Combinatorics of RNA Hairpins and Cloverleaves

By *Michael S. Waterman*

For an abstract single-stranded RNA, a combinatorial analysis is given for two important structures, hairpins and cloverleaves. The number of possible hairpins is $O(2^n)$, while the number of possible three-petal cloverleaves is $O(n^3 2^n)$, and the number of possible general cloverleaves is $O((2.21)^n)$.

1. Introduction

Determining the shape a single-stranded RNA takes in solution is an important problem in molecular biology. The given for the problem is the primary structure or sequence of bases of the RNA. Although RNA shape has been inferred from x-ray diffraction data [3], usually the next step is to study the secondary structure of the RNA. Secondary structure is a class of two-dimensional graphs in which the sequence has formed helical regions. The final step is to infer tertiary structure of the RNA, which is the three-dimensional shape.

Although no algorithms have been proposed for prediction of tertiary structure, several have been devised for secondary structure. Most of these secondary-structure algorithms fail to search all possible graphs for the optimal structure. Recent work of the author [4] does present an algorithm to efficiently search secondary structure, but that work will not be presented here.

Instead, we explore in detail the number of the two major secondary structures: hairpins and cloverleaves. An example of each of these structures is given in Fig. 1. The problem we will address and answer is the number of each type of structure for an arbitrary-length RNA. The indicated graphs must be labeled, since these molecules have a polarity (the 3′ end to the 5′ end), and only bases of opposite polarity bond. For a real RNA, base i may or may not be able to bond with base j. In our discussion, we allow all possible bondings.

Dr. Michael S. Waterman, S-1 Mail Stop 606, Los Alamos Scientific Laboratory, P.O. Box 1663, Los Alamos, New Mexico 87545.

Copyright © 1979 by the Massachusetts Institute of Technology
Published by Elsevier North Holland, Inc. 0022-2526/79/020091+06$01.75

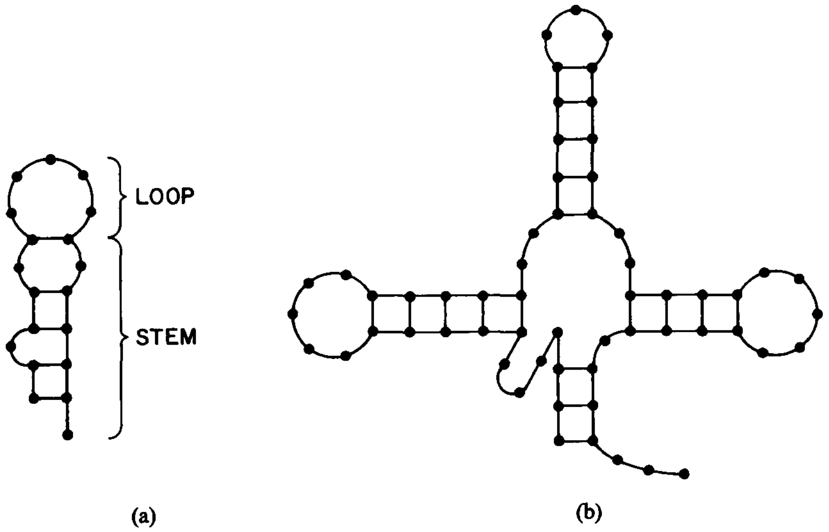

Figure 1. Secondary structure: (a) hairpin, (b) cloverleaf. The primary structure can be found by beginning at the lower left hand corner and traversing the outside of the diagrams in a clockwise direction.

This study shows that there are a large number of configurations to be searched over. Thus, a study such as [2] can be considered to give lower bounds to the amount of possible secondary structure.

2. Hairpins

In this section, we derive the number of hairpins for an RNA (or string) of length n. As Fig. 1 indicates, there are two major components to a hairpin: the loop and the stem. The only (physical) constraint we impose is that any loop must have m or more bases. Other meaningful constraints could be imposed, such as requiring all helical regions to have at least two bonds, but that will not be done here.

First of all, we prove a useful lemma on the number of stems. Let $K(a,b)$ be the number of stems with a bases on one side, b bases on the other, and the lead (or topmost) bases bound.

