Trends in Applied Sciences Research

ISSN 1819-3579

Academic
Journals Inc.

www.academicjournals.com

Trends in Applied Sciences Research 1 (3): 273-277, 2006
ISSN 1819-3579

General Combinatorics of RNA Secondary Structure

Hong Yang and Tianming Wang
Department of Applied Mathematics,
Dalian University of Technology
Dalian 116024, People's Republic of China

Abstract: For an abstract single-strand RNA, a combinatorial analysis is given for two important structures, loops and stacks. The total number of secondary structures that have given loops with minimal length m(m≥ 0 or m≥ 1) and minimal stack length l(*l*≥ 0 or *l*≥ 1) is computed, under the assumption that all base pairs can occur.

Key words: RNA secondary structure, loops, stems, recursion formula

Introduction

Determining the shape a single-strand RNA takes in solution is an important problem in computation biology, Zuker and Sank Off (1984). The given for the problem is the primary structure of the RNA. Although no algorithms have been proposed for prediction of tertiary structure, several have been devised for secondary structure. Especially, constrained secondary structures are of primary importance in biology since not every conceivable element of a secondary structure will be found in reality, (Liao and Wang, 2003; Hofacker *et al*., 1998).

Here we explore in detail the number of the two major secondary structures: loops and stacks. An example of each of these structures is given in Fig. 1. The problem we will address and answer is the total number of structures, which have exact b loops with minimal loop length m(m≥ 0 or m≥ 1) and minimal stack length *l*(*l*≥ 0 or *l*≥ 1). This sort of studies has a long history starting with the investigations of Waterman and Smith (1978a) and Waterman and Smith (1978b) who gave the first formal frame-work for the topic (Waterman, 1995). The prediction algorithms are combinatorial. Here we are concerned on the enumeration problem of RNA loops and stems. Previous results on the number of different loops of RNA molecules are due to Hofacker *et al*. (1998). They obtained some

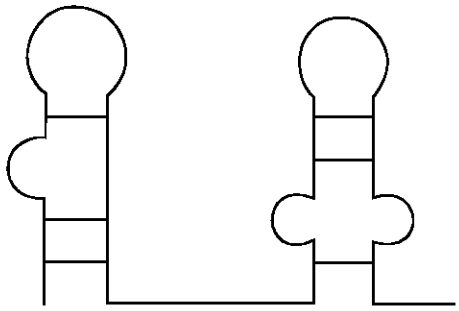

Fig. 1: There are two major secondary structures: loops and stacks

Corresponding Author: Hong Yang, Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, People's Republic of China

enumeration results for loops with degree d and b unpaired digits in the set of all secondary structures. Similarly, we will obtain some enumeration results for structures, which have given loops with minimal length m($m \geq 0$ or $m \geq 1$) and minimal stack length l($l \geq 0$ or $l \geq 1$).

In this study, we will make a further discussion about the number of constrained loops.

The Basic Definition

Definition 1

Let $R = r_1 r_2 \Lambda r_n$, $r_1 \in \{ACGU\}$, ($i = 1, 2, \Lambda\ n$) be the RNA sequence. The secondary structure is a vertex-labelled graph on n vertices with an adjacency matrix $A = (r_{ij})$ fulfilling: (1) $r_{i,j+1} = 1$, $1 \leq i \leq n-1$ (2) If $r_{i,k} = 1$, $k \neq i-1$, r_i pairs with r_k; (3) For each i there is at most a single $k \neq i-1$, $i+1$ and $|i-k| \geq m$, where $m \geq 1$, such that $r_{i,k} = 1$; (4) If $r_{i,j} = r_{k,l} = 1$ and $i<k<j$, then $i<l<j$.

Definition 2

A stack consists of subsequent base pairs

$$(p-k, q+k), (p-k+1, q+k-1), \cdots (p, q)$$

such that neither (p-k-1, q+k+1) nor (p+1, q-1) is a base pair. k+1 is the length of the stack. (p-k, q+k) is the terminal base pair of the stack.

Definition 3

The sequence I+1, I+2, Λ j-1 is a loop, if I+1, I+2, Λ j-1 are all unpaired and $a_{ij} = 1$. The pair (i,j) is said to be the foundation of the loop and the loop length is j-i.

Definition 4

A free base is a base paired with none. A matched base is a base paired with another base.

