NESTED ARC-ANNOTATED SEQUENCES AND STRONG FRAGMENTS

Furletova E.[1], Roytberg M.*[1], Starikovskaya T.[2]
[1]Institute of Mathematical Problems of Biology, Pushchino, Moscow Region, Russia
[2]Lomonosov Moscow State University, Moscow, Russia
*Corresponding author: e-mail: mroytberg@impb.psn.ru

Motivation and Aim: The nested arc-annotated sequences (NAAS) represent RNA secondary structures. Informally, a NAAS is a word in the alphabet {A, U, G, C} and a set of nested arcs connecting its letters. An *optimal structure* for a given word is a NAAS on the word having maximal possible number of arcs among the NAASs. A word is *strong*, if any of its optimal structures has an arc between the first and the last characters of the fragment. The run-time bound of a dynamic programming algorithm finding an optimal structure for a given word *w* can be improved by replacement of the number of all fragments of a word w (~ n^2, where $n = n(w)$ is a length of the word *w*) by the number *F(w)* of strong fragments of the word *w* [1]. In [1] the authors claim that the average value F(w) is linear if we will consider only sequences meeting some physical restrictions. Our aim was to study the behavior of *F*(*n*) where *F*(*n*) is an average number of strong fragments of a random word w of length *n* (all letters of *w* are iid variables).

Methods and Algorithms: We have proved the following statement. Let *G(n)* be a number of strong words of length n; $g(n) = G(n)/4^n$. Then for all $n \geq 0$

$$F(n+1) = n * g(2) + (n-1) * g(3) + \ldots 1 * g(n+1) . \qquad (1)$$

In order to understand the behavior of *F*(*n*) we will study *g*(*n*). If $g(n) \geq c > 0$ for all *n*, then the function *F*(*n*) is quadratic.

Experiments and results: To investigate the average number of strong fragments of a random word we have performed Monte-Carlo computer experiments. For each $n \in \{1,\ldots, 1000\}$ we have created a set R[n] consisting of 10000 random sequences of length n. Besides this, we have calculated the experimental values of *F*(*n*) and g(*n*) over the set. The experiments show that $g(n) \approx 0.02$ for all $n \geq 30$. Using (1) and information about g(*n*) we have approximated *F*(*n*) by the formula $F^*(n) = 0.01n^2 + 0.64n - 3$ (n > 30). For all $n > 30$ we have $F(n) > F^*(n)$ and the difference $F(n) - F^*(n)$ grows monotonically for $n > 100$.

Conclusion: Formula (1) and computer experiments show that the average number *F(n)* of strong segments of a random sequence of length *n* growth ~n^2 .

References:

1. Wexler Y., Zilberstein C., Ziv-Ukelson M. A Study of Accessible Motifs and RNA Folding Complexity. Proceedings of RECOMB 2006: 473-487.