Lemma. *For $K(a,b)$ defined as above,*

$$K(a,b)=\binom{a+b-2}{a-1}.$$

Proof: After the lead bases are bound, there are $a-1$ bases on one side and $b-1$ bases on the other. Since the bonds cannot "cross" each other, achieving k bonds is equivalent to choosing k elements from $a-1$ and k elements from $b-1$. Thus

$$K(a,b)=\sum_{k>0}\binom{a-1}{k}\binom{b-1}{k}=\binom{a+b-2}{a-1}.$$

Next the lemma is utilized to count the hairpins.

Theorem 1. *Let $H(n)$ be the number of hairpins for a string of length n with m*

or more bases in the loop. Then

$$H(n) = 2^{n-m-1} - 1, \qquad n \geqslant m+1.$$

Proof: Let the string of length n be denoted by $1-2-3-\cdots-n$. $H(n)$ is the sum of the number of hairpins whose stem joins $1-2-\cdots-i$ and $j-(j+1)-\cdots-n$ (i and j are bonded). Of course, $j-i-1 \geqslant m$ to satisfy the loop constraint. Thus

$$\begin{aligned} H(n) &= \sum_{i=1}^{n-m-1} \sum_{j=i+m+1}^{n} \binom{n-j+i-1}{n-j} \\ &= \sum_{i=1}^{n-m-1} \sum_{l=0}^{n-i-m-1} \binom{l+i-1}{i-1} \\ &= \sum_{i=1}^{n-m-1} \binom{n-m-1}{i} = 2^{n-m-1} - 1. \end{aligned}$$

3. Cloverleaves

Next we consider the number of cloverleaves. Notice that a cloverleaf is a modified hairpin, with a stem as in a hairpin but with a "loop" that is a sequence of three hairpins joined by the primary structure. It is possible to allow more than three hairpins in the "loop." These possibilities are indicated in Fig. 2.

Due to difficulties in calculation, we set $m=1$ in this section. Below we refer to the three hairpin "loop" as a *three-petal cloverleaf* and to the more general situation as a *general cloverleaf.*

THEOREM 2. *The number of three-petal cloverleaves, $C_3(n)$, satisfies*

$$C_3(n) \sim \tfrac{1}{3} n^3 2^{n-11},$$

where $a_n \sim b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$.

Proof: Assume the loop of the cloverleaf has ν bases, and k_i is bonded to l_i, where $1 \leqslant k_1 \leqslant l_1 < k_2 \leqslant l_2 < k_3 \leqslant l_3 \leqslant \nu$. Then $H(l_i - k_i - 1) + 1$ structures can form for each of these bonds. Since $H(n)$ is of convenient form when $n \geqslant m+1 = 2$, assume $l_i - k_i \geqslant 3$. The number of loop structures, $L_3(\nu)$, is then seen to be

$$\begin{aligned} L_3(\nu) &= \sum_{l_3=12}^{\nu} \sum_{k_3=9}^{l_3-3} \sum_{l_2=8}^{k_3-1} \sum_{k_2=5}^{l_2-3} \sum_{l_1=4}^{k_2-1} \sum_{k_1=1}^{l_1-3} 2^{l_1-k_1-3} 2^{l_2-k_2-3} 2^{l_3-k_3-3} \\ &= \nu^2 2^{\nu-9} - 29\nu 2^{\nu-9} + 1092^{\nu-8} - \frac{\nu^3}{6} + 2\nu^2 - \frac{65\nu}{6} + 19. \end{aligned}$$

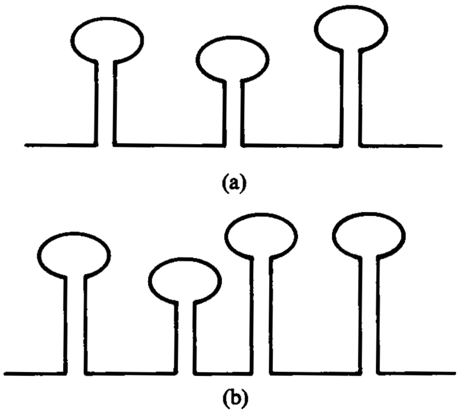

Figure 2. Cloverleaf "loops": (a) 3 hairpins, (b) more than 3 hairpins.