Definition 5

An external vertex is an unpaired vertex that does not belong to a loop. A collection of adjacent external vertices is called an external element.

The Recursion Formulas for Rna Secondary Structures

The number of unpaired vertices is the length of the loop. Generally, the length of loops is positive integer. Furthermore, We develop the conception of length of loops including 0.

Theorem 1

Any secondary structure Φ can be uniquely decomposed into loops and external vertices.

Proof

Each vertex that is contained in a base pair belongs to a unique stack and a stack is some loops whose length is 0. Since an unpaired vertex is either external or immediately interior to a unique base pair. The decomposition is unique: each loop is characterized uniquely by its "closing" base pair.

Theorem 2

Let $L_n(b)$ denote the number of structures on n vertices with exactly b loops whose length m is nonnegative (including 0), then

$$L_{n+1}(b) = L_n(b) + \sum_{k=m}^{n-1}\sum_{t=0}^{b-1} L_k(t)L_{n-k-1}(b-1-t) \qquad n \geq m+1$$

$$L_n(b) = 0 \quad b > 0, n \leq m+1;\ L_n(0) = 1 \qquad n \geq 0$$

Proof

The number $L_{n+1}(b)$ of structures on n+1 vertices with exactly b loops whose length is nonnegative can be computed as follows: (i) adding an unpaired vertex to any structures on n vertices, we obtain $L_n(b)$ structures with exactly b loops whose length is nonnegative; (ii) inserting an additional pair (1, k+2) we have $L_k(t)$ times all the structures with exactly b-1-t loops in the reminder of the sequence, where (l, k+2) is a loop whose length is 0. Summing over k, we can get the recursion.

Let $\Psi(n)$ be the number of structures with minimal stack length l and let $\Psi^*(n)$ be the number of structures on n vertices which have only stacks of length at least l if an additional terminal base pair is attached. Furthermore, let $\Psi^{**}(n)$ be the number of structures on n vertices which have all stacks of length at least l for which (l,n) is not a base pair. These three numbers fulfill for l>1 the coupled recursions.

$$\Psi_{n+1}(l) = \Psi_n(l) + \sum_{k=m+2l-2}^{n-1} \Psi_k^*(l)\Psi_{n-k-1}(l)$$

$$\Psi_n^*(l) = \sum_{p=l-1}^{(n-m)/2} \Psi_{n-2p}^{**}(l)$$

$$\Psi_n^{**}(l) = \Psi_n(l) - \Psi_{n-2}^*(l)$$

$$\Psi_n(l) = \Psi_{n=1}^{**}(l) = 1 \qquad n < m + 2l$$

$$\Psi_n^*(l) = 0, \qquad n < m + 2l$$

Theorem 3

Let $N_n(l)$ denote the number of loops whose length is nonnegative (including 0) on n vertices with minimal stack length l(l>0), then

$$N_{n+1}(l) = N_n(l) + \sum_{k=m+2l-2}^{n-1} \Psi_k^*(l)N_{n-k-1}(l) + \Psi_{n-k-1}(l)[N_k(l) + \Psi_k(l)] \qquad n \geq m + 2l$$

Proof

(i) Adding an unpaired base to each structure on n vertices, we obtain $N_n(l)$ loops. (ii) Inserting a base pair (l,k+2), we have $\Psi_k^*(l)$ times the number of all the loops in the reminder of the sequence plus the number of all loops within the newly formed base pair (l,k+2) times the number of structures in the tail. Summing over k, we can get the recursion.

A Method to Compute Special Structures

There are two major components: the loop and the stack for an abstract single-strand RNA. The constraint we imposed is that any structures have h loops and any stack has l(l>0) base pairs.

Theorem 4

Let $I_n(h)$ denote the number of structures with h loops and with minimal end loop length m(m>0), let $Y_n(h)$ denote the number of structures with minimal end loop length m(m>0) and given h loops that 3'and 5' ends are paired.

$$I_{n+1}(h) = I_n(h) + \sum_{k=m}^{n-1}\sum_{t=1}^{h} Y_{k+2}(t) I_{n-k-1}(h-t) \qquad h > 0, n \geq m+1 \tag{1}$$

$$I_n(0) = 1; I_n(h) = 0 \qquad h > 0, n \leq m+1$$

$$Y_n(h) = Y_{n-2}(h) + I_{n-2}(h-1) - Y_{n-2}(h-1) \tag{2}$$

$$Y_0(h) = Y_1(h) = Y_2(h) = 0, Y_n(0) = 0$$

Proof

(i) Adding an unpaired base to each structure on n vertices, we obtain $I_n(h)$; (ii) inserting a base pair (l, k+2), we have $Y_{k+2}(t)$ times all the structures with exactly h-t loops in the reminder of the sequence. Summing over k, we can get the recursion.

