RNA secondary structures and bicoloured ordered trees

Chunlin Liu
Department of Mathematics, National University of Defense Technology,
Changsha, Hunan 410073 P. R. China
E-mail: liuchunlin77@eyou.com

Abstract. This paper introduces a bijection between RNA secondary structures and bicoloured ordered trees.
MSC: 05C05
Keywords: tree, bijection, RNA

1 Introduction

RNA is described by a linear sequence of bases from the set $\{A$(adenine), C(cytosine), G(guanine), U(uracil)$\}$. An RNA folds back on itself and forms bonds between pairs of bases, creating helical regions, by the well-known Watson-Crick rules: A pairs with U and C with G. This bonded folding of RNA is called secondary structure. RNA secondary structures of a given length with a fixed number of base pairs have been studied ([2], [3]), where the specific identity of the bases is ignored. This paper wants to consider such structures as well, which can be drawn as planar graphs on the upper halfspace in R^2 with vertices on the horizontal axis representing for bases and arcs for pairs. For example, Fig.1 illustrates all RNA secondary structures with 6 bases and 2 pairs. Here another restriction is imposed on the structures that any two consecutive bases can't form a base pair and a base appears in at most one pair.

An ordered tree is a rooted unlabeled tree in which the subtrees of each vertex are linearly ordered, and a bicoloured ordered tree is an ordered tree in which even height vertices are assigned by one colour and odd height ones by the other, where height of a vertex is the distance from it to the root.

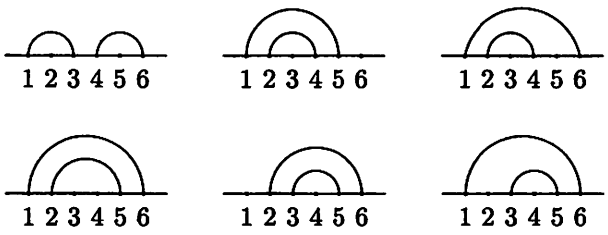

Figure 1

Bijections between RNA secondary structures with $n+k-2$ bases and $k-1$ pairs and ordered trees on n vertices with k internal vertices were presented ([2], [3]), meanwhile the number of such ordered trees equals that of bicoloured ordered trees on n vertices with k even height vertices ([1]), which implies that there could be a connection between RNA secondary structures and bicoloured ordered trees. This paper was motivated by such observations and wants to establish a bijection between them.

2 Bijection

In this paper, we define a secondary structure for a RNA linear sequence of length n as a finite set S (possibly empty) of ordered pairs (i,j) from ordered set $[n] = \{1,2,\ldots,n\}$ where $1 \leq i < j-1 \leq n-1$ satisfying the knot constraint: if $i \leq i^{'} \leq j \leq j^{'}$, then (i,j) and $(i^{'},j^{'})$ can't be two distinct elements of S. Here $1,2,\ldots,n$ represent the linearly ordered bases in RNA. Suppose $1 \leq i \leq n$, we call i a free base if there isn't a pair in S containing i, otherwise a paired one. The definition of secondary structure implies that if i is paired, then there is either a pair (i,j) $(i < j-1)$ or (j,i) $(j < i-1)$ in S, in which cases i will be called resp. as a left base and right base. To a right base i, let $A_1(i) = \{k | k \geq i+1$ and for $i+1 \leq l \leq k$, l is not a right base $\}$ and $A_2(i) = \{m | \exists\ k \in A_1(i)$, s.t. $(k,m) \in S\}$. Obviously the subset of S $\{(k,m) | k \in A_1(i), m \in A_2(i), (k,m) \in S\}$ can constitute also a secondary structure (denoted by $S(i)$) on $A_1(i) \bigcup A_2(i)$, which is called a substructure of S or i's accessible substructure. If $i = n$ or i is followed by a right base, then $A_1(i) = \emptyset$, $A_2(i) = \emptyset$ and i has no accessible substructure. Since the base 1 can't belong to any substructure accessible from a right base, let $A_1(0) = \{k | k \geq 1$ and for $1 \leq l \leq k$, l is not a right base $\}$, $A_2(0) = \{m | \exists\ k \in A_1(0)$, s.t. $(k,m) \in S\}$ and define the leading substructure $S(0)$ of S to be $\{(k,m) | k \in A_1(0), m \in A_2(0), (k,m) \in S\}$ on $A_1(0) \bigcup A_2(0)$. It's easy to see that S decomposes the RNA sequence uniquely into substructures $S(0)$, $S(i_1)$, ..., $S(i_s)$, where $i_1,\ldots,i_s$ are right bases having accessible substructures. For

example, from the RNA in Example 2.2, we get substructures $S(0)$, $S(10)$, $S(17)$, $S(24)$, $S(26)$: $A_1(0) = \{1, 2, \ldots, 9\}$, $A_2(0) = \{10, 15, 16, 17\}$; ...; $A_1(26) = \{27, 28\}$, $A_2(26) = \emptyset$.