Now, $C_3(n)$ satisfies

$$\begin{aligned} C_3(n) &= \sum_{k_2=12}^{n-2} \sum_{k_1=0}^{n-k_2} L_3(k_2)K(k_1,n-k_1-k_2) \\ &\sim \sum_{k_2=0}^{n} \sum_{k_1=0}^{n-k_2} L_3(k_2)K(k_1,n-k_1-k_2) \\ &\sim 2^{-9} \sum_{k_2=0}^{n} \sum_{k_1=0}^{n-k_2} k_2^2 2^{+k_2} \binom{n-k_1-k_2+k_1-2}{k_1-1} \\ &= 2^{-9} \sum_{k_2=0}^{n} k_2^2 2^{k_2} 2^{n-k_2-2} \\ &= 2^{-9} 2^{n-2} \frac{n(n+1)(2n+1)}{6} \sim \tfrac{1}{3} n^3 2^{n-11}. \end{aligned}$$

To proceed to the general cloverleaf, the number of general loop structures $L_g(\nu)$ will be needed. $L_g(\nu)$ was calculated in [4] to be

$$L_g(\nu) \sim \mu^{-1}\lambda^{\nu},$$

where $\lambda = 2.2055\cdots \in (2,3)$ is a solution of $x^3 - 2x^2 - 1 = 0$ and $\mu = \lambda^{-1} + 3\lambda^{-3} + 2^{-4}[2\lambda(\lambda-2)^{-2} - 2\lambda^{-1} - 8\lambda^{-2} - 24\lambda^{-3}]$. This calculation was made by application of the renewal theorem of Feller [1, p. 330].

THEOREM 3. *The number of general cloverleaves, $C_g(n)$, satisfies*

$$C_g(n) \sim [4\mu(\lambda-2)]^{-1}\lambda^{n+1}.$$

Proof: $C_g(n)$ satisfies

$$C_g(n) = \sum_{k_2=1}^{n-2} \sum_{k_1=0}^{n-k_2} L_g(k_2)K(k_1, n-k_1-k_2)$$

$$\sim \sum_{k_2=0}^{n} \sum_{k_1=0}^{n-k_2} L_g(k_2)K(k_1, n-k_1-k_2)$$

$$\sim \sum_{k_2=0}^{n} \sum_{k_1=0}^{n-k_2} \mu^{-1}\lambda^{k_2}\binom{n-k_2-2}{k_1-1}$$

$$= \mu^{-1} \sum_{k_2=0}^{n} \lambda^{k_2}2^{n-k_2-2}$$

$$= \frac{2^n}{4\mu} \sum_{k_2=0}^{n} \left(\frac{\lambda}{2}\right)^{k_2}$$

$$= \frac{1}{4\mu}\left(\frac{\lambda^{n+1}-2^{n+1}}{\lambda-2}\right)$$

$$= \frac{1}{4\mu}(\lambda^n+\lambda^{n-1}2+\cdots+\lambda 2^{n-1}+2^n) = s_n.$$

Now, let

$$v_n = \frac{4\mu s_n}{(2\lambda)^n} = \frac{1}{2^n} + \frac{1}{\lambda}\frac{1}{2^{n-1}} + \cdots + \frac{1}{\lambda^{n-1}}\frac{1}{2} + \frac{1}{\lambda^n}.$$

Then

$$v_{n+1} = \frac{1}{\lambda}v_n + \frac{1}{2^{n+1}},$$

and, applying the renewal theorem of Feller [1, p. 330] with $b_n = 1/2^n$, $f_1 = 1/\lambda$, $f_i = 0$ for $i > 1$,

$$v_n \rightarrow \frac{\sum_{n>0} \frac{1}{2^n}}{\sum nf_n} = \lambda.$$

Therefore

$$s_n \sim \frac{1}{\mu} 2^n \lambda^{n+1},$$

and the proof is complete.

Acknowledgments

The author wishes to thank Professor T. F. Smith of Northern Michigan University for helpful discussions on this topic, and Dr. J. A. Howell of Los Alamos Scientific Laboratory for performing the symbolic computation to sum $L_3(\mu)$.

References

1. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd ed., Wiley, New York, 1968.
2. J. Gralla and C. DeLisi, m-RNA is expected to form stable secondary structure, *Nature* 248:330–332 (1974).
3. A. Rich and S. Kim, The three-dimensional structure of transfer RNA, *Sci. Amer.* 238:52–62 (1978).
4. M. Waterman, Secondary structure of single-stranded nucleic acids, Studies in Foundations and Combinatorics, *Advances in Math.; Supplementary Volume I* (1978)

Los Alamos Scientific Laboratory

(Received May 10, 1978)