The Enumeration of Various Types of Constrained Secondary Structure[*]

Wenwen Wang [‡§] Tianming Wang Yanchun Yang
Department of Applied Mathematics, Dalian University of Technology
Dalian 116024, P.R.China

Abstract

For the exact enumeration of various types of constrained secondary structures, this paper presents some recursion formulas and derives some properties which are based on the recursion relation about $S(n)$ given by Waterman in [7]. And a classification of secondary structure by complexity is discussed. Furthermore, we obtain some relations on RNA secondary structures of a given order.

Keywords: RNA secondary structure; Combinatorial enumeration; Recurrence relation; Planar graph

1. Introduction

Determining the shape a single-stranded RNA takes in solution is an important problem in molecular biology [10]. The primary structure of a single-stranded RNA is the sequence of nucleotides or bases making up the molecule. After the RNA primary structure was known [8], second structure has received much attention [9].

For an abstract single-stranded RNA, a combinatorial analysis is given to enumerate the number of RNA structures with certain properties [1,3]. Here we will focus on enumeration problems, which are related to the secondary structure of RNA. In these enumeration studies, the specific identities of the bases are ignored, in effect all possible base pairs are allowed. This sort of studies has a long history which started from the investigations of Waterman [2,4,5,6,7,11].

This paper introduces some basic definitions in section 2. To give the exact enumeration of various types of constrained secondary structures, in section 3 we present some recursion formulas and derive some properties which are based on the definition about $S(n)$ given by Waterman in [5]. And in section 4 a classification of secondary structure by complexity is discussed. Furthermore, some relations on RNA secondary structures of a given order are obtained.

[*] It is supported by National Science Foundation of China (10571019).

[†] Corresponding author. Tel.: +86 411 8470 8351 8408; fax: +86 411 8470 6100.

2. The basic definition

Definition 1(Waterman [6]) Let $R = r_1 r_2 \cdots r_n, r_i \in \{A, C, G, U\}, i = 1, 2, \cdots, n$, be the RNA sequence. The secondary structure is a vertex-labelled graph on n vertices with an adjacency matrix $A = (r_{ij})$ fulfilling : (1) $r_{i,i+1} = 1, 1 \leq i \leq n-1$; (2) If $r_{i,k} = 1, k \neq i-1, i+1$, r_i pairs with r_k; (3) For each i there is at most a single $k \neq i-1, i+1$ such that $r_{i,k} = 1$; (4) If $r_{i,j} = r_{k,l} = 1$ and $i < k < j$, then $i < l < j$.

We will call an edge (i, j), $|i - k| \neq 1$ a bond or a base pair. A vertex i connected only to $i-1$ and $i+1$ will be called unpaired. A vertex i is said to be interior to the base pair (k, l) if $k < i < l$. If, in addition, there is no base pair (p, q) such that $k < p < i < q < l$, we will say that i is immediately interior to the base pair (k, l).

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 A bonding loop consists of a terminal base pair and unpaired vertices. The number of unpaired vertices is the length of the bonding loop.

Definition 4 A stack $[(p, q), \cdots (p+k, q-k)]$ is called terminal if $p - 1 = 0$ or $q + 1 = n + 1$ or if the two vertices $p-1$ and $q+1$ are not interior to any base pair. The sub-structure enclosed by the terminal base pair (p, q) of a terminal stack will be called a component of the secondary structure. We will say that a structure on n vertices has a terminal base pair if $(1, n)$ is a base pair.

Definition 5 A external vertex is an unpaired vertex which dose not belong to a loop. A collection of adjacent external vertices is called an external element. If it contains the vertex 1 or n it is a free end, otherwise it is called joint.

Definition 6 A internal vertex is an unpaired vertex which is interior to a base pair.

From the combinatorial point of view, it makes perfect sense to consider the general problem with a minimum number $m (m > 0)$ of unpaired vertices in each bonding loop. We now present the recursion formulas for the exact enumeration of various types of constrained secondary structures as well as their structural elements.

3. Recurrence relations

For a secondary structure on n digits, if we add a digit $n+1$, then $n+1$ either is a free end or is paired with k. We will use the above procedure to discuss the following problems.

