ELSEVIER

Discrete Applied Mathematics 51 (1994) 317–323

DISCRETE APPLIED MATHEMATICS

Linear trees and RNA secondary structure[†,‡]

William R. Schmitt*[,a], Michael S. Waterman[b]

[a]*University of Memphis, Memphis, TN 38152, USA*
[b]*University of Southern California, Los Angeles, CA 90089-1113, USA*

(Received 3 July 1991; revised 3 March 1992)

Abstract

The total number of RNA secondary structures of a given length with a fixed number of base pairs is computed, under the assumption that all base pairs can occur. This is done by establishing a one-to-one correspondence between secondary structures and trees. A duality operator on trees is presented, which explains a symmetry in the numbers counting secondary structures.

1. Introduction

The now classical helical structure of DNA was proposed in 1953 by Watson and Crick. In their model, DNA is double-stranded, where each strand is specified by the sequence of bases attached to a sugar–phosphate backbone. One strand determines the other by the so-called Watson–Crick rules: A (adenine) forms base pairs with T (thymine), and G (guanine) forms base pairs with C (cytosine). These base pairs are formed by hydrogen bonds, two hydrogen bonds forming an A–T pair and three forming a G–C pair. In this paper we study helical structures of another nucleic acid, single-stranded RNA.

RNA is described by its sequence of bases, which are the same as for DNA except that T is replaced by U (uracil), and in RNA, A pairs with U. These single-stranded molecules fold onto themselves, forming helical structures. The sequence of bases of the RNA molecule is known as *primary structure*, and it is determined experimentally. A subset of helical structure consistent with a planar graph is known as *secondary structure* and is carefully defined in Section 2. The complete three-dimensional structure of the RNA molecule is known as *tertiary structure*. Secondary and tertiary

* Corresponding author.
† Research supported by the National Institutes of Health and the National Scientific Foundation.
‡ This work is dedicated to the memory of Paul R. Stein.

SSDI 0166-218X(92)00038-F

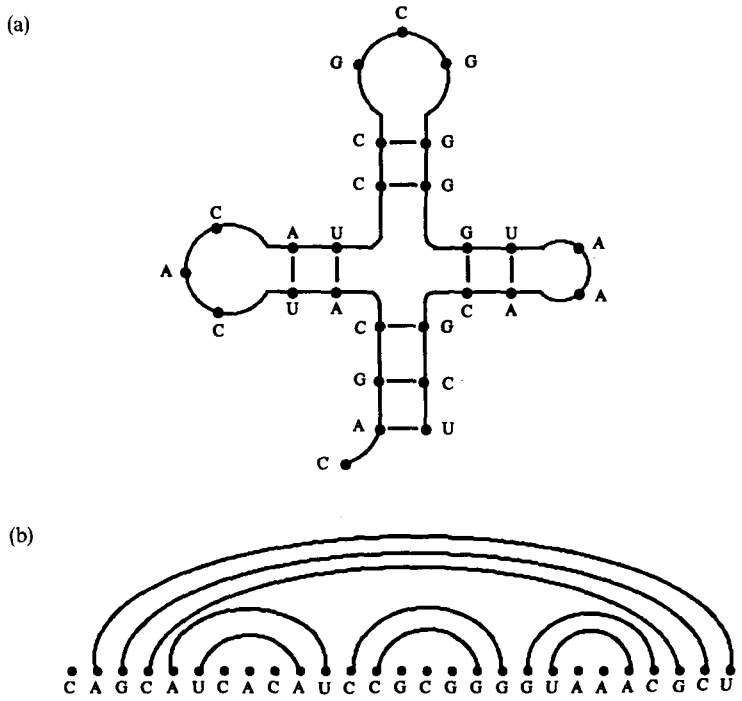

Fig. 1. Two representations of secondary structure.

structures determine the shape and hence the function of these important biological molecules. The structure of RNA is utilized in regulating the expression of genes, in the assembly of protein molecules, and in many other fundamental biological processes. For an excellent general reference to molecular biology, see Lewin [3].