(2) If (2,n-1) is a pair, we obtain $Y_{n-2}(h)$; If (2n-1) is not a pair, we obtain $I_{n-2}(h-1)$-$Y_{n-2}(h-1)$.

Theorem 5

Let $I_n(h, l)$ denote the number of structures with h loops and with minimal end loop length m(m>0) and minimal stacks length l(l>0), let $Y_n(h, l)$ denote the number of structures with given h loops that 3'and 5' ends are paired and with minimal end loop length m(m>0) and minimal stack length l(l>0).

$$I_{n+1}(h, l) = I_n(h, l) + \sum_{k=m}^{n-1}\sum_{t=1}^{h} Y_{k+2}(t, l) I_{n-k-1}(h-t, l) \qquad h > 0, n \geq m+2l \tag{3}$$

$$I_n(0, l) = 1, I_n(h, l) = 0 \qquad h > 0, n \leq m+2l \tag{4}$$

$$Y_n(h, l) = Y_{n-2l}(h, l) + I_{n-2l}(h-1, l) - Y_{n-2l}(h-1, l)$$

$$Y_0(h, l) = Y_1(h, l) = Y_2(h, l) = \cdots = Y_{l+1}(h, l) = 0, Y_n(0, l) = 0$$

Proof

(I) Adding an unpaired base to each structure on n vertices, we obtain $I_n(h, l)$; (ii) inserting a base pair (l, k+2), we have $Y_{k+2}(t,l)$ times all the structures with exactly h-t loops in the reminder of the sequence. Summing over k, we can get the recursion.

(2) If (l,n), (2,n-1)Λ (l,n-l-1) is a stack whose length is l and (l+1, n-1) is a pair, we obtain $Y_{n-2l}(h,l)$; if (l,n), (2,n-1)Λ(l,n-l-1) is a stack whose length is l and (l+n, n-1) is not a pair, we obtain $I_{n-2l}(h\text{-}1,l)$-$Y_{n-2l}(h\text{-}1,l)$.

Theorem 6

Let $M_n(l)$ denote the number of loops whose length is no less than 1 with minimal stack length l, then

$$M_{n+1}(l) = M_n(l) + \sum_{k=m+2l-2}^{n-1} \{\Psi_k^*(l) M_{n-k-1}(l) + \Psi_{n-k-1}(l)[M_k(l) + \Psi_k(l)]\} - \sum_{k=m+2l-2}^{n-1} \Psi_{k-2}^*(l)\Psi_{n-k-1}(l)$$

$$n \geq m+2l$$

$$M_n(l) = 0, n < m+2$$

Proof

Because the number $M_{n+1}(l)$ of loops whose length is no less than 1 with minimal stack length l on n+1 vertices consists of all loops on n+1 vertices plus all loops in the tail times the number of structures with the newly introduced base pair plus all loops within the newly formed base pair times the number of structures in the tail. The newly formed base pair introduces an additional loop for all the $\Psi_k(l)$-$\Psi^*_{k-2}(l)$ structures in its interior without a terminal base pair.

References

Hofacker, I.L., P. Schuster and P.F. Stadler, 1998. Combinatorics of RNA secondary structures. Discrete Applied Math., 88: 207.

Liao, B. and T.M. Wang, 2003. General combinatorics of RNA hairpins and cloverleaves. J. Chem. Inf. Comput. Sci., 43: 1138.

Waterman, M.S. and T.F. Smith, 1978. Combinatorics of RNA hairpins and cloverleaves. Stud. Applied Math., 60: 91.

Waterman, M.S., 1995. Introduction to Computational Biology: Maps, Sequences and Genomes, Chapman and Hall, London.

Waterman, M.S. and T.F. Smith, 1978b. RNA secondary structure: A complete mathematical analysis, Math. Biosci., 42: 257.

Zuker, M. and D. Sank off, 1984. RNA secondary structures and their prediction Bull. Math. Biol., 46: 591.