Lemma 1 [5]. *Let ordered set $[n] := \{1, 2, \cdots, n\}$ and S_n be the number of structures on $[n]$ which have a minimum number $m (m > 1)$ of unpaired vertices in each bonding loop. S_n satisfies the recurrence relation: $S_n = S_n + \sum_{j=1}^{m} S_{n-1-j} + \sum_{j=m+1}^{n-m-1} S_j S_{n-2-j}$, with the boundary values $S_0 = S_1 = \cdots = S_{m-1} = 0, S_m = 1$, $S_n = 0$ for $n < 0$.*

Theorem 1. *Let $N_n(b)$ denote the number of secondary structures with b stacks, and $Z_n(b)$ denote the number of secondary structures with b stacks given that the $3'$ and $5'$ ends are paired,* *then*

$$N_{n+1}(b) = N_n(b) + \sum_{k=1}^{m+2} Z_{n-k+2}(b) + \sum_{k=m+3}^{n-m} \sum_{l=0}^{b} N_{k-1}(b-l) Z_{n-k+2}(l), \ n \geq m+1;$$

$$N_n(0) = 1, \ N_n(b) = 0, \ b > 0, \ n \leq m+1. \qquad (1)$$

Proof. Now we consider the sequence $[1, n+1]$, and there are two cases to be considered: $n+1$ either is a free base or is paired with k, where $1 \leq k \leq n-1$. If $n+1$ is unpaired, there are $N_n(b)$ secondary structures with b stacks. Otherwise, three subcases are introduced: if $1 \leq k \leq m+2$, the number of satisfied structures is $Z_{n-k+2}(b)$ in $[k, n+1]$,; if $m+2 < k \leq n-m$, the number is $\sum_{l=0}^{b} Z_{n-k+2} N_{k-1}(b-l)$; if $n-m < k \leq n-1$, the structure is unsatisfied. This complete the proof.

The auxiliary variable $Z_n(b)$ satisfies the recurrence

$$Z_n(b) = Z_{n-2}(b) + N_{n-2}(b-1) - Z_{n-2}(b-1), Z_0(b) = Z_1(b) = 0, Z_n(0) = 0, n \geq 0. \qquad (2)$$

We can consider the sequence $[1, n]$. Of course, 1 is paired with n. If 2 is paired with $n-1$, the number is $Z_{n-2}(b)$; otherwise, there are $N_{n-2}(b-1) - Z_{n-2}(b-1)$ structures. And it is clear that $Z_n(1) = [\frac{n-m}{2}], n \geq m+1. \qquad (3)$

Corollary 1. $N_n(1) = \frac{1}{4}[\frac{1}{6}n(n-1)(2n-1) - [\frac{n}{2}]]$, $n \geq m+1$.

Proof. By Theorem 1, let $b = 1, m = 1$, we can get the following relation

$$N_{n+1}(1) = N_n(1) + \sum_{k=m}^{n-1} Z_{k+2}(1), \qquad (4)$$

According to (1)(2)(3)(4), we obtain

where

$$N_{n+1}(1) - N_n(1) = \begin{cases} \frac{1}{4}(n^2-1), & \textit{if } n \textit{ is odd;} \\ \frac{1}{4}n^2, & \textit{if } n \textit{ is even.} \end{cases}$$

Furthermore, we can get $N_n(1) = \frac{1}{4}[\frac{1}{6}n(n-1)(2n-1) - [\frac{n}{2}]], n \geq 2$.

The proofs of the following Theorems are all similar to Theorem 1.

Theorem 2. *Let $J_n(b)$ denote the number of structures on n vertices with exactly b components,then*

$$J_{n+1}(b) = J_n(b) + \sum_{k=b(m+2)-(m+1)}^{n-m} S_{n-k} J_{k-1}(b-1), \ n \geq b(m+2);$$

$$J_{b(m+2)}(b) = 1, \ \textit{if } n \leq b(m+2)-1, \textit{then } J_n(b) = 0, \ \textit{for } b > 0, J_n(0) = 1.$$

Theorem 3. *Let $A_n(b)$ be the number of structures with exactly b hairpins, then it satisfies the recurrence*

$$A_{n+1}(b) = A_n(b) + \sum_{k=1}^{m+2} A_{n-k}(b) + \sum_{k=m+3}^{n-m} \sum_{l=0}^{b} A_{n-k}(l) A_{k-1}(b-l), \ n \geq m+1;$$

$$A_n(b) = \delta_{0,b}, n \leq m+1, A_n(0) = 1 \ for \ all \ n, A_n(b) = 0, b > 0 \ for \ n \leq m+1.$$

Corollary 2. $A_n(1) = J_n(1) = 2^{n-m-1} - 1$, $n \geq m+1$.

Proof. By Theorem 3, let $b = 1$, then $A_{n+1}(1)$ satisfies the recurrences, $A_{n+1}(1) = \sum A_k(1) + n - m$, i.e., $A_{n+1}(1) = 2A_n(1) + 1$. The result is clear. For the same analysis, we know $A_n(1) = J_n(1)$. This complete the proof.