As an example of secondary structure, we consider the sequence

CAGCAUCACAUCCGCGGGGUAAACGCU

and give one secondary structure for this sequence in Fig. 1(a), where the base pairs are indicated by dashes. This structure is referred to in the biological literature as a cloverleaf and is the secondary structure assumed by transfer RNA molecules. In Fig. 1(b), the same structure is presented, where the primary structure is given along the horizontal axis and the base pairs are shown as arcs. This representation is used by Stein [4].

The prediction of the shapes of biological molecules is an important topic in computational biology; see [9] for a review. In [8] the problem of prediction is shown to be polynomial. The prediction algorithms are combinatorial, and the enumeration of secondary structure is a natural problem; see [5–7]. In these enumeration studies, the specific identity of the bases is ignored, which in effect allows all possible base pairs. In this way, the entire set of possible RNA secondary structures is studied.

In this paper, we compute the total number of secondary structures of a given length, under the assumption that all base pairs can occur. This is done by establishing a one-to-one correspondence between the set of all secondary structures of a specific length with a fixed number of base pairs, and a particular set of plane trees which is easily counted. We also present a duality operator on these sets of trees which explains a symmetry that we observed in the numbers which count secondary structures. This work was initiated with Paul Stein, whose untimely death prevented him from seeing its conclusion.

2. Enumeration

Our model for an RNA sequence is the ordered set $[n] = \{1,2, \ldots, n\}$. A *secondary structure* on $[n]$ is a simple graph on $[n]$, i.e. a set B of (unordered) pairs of elements of $[n]$, satisfying

(a) $\deg i \leqslant 3$ for $1 \leqslant i \leqslant n$,

(b) if $(i,j) \in B$, then $|i-j| \geqslant 2$,

(c) if $(i,j),(k,l) \in B$, where $i<j$ and $k<l$, and $[i,j] \cap [k,l] \neq \emptyset$ then either $[i,j] \subset [k,l]$ or $[k,l] \subset [i,j]$ (where $[i,j]$ denotes the interval $\{r: i \leqslant r \leqslant j\}$).

Let $s(n)$ be the total number of secondary structures on $[n]$. Howell et al. [2] show that these numbers satisfy the Catalan-like recurrence relation

$$s(n+1) = s(n) + \sum_{j=1}^{n-1} s(j-1)s(n-j), \tag{1}$$

for $n \geqslant 2$, and $s(0) = s(1) = s(2) = 1$. Asymptotic formulae for $s(n)$ and related sequences determined by a recurrence relation generalizing the above are given by Stein and Waterman [5].

Let $\mathscr{S}_{n,k}$ be the set of secondary structures on $[n]$ which have exactly k pairs, and let $s(n,k) = |\mathscr{S}_{n,k}|$. For example, the set $\mathscr{S}_{6,2}$ is shown in Fig. 2.

The numbers $s(n,k)$ are also studied in [2], where it is shown that they satisfy the recurrence

$$s(n+1,k+1) = s(n,k+1) + \sum_{j=1}^{n-1} \left[\sum_{i=0}^{k} s(j-1,i)\, s(n-j,k-i) \right], \tag{2}$$

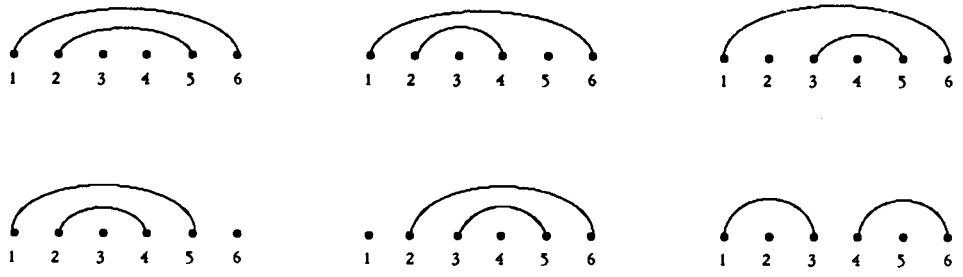

Fig. 2. The set $\mathscr{S}_{6,2}$.

