RNA secondary structure, permutations, and statistics

Robert Willenbring

Department of Mathematics, Michigan State University, United States

ARTICLE INFO

ABSTRACT

We construct a permutation representation for RNA secondary structure. We also introduce some basic combinatorial statistics for RNA secondary structure and relate them to permutation statistics when appropriate. These statistics allow us to quantify some structural phenomena in RNA secondary structure.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

RNA is a single stranded nucleic acid composed of four nitrogen bases: adenine, guanine, cytosine, and uracil. The primary structure of an RNA molecule is the order in which these bases occur. Additionally, adenine bases may form hydrogen bonds with uracil bases, and guanine bases may form hydrogen bonds with cytosine bases. However, adjacent bases never form hydrogen bonds. The secondary structure of an RNA molecule identifies the positions of bases that form hydrogen bonds.

RNA secondary structure was first discussed from a graph theoretic point of view by Waterman [1]. This representation (Fig. 1) is clearly equivalent to the non-crossing set partition representation (Fig. 1, right) he uses in later articles [2]. We will use the non-crossing set partition as our basic representation. We note that a non-crossing set partition represents an RNA secondary structure under the Waterman definition if all its blocks are of size 2 or 1 and if the minimum distance between numbers in the same block is 2.

We denote the set of non-crossing set partitions that represent RNA secondary structures by SS_n, where n is the number of bases. We use $SS_{n,k}$ to denote the set of all RNA secondary structures (RNA SS) with n bases and k bonds, where a bond is a block of size two. We refer to the smaller number in a bond as a *left bond*, the larger number in a bond as a *right bond*, and bases in blocks of size one as *unpaired bases*.

In our discussion of statistics on RNA SS, we will differentiate between single and multiple hairpin RNA SS. *Single Hairpin* RNA SS are those RNA SS with the property that every right bond is greater than any of its left bonds. It is known that the total number of single hairpin RNA SS on n bases is $2^{n-2} - 1$ [3]. All other RNA SS have multiloops or multiple components and we refer to these RNA SS as *multiple hairpin* RNA SS.

The permutation model for RNA SS presented here was developed while studying two different combinatorial representations of RNA SS. One is the Schmitt and Waterman linear tree model, which is produced by a bijection between $SS_{n,k}$ and the set of linear trees $LT_{n-k+1,k+1}$ [2]. A linear tree in $LT_{n-k+1,k+1}$ has $n - k + 1$ vertices, $k + 1$ of which are non-terminal. The other is Nkwanta's lattice path model, which associates a left bond with a north edge, a right bond with a south edge, and an unpaired base with an east edge [4].

Attempts to make the number of vertices in the Schmitt/Waterman model dependent on the single parameter n eventually produced a lattice path model almost identical to Nkwanta's, but with only north and east edges. In this model, a
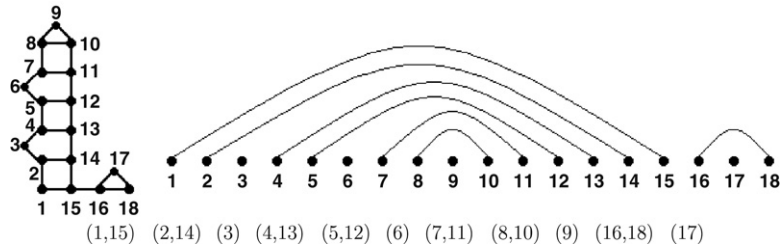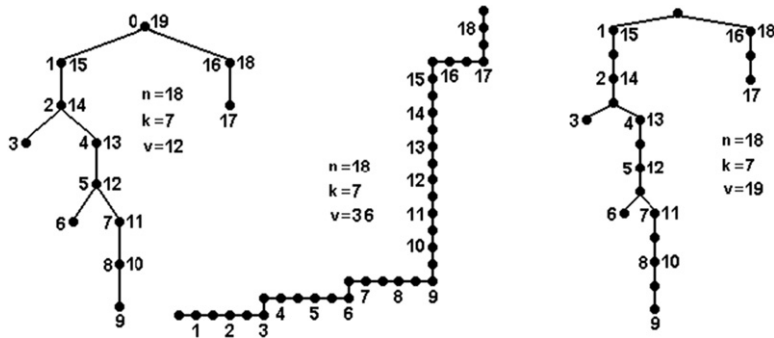
Fig. 1.

Fig. 2.

left bond is associated with two east edges, a right bond by two north edges, and an unpaired base by an east edge followed by a north edge. The set of lattice paths produced by this construction, $LP_{n,k}$, forms a subset of the set of all lattice paths, which is counted by the Catalan number, C_n. Adaptations of bijections described by Stanton and White [5] lead to other restricted Catalan models for RNA secondary structure using standard Young tableaux, integer sequences, linear trees, and other objects. For example, Fig. 2 shows the Schmitt/Waterman model, the adapted lattice path model, and the adapted linear tree model for the RNA secondary structure described in Fig. 1.

In Section 2 of this paper, we present a bijection from $SS_{n,k}$ to a certain class of 3-2-1 avoiding permutations. In Section 3, we introduce several combinatorial statistics on $SS_{n,k}$ motivated by statistics on permutations and other objects.

2. A permutation representation

This section considers a mapping of $SS_{n,k}$ to a subset of 3-2-1 avoiding permutations, $\Pi_{n,k}$. This bijection is an adaptation of the Robinson–Schensted–Knuth correspondence as described by Stanley [6].

We will need the following definitions:

Definition 2.1. Let $\pi = \pi_1, \pi_2, \ldots, \pi_n$ be a permutation.

(1) A *substring* in π is any string of numbers $(\pi_{i_1}, \pi_{i_2}, \ldots, \pi_{i_m})$ such that $i_1 < i_2 < \cdots < i_m$.
(2) π is 3-2-1 *avoiding* if it has no decreasing substrings of length 3.
(3) π_k is an *excedance* if $\pi_k > k$.
(4) π_k is a *fixed point* if $\pi_k = k$.
(5) π_k is a *fall* if $\pi_k < k$.
(6) An *inversion* in π is any substring (π_i, π_j) such that $\pi_i > \pi_j$. We use *inv* to denote the number of inversions in π.
(7) k is a *descent* position if $\pi_k > \pi_{k+1}$.
(8) The *major index* of π, *maj*, is the sum of the descent positions in π.

Definition 2.2. Let Π_n be the set of all permutations π such that:

(1) π is 3-2-1 avoiding.
(2) If position i in π contains a fall, position $i + 1$ does not.
(3) If c is a fall in π, $c + 1$ is not.
(4) Each fall c is the second element of at least two inversions.

Definition 2.3. Let $\Pi_{n,k}$ be the set of all permutations in Π_n that have exactly k falls.

Theorem 2.1. *There exists a bijection from* $SS_{n,k}$ *to* $\Pi_{n,k}$.