Theorem 4. *Let V_n denote the total number of internal vertices, and let U_n be the total number of unpaired bases, then*

$$V_{n+1} = V_n + \sum_{k=1}^{m} U_{n-k} + \sum_{k=m+1}^{n-m} [S_{n-k} V_{k-1} + S_{k-1} U_{n-k}], \ n \geq m+1; V_n = 0, n \leq m+1, V_0 = 0.$$

Theorem 5. *Let U_{n+1} denote the total number of unpaired bases, then*

$$U_{n+1} = U_n + S_n + \sum_{k=1}^{m} [(k-1) S_{n-k} + U_{n-k}] + \sum_{k=m+1}^{n-m} [S_{n-k} U_{k-1} + S_{k-1} U_{n-k}], \ n \geq m+1;$$
$$U_n = n, n \leq m+1, U_0 = 0.$$

By Definition 5 and 6, the total number of external vertices is denoted by E_n. It is clear that $V_n + E_n = U_n$. For sake of completeness, we state the relation for E_n,

$$E_{n+1} = E_n + S_n + \sum_{k=1}^{m} (k-1) S_{n-k} + \sum_{k=m+1}^{n-m} S_{n-k} E_{k-1}, \ n \geq m+1;$$
$$E_n = n, n \leq m+1, E_0 = 0.$$

Theorem 6. *Let P_{n+1} denote the total number of base pairs, then*

$$P_{n+1} = P_n + \sum_{k=1}^{m} [P_{n-k} + S_{n-k}] + \sum_{k=m+1}^{n-m} [S_{n-k} P_{k-1} + S_{k-1}(P_{n-k} + S_{n-k})], n \geq m+1;$$
$$P_n = 0, n \leq m+1.$$

By Th5 and Th6, we can easily obtain the following relation: $U_n + 2P_n = nS_n$.

Theorem 7. *Let I_{n+1} denote the total number of components, then*

$$I_{n+1} = I_n + \sum_{k=1}^{m} S_{n-k} + \sum_{k=m+1}^{n-m} S_{n-k} [I_{k-1} + S_{k-1}], n \geq m+1; I_n = 0, n \leq m+1.$$

Theorem 8. *Let N_{n+1} denote the total number of stacks, and Z_n be the total number of secondary structures given that $3'$ and $5'$ ends are paired, then*

$$N_{n+1} = N_n + \sum_{k=1}^{m} Z_{n-k+2} + \sum_{k=m+1}^{n-m} [S_{n-k}N_{k-1} + S_{k-1}Z_{n-k+2}], n \geq m+1; N_n = 0, n \leq m+1.$$

For the auxiliary variable, we find

$$Z_{n+2} = N_n + S_n - S_{n-2}, n \geq m+1; Z_m = Z_{m+1} = 1, Z_n = 0, n < m.$$

4. Secondary structures of a given order

Secondary structures are classified by a certain complexity criterion. A simple lemma is necessary to make certain this definition.

Lemma 2 [6]. *If $A = (a_{ij})$ is the adjacency matrix for some secondary structure and, if $A' = (a'_{ij})$ is formed from $A = (a_{ij})$ by setting $a'_{ij} = a'_{ji} = 0$ for any set of choices of i and j $(i \neq j \pm 1)$, then A' is the adjacency matrix for another secondary structure.*

Definition 7 Let $A = (a_{ij})$ be the adjacency matrix for a secondary structure. A sequence $A^{(i)}$ of adjacency matrices of secondary structure is formed as follows:

(i) $A^{(0)} = A$.

(ii) From $A^{(i+1)}$ from $A^{(i)}$ by setting $a_{kl}^{(i+1)} = a_{lk}^{(i+1)} = 0$ whenever $a_{kl}^{(i)} = a_{lk}^{(i)} = 1$, k and l are members of some hairpin, and $k \neq l \pm 1$.

The secondary structure for A is said to be ωth order if $A^{(\omega)}$ is the first matrix in the sequence $\{A^{(\omega)}\}_{\omega=0}^{\infty}$ such that the secondary structure for $A^{(\omega)}$ has no hairpins.

Of course, the open structure has order $\omega = 0$ and any structure without a multiloop has order $\omega = 1$. A bulge or interior can not change the order of secondary structure, but a multiloop.

In [1], the number of secondary structures with c components and order ω is discussed. Now we will give a further discussion about the constrained secondary structure with a given order.

Let $S_n(k, \omega)$ be the number of secondary structures with k paired bases and order ω. Furthermore, let $S_n^*(k, \omega)$ be the number which yield a structures of order ω and k paired bases when enclosed by an additional base pair.