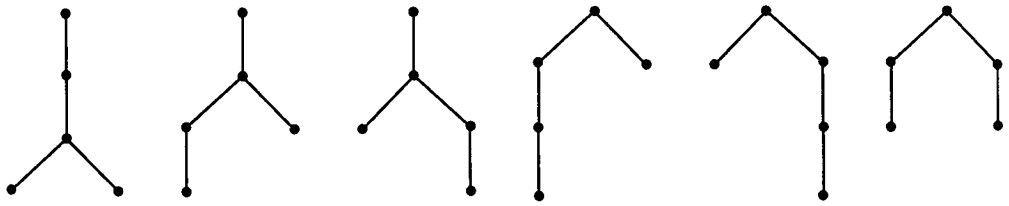

Fig. 3. The set $\mathscr{T}_{5,3}$.

for $n \geqslant 2$, where $s(n,0) = 1$, for all n, and $s(n,k) = 0$ for $k \geqslant \frac{1}{2}n$. Using vector notation, this recurrence can be written in a form identical to that of Eq. (1). Hence, the sequences $\{s(n,k): k \geqslant 0\}$ can be viewed as vector generalizations of Catalan numbers.

It turns out that the numbers $s(n,k)$ are much easier to compute than the $s(n)$. Our main result is a remarkably simple formula for $s(n,k)$, which is obtained by constructing a bijection from $\mathscr{S}_{n,k}$ onto a certain set of trees.

A *linear tree* is a rooted tree together with a linear ordering on the set of children of each vertex in the tree. An *isomorphism* of linear trees is a root-preserving tree isomorphism which also preserves linear orders. An *unlabelled* linear tree is an isomorphism class of linear trees. Let $\mathscr{T}_{n,k}$ be the set of unlabelled linear trees having n vertices, k of which are nonterminal, and let $t(n,k) = |\mathscr{T}_{n,k}|$. The set $\mathscr{T}_{5,3}$ is shown in Fig. 3, with the roots at the tops of the trees.

Linear trees differ from rooted planar trees in that the children of the root in a linear tree are linearly ordered (from left to right), while the children of the root in a rooted planar tree are merely cyclically ordered. For example, the fourth and fifth trees shown in Fig. 3 are equal as rooted planar trees, but differ as linear trees.

Proposition 2.1. *For all $n,k \geqslant 1$, there exists a bijection $\phi: \mathscr{S}_{n+k-2,k-1} \to \mathscr{T}_{n,k}$.*

Proof. Suppose $n,k \geqslant 1$, and B is a secondary structure on $[n+k-2]$ having $k-1$ pairs. Let $F \subseteq [n+k-2]$ be the set of unpaired elements of $[n+k-2]$ (note that $|F| = n-k$). Let V be the set $\{[i,j]:(i,j) \in B\} \cup F \cup \{*\}$. Partially order V by letting $*$ be maximal, and otherwise ordering by set inclusion, where we identify $k \in F$ with the set $\{k\}$. The Hasse diagram of this poset is a rooted tree, with root $*$, having a total of n vertices, $n-k$ of which are terminal. The linear order of the set F of terminal vertices gives this tree a linear structure. Letting $\phi(B)$ be the isomorphism class of this linear tree, we have $\phi(B) \in \mathscr{T}_{n,k}$. For example, ϕ maps the secondary structure shown in Fig. 4(a) to the unlabelled linear tree given in Fig. 4(b).

Suppose T_1 is a linear tree having isomorphism class $T \in \mathscr{T}_{n,k}$. Search T_1 by (leftmost) depth-first search, labelling vertices as they are encountered by consecutive integers (starting with 0 at the root), but label internal vertices only when they are first encountered and last encountered. The resulting pairs of labels on the (nonroot) internal vertices are the pairs of a secondary structure on $[n+k-2]$. This secondary structure is clearly independent of the particular choice of T_1 representing T; hence, we denote it by $\alpha(T)$. For example, the linear tree shown in Fig. 5 gives rise to the set of pairs $\{(1,12),(2,7),(4,6),(9,11)\}$, which is the secondary structure of Fig. 4.