To an even height internal vertex v in a bicoloured ordered tree T with root u, define the bicoloured ordered subtree $T(v)$ to be the subgraph of T induced by $\{w | d(u,w) = d(u,v) + d(v,w), d(v,w) \leq 2\}$, where $d(u,v)$ is the distance between u and v.

Let $R_{n,k}$ be the set of RNA secondary structures with n bases and k pairs, and $T_{n,k}$ the set of bicoloured ordered trees with n vertices including k even height ones. The bijection below between $R_{n+k-2,k-1}$ and $T_{n,k}$ is to establish a one-to-one correspondence between substructures of a secondary structure S and subtrees of a bicoloured ordered tree T, such that a free base in S corresponds to an odd height vertex in T and two paired bases (i.e., the base pair) in S to an even height one in T.

Theorem 2.1 *There is a bijection between $R_{n+k-2,k-1}$ and $T_{n,k}$.*

Proof. The procedure to construct a tree T in $T_{n,k}$ from a secondary structure S in $R_{n+k-2,k-1}$ can be described inductively as follows.
(1) Consider the leading substructure $S(0)$ of S. Let the free bases in $S(0)$ be sons of the root in T preserving the same linear order as in $S(0)$.
(2) If there are pairs in $S(0)$, then consider free bases in $S(0)$ which are preceded by left bases. To such a free base i, select out the consecutive left bases preceding i, i.e. $i-l, i-l+1, \ldots, i-1$ if they are left bases while $i-l-1$ is not, and their paired right bases, i.e. $i+i_l, \ldots, i+i_1$ $(i_1 < \cdots < i_l)$. Then let $i-1, \ldots, i-l$ (or $i+i_1, \ldots, i+i_l$) be sons of the vertex i in T ordered from left to right. When all such free bases are considered, we get the subtree of T relative to $S(0)$.
(3) Consider those chosen right bases having accessible substructures. E.g., suppose $i+i_m$ $(1 \leq m \leq l)$ has $S(i+i_m)$, then as in (1) let the free bases in $S(i+i_m)$ be the sons of vertex $i-m$ in T, and as in (2) construct the height 2 vertices in the subtree $T(i-m)$.
(4) Repeat above procedure until all substructures in S are considered.

Since a base is contained in a unique substructure, and from the procedure a free base in a substructure corresponds to an odd height vertex in the subtree and a left base (or the right base) to an even height one, we eventually get a bicoloured ordered tree in $T_{n,k}$. Moreover, free bases preceded by left bases correspond to odd height internal vertices and right bases having accessible substructures to even height internal ones. To the reverse procedure, we have had subtrees of T and construct their corresponding substructures which may be put in the suitable place to get the desired secondary structure. Suppose the root of T is u.
(1) Consider $T(u)$, let the height 1 vertices in $T(u)$ be free bases in $S(0)$.

(2) Consider height 1 internal vertices in $T(u)$, say v. Suppose v has ordered sons $w_1, w_2, \ldots$. Then in $S(0)$ construct left bases, corresponding resp. to $w_1, w_2, \ldots$, preceding consecutively of the free base v. Secondly, from these left bases construct the base pairs in $S(0)$ with their right bases locating after the rightmost free base satisfying: if i_1, i_2 $(i_1 > i_2)$ are left bases and $(i_1, j_1), (i_2, j_2)$ are base pairs, then $j_1 < j_2$. When all height 1 internal vertices in $T(u)$ considered, we obtain the leading substructure $S(0)$ corresponding to $T(u)$.
(3) To height 2 internal vertex w in T, suppose the pair in $S(0)$ relative to w being (i_1, j_1), consider subtree $T(w)$, use (1), (2) and (3) on $T(w)$ to get a substructure, call it $S(j_1)$, and insert it in S by following j_1, which wouldn't violate the knot constraint obviously.
(4) Repeat (3) until all height 2 internal vertices in T are researched, then consider internal vertices of height 4, 6, ... consequently. Since a free base and a pair are added according to an odd height vertex and an even height one resp., when all vertices are considered, we get a secondary