Theorem 9. *$S_n(k, \omega)$ satisfies the recursion*

$$S_{n+1}(k, \omega) = S_n(k, \omega) + \sum_{i=1}^{m} S_{n-i}^*(k-1, \omega) + \sum_{i=m+1}^{n-m} \sum_{j=1}^{k-1} \{S_{n-i}^*(j, \omega) \sum_{i=0}^{\omega} S_{i-1}(k-j-1, l) +$$

$$S_{i-1}(k-j-1, \omega) \sum_{l=0}^{\omega-1} S_{n-i}^*(j, l)\}; S_n(0,0) = 1,\ S_n(0, \omega) = S_n(k, 0) = 0,\ n \leq m+1.$$

Adding a terminal base pair to a sequence doesn't change the number of the structure, so we get the relation $S_n(k,\omega) = S^*_n(k,\omega)$.

Theorem 10. *Let $\tilde{S}_n(\omega)$ be the total number of secondary structures with order ω, then $\tilde{S}_n(\omega)$ satisfy the recursion relation:*

$$\tilde{S}_{n+1}(\omega) = \tilde{S}_n(\omega) + \sum_{k=1}^{m} \tilde{S}_{n-k}(\omega) + \sum_{k=m+1}^{n-m} \{\tilde{S}_{n-k}(\omega) \sum_{l=0}^{\omega} \tilde{S}_{k-1}(l) + \tilde{S}_{k-1}(\omega) \sum_{l=0}^{\omega-1} \tilde{S}_{n-k}(l)\};$$

$$\tilde{S}_{n+1}(0) = 1;\ \tilde{S}_{n+1}(\omega) = 0 \textit{ for } \omega \geq 1, n \leq m+1.$$

Let $N_n(b,\omega)$ be the number of secondary structures with b stacks and order ω. The auxiliary variable $Z_n(b,\omega)$ denote the number of secondary structures with exactly b stacks and order ω given that the $3'$ and $5'$ ends are paired.

Theorem 11. *The numbers $N_n(b,\omega)$ satisfy the recursion:*

$$N_{n+1}(b,\omega) = N_n(b,\omega) + \sum_{k=1}^{m+2} Z_{n-k+2}(b,\omega) + \sum_{k=m+3}^{n-m} \sum_{i=0}^{b} \{N_{k-1}(b-i,\omega) \sum_{l=0}^{\omega} Z_{n-k+2}(i,l)$$
$$+ Z_{n-k+2}(i,\omega) \sum_{l=0}^{\omega-1} N_{k-1}(b-i,l)\}; N_n(0,0) = 1,\ N_n(0,\omega) = N_n(b,0) = 0.$$

For the auxiliary variable $Z_n(b,\omega)$ recursion is

$$Z_n(b,\omega) = Z_{n-2}(b,\omega) + N_{n-2}(b-1,\omega) - Z_{n-2}(b-1,\omega),\ n \geq m+1.$$

The proof is similar with (2) in Theorem 1.

References

[1] I.L. Hofacker, P. Schuster, P.F. Standler, Combinatorics of RNA Secondary Structures, Disc. Appl. Math. 88 (1998) 207.

[2] J.A. Howell, T.F. Smith, M.S. Waterman, Computation of generating function for biological molecules, SIAM J. Appl. Math. 39 (1980) 119.

[3] B. Liao, T.M. Wang, General Combinatorics of RNA Secondary Structure Mathematical Biosciences, 191 (2004) 69.

[4] W.R. Schmitt, M.S. Waterman, Liner trees and RNA secondary structure, Discr. Appl. Math. 12 (1994) 412.

[5] P.R. Stein, M.S. Waterman, On some new sequences generalizing the Catalan and Motzkin numbers, Disc. Math. 26 (1978) 261.

[6] M.S. Waterman, Secondary structure of single-stranded nucleic acids, Adv. Math. Suppl. Stud. 1 (1978) 167.

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

[8] R.R. Holley, J. Apgar, G.A. Everett, J.T. Madison, M. Marquisee, S.H. Merrill, J.R. Penswich, and A. Zamer, Structure of a ribonucleic acid, Science, 147 (1965).

[9] V.A. Bloomfield, D.M. Crothers, and I. Tinoco, Physical Chemistry of Nucleic Acids, Harper, New York, 1974.

[10] M. Zuker, D. Sank off, RNA secondary structures and their prediction, Bull.Math.Biol. 46(4) (1984) 591.

[11] R.C. Penner, M.S. Waterman, Spaces of RNA secondary structures, Adv. Math. 101 (1993) 31.