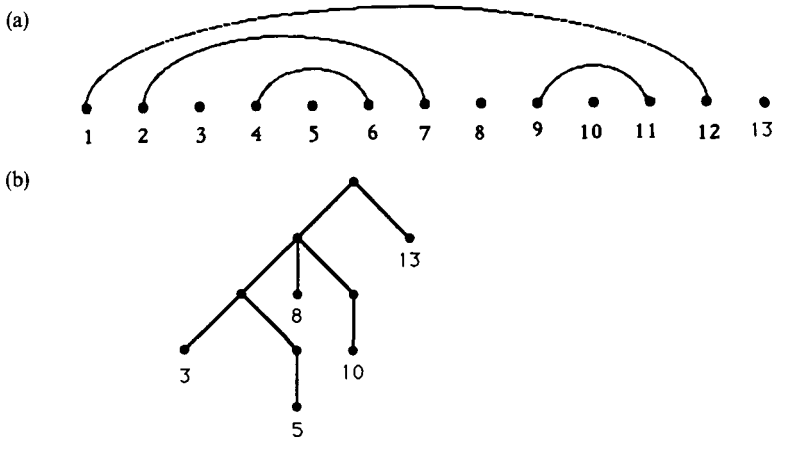

Fig. 4. A secondary structure and its corresponding tree.

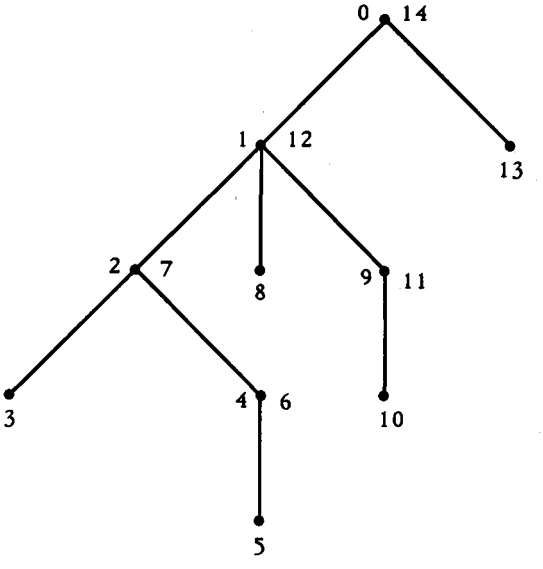

Fig. 5. Labelling of a linear tree, determining corresponding secondary structure.

It is straightforward to verify that the maps ϕ and α are inverses of one another. □

Theorem 2.2. *The number of secondary structures over a sequence of length n having exactly k pairs is given by*

$$s(n,k) = \frac{1}{k}\binom{n-k}{k+1}\binom{n-k-1}{k-1}$$

for $n, k \geqslant 0$.

Proof. It is a standard exercise in generating functions to show that the numbers $t(n,k)$ are given by

$$t(n,k) = \frac{1}{k-1}\binom{n-1}{k}\binom{n-2}{k-2}.$$

Also, a bijective proof of this result was recently given by Chen [1]. Thus, the theorem follows from the identity $s(n,k) = t(n-k+1,k+1)$, implied by Proposition 2.1. □

It is well known that the total number of linear trees on n vertices $t(n) = \sum_k t(n,k)$ is equal to the Catalan number

$$n^{-1}\binom{2n-2}{n-1}.$$

Hence, the sequences of numbers $s(n+k-2,k-1) = t(n,k)$, for all fixed n, are, in a sense, vector generalizations of Catalan numbers.

3. Symmetry

The symmetry $s(n+k,k) = s(2n-k-1,n-k-1)$, which was originally observed in data computed from the recurrence relation (2), is easily verified using Theorem 2.2. By the correspondence of Proposition 2.1, this symmetry can be written as $t(n,k) = t(n,n-k)$, suggesting the existence of a kind of duality operator on the set $\mathscr{T}_n$ of all unlabelled linear trees on n vertices.

