H. Cao), x_hz@163.com (H.-Z. Xie), zw810@hotmail.com (W. Zhang), wangkan@ynu.edu.cn (K. Wang), liwei2006@mail.ynu.edu.cn (W. Li), liucq@ynu.edu.cn (C.-Q. Liu).

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.07.027

In vitro experiments of *Tetrahymena* ribozyme suggested (Heilmann-Miller and Woodson, 2003) that the co-transcriptional folding rate is two-folds of the full-length refolding rate, and both lead to the same functional folding. This implies that full-length folding and co-transcriptional folding are correlated. For those secondary structure units formed during the co-transcriptional folding, which are conserved (or thermodynamically, stable)? As soon as they are formed at certain time point, they will present in all the succeeding folding states. Which units are relatively conserved (the basic frame is constant, the nucleotide may vary)? Which units are fluctuated (appearing from time to time)? Which are the flash in the pan (appearing for only once or twice)? Moreover, how these temporal secondary structures with different "life-span" link to their functionalities? Here functionality has two meanings: one is the correct RNA folding and the follow-up splicing and processing, another is the possible reflections during the period of protein synthesis.

In experiments, we can hardly isolate the RNA molecules at different co-transcriptional folding stages. In order to understand the "dynamic process" of the co-transcriptional folding, associated with some novel mathematical approaches and physical models or concepts, such as complexity measure factor (Xiao et al., 2005), cellular automaton (Xiao et al., 2006), low-frequency collective motions (Chou, 1988, 1989; Delarue and Dumas, 2004); solitons (Sinkala, 2006), and ensemble classifier (Shen and Zhou, 2009), which can significantly stimulate the development of biological science, in this paper we would like to introduce a novel dynamic model for simulating RNA secondary structure in transcription. Enlighten from the "sliced window" segmented folding approaches (Washio et al., 1998), we have constructed a so-called "dynamic extended folding simulation" approach. Considering the pre-mRNA of the human hemoglobin (Hb) genes as the sample, we simulated its co-transcriptional folding process in eukaryotic cells, and analyzed the folding secondary structure units in different lengths (which correspond to the real process of different periods of transcription). As expected, we found the secondary structure units with different roles in the dynamic process. These units supported the thermodynamically stable structure of the full-length RNA. It is very likely to that they play a role in RNA splicing process, as well as correlate to the protein folding structure, which is formed during the mature mRNA directed protein synthesis.

2. Method details

2.1. Interpretation of the samples

The human Hb genes are among the most representatives in the genome. They are ideal for exploring the relationships of genomic organization and function (Pauling et al., 1949; Marks et al., 1986; Goh et al., 2005; Giardine et al., 2007). Human Hb genes are clustered tightly in α and β regions. The α-globin gene region, at the short arm of chromosome 16, spans 28.3 kb and includes five functional genes and two pseudogenes. The β-globin gene region, at the short arm of chromosome 11, spans 44.7 kb and contains five functional genes and one pseudogene. These globin genes share the same structure, two introns and three exons, with the 5′ promoter sequence and the 3′ untranslated region (UTR). Our samples are the first functional α gene *HBZ* and the fifth functional β gene *HBB*.

2.2. Symbol string representation of RNA secondary structure

RNA folding is characterized as the self-folding of a single chain, and forms alternative stems and loops with partial helices and unpaired nucleotides. The secondary structure elements include: terminal loops and their helix stems (which forms simple hairpins), internal loops, bulge loops, branched loops and single chain regions. We define L = terminal loop, S = single chain, and J = branched loop. Notice that nucleotide sequence are in a direction of 5′ terminal to 3′ terminal, helix stems, internal loops, and bugle loops are marked D and d, I and i, and B and b, respectively. By setting so, RNA folding secondary structure can be mapped to the symbol string representation (Fig. 1).

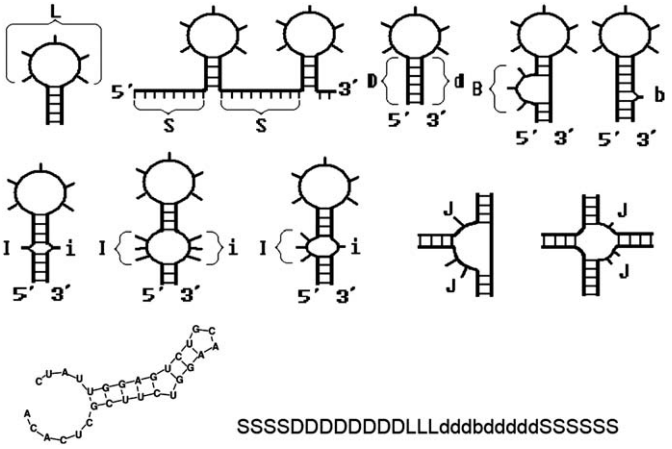

Fig. 1. RNA secondary structure elements and a symbol string representation.

2.3. Dynamic extended folding

For the full-length 1651 nucleotides (nts) of *HBZ* pre-mRNA, we started from 30 nts (i.e., 1–30 nts), and increase 30 nts each time (1–30, 1–60, 1–90, etc.), until the full-length, and finally we got 56 sequences, named as EFUs (elongate folding units). For *HBB*, there were 54 EFUs. Considering each EFU as a folding unit, we fold it with RNAstructure secondary structure prediction program and drew its secondary structure with RNAStudio program (Zhang and Liu, 2002). The hairpin structure, which is the common component and the basis of all RNA folding secondary structures, is marked by the start of the hairpin terminal loop (the site position in the full-length sequence, 1651 nts for *HBZ* and 1606 nts for *HBB*) and the unpaired bases of the hairpin terminal loop. Hairpins should have at least three base pairs (bps) and their occurrences in the EFUs are analyzed (see Fig. 2).

2.4. Defining the conservations of hairpin and chain conformation

The total number of the EFUs is denoted as n. Each EFU has a secondary structure (n structures in all). Consider the hairpins defined in Section 2.3 as the same hairpin (marked as 1, 2, …, k). The number of hairpins in all the EFUs is counted as m (k) and the following calculation is conducted (Zhang et al., 2005; Zhou et al., 2005).

$$p = \frac{m}{n} \tag{1}$$

The frequency of hairpin k is defined as: $m/n \geq 0.90$, over-conserved hairpin (over-CH); $m/n \geq 0.80$ and < 0.90, conserved hairpin (CH); $m/n \geq 0.50$ and < 0.80, sub-conserved hairpin (sub-CH); $m/n \leq 0.50$, fluctuated hairpin (FH).

In each of the EFU folding secondary structures, the conformation of an RNA strand consists of terminal loops (L), helical stems (⟨/⟩), internal loops (I/i), bugle loops (B/b), branched loops (J), and single chains (S). They can also be classified to over-