Proposition 3.1. *For all n, there is an involution $T \to T^*$ of the set $\mathscr{T}_n$, which maps $\mathscr{T}_{n,k}$ bijectively onto the set $\mathscr{T}_{n,n-k}$, for $1 \leqslant k \leqslant n-1$.*

Proof. Define a *linear partition* of a set S to be a partition π of S, together with a linear ordering of each block of π. Let T_1 be a linear tree representing $T \in \mathscr{T}_n$, where T_1 has vertex set V, terminal vertex set F, and edge set E, and suppose all edges of T_1 are directed away from the root. For each $v \in V - F$, let B_v be the set of all edges directed out of v. B_v is linearly ordered, from left to right, and the set $\beta(T_1) = \{B_v : v \in V - F\}$ is a linear partition of E (into *bushes*). For all $e \in E$, let $l(e) \in F$ be the leftmost terminal vertex which can be reached by a directed path through e. For each $w \in F$, let $P_w = \{e \in E : l(e) = w\}$. P_w is linearly ordered ($e < f$ means that f is farther from the root than e), and the set $\pi(T_1) = \{P_w : w \in F\}$ is a linear partition of E (into *leftmost paths*).

The dual T^* of the unlabelled tree T is defined to be the isomorphism class of a linear tree T_2 having edge-set E which satisfies $\beta(T_2) = \pi(T_1)$ and $\pi(T_2) = \beta(T_1)$, as linear partitions. It is easy to check that such T_2 exists, and is determined uniquely, up to vertex labelling, by these conditions. It follows immediately from the definition that $(T^*)^* = T$, and that $T \in \mathscr{T}_{n,k}$ if and only if $T^* \in \mathscr{T}_{n,n-k}$. □

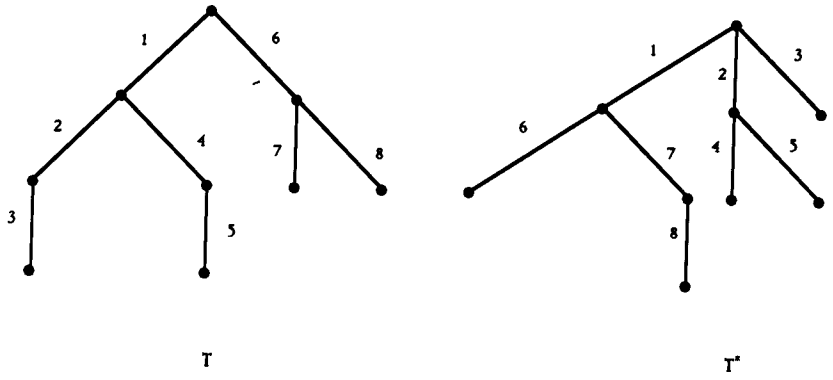

Fig. 6. A linear tree and its dual.

An example of a linear tree T and its dual T^* is shown in Fig. 6 with edge-labellings. In this example, $E = [8]$, $\pi(T) = \{\{1,2,3\},\{4,5\},\{6,7\},\{8\}\} = \beta(T^*)$ and $\beta(T) = \{\{1,6\},\{2,4\},\{7,8\},\{3\},\{5\}\} = \pi(T^*)$.

References

[1] W.Y.C. Chen, A general bijective algorithm for trees, Proc. Nat. Acad. Sci. USA 87 (1990) 9635–9639.
[2] J.A.Howell, T.F. Smith and M.S. Waterman, Computation of generating functions for biological molecules, SIAM J. Appl. Math. 39 (1980) 119–133.
[3] B. Lewin, Genes IV (Oxford Univ. Press, Oxford, 1990).
[4] P.R. Stein, On a class of linked diagrams, I. Enumeration, J. Combin. Theory A 24 (1978) 357–366.
[5] P.R. Stein and M.S. Waterman, On some new sequences generalizing the catalan and Motzkin numbers, Discrete Math. 26 (1978) 261–272.
[6] M.S. Waterman, Secondary structure of single-stranded nucleic acids, Adv. Math., I (suppl.) (1978) 167–212.
[7] M.S. Waterman, Combinatorics of RNA hairpins and cloverleaves, Stud. Appl. Math. 60 (1979) 91–96.
[8] M.S. Waterman and T.F. Smith, Rapid dynamic programming algorithms for RNA Secondary structure, Adv. Appl. Math. 7 (1986) 455–464.
[9] M. Zuker and D. Sankoff, RNA Structures and their prediction, Bull. Math. Biol. 46 (1984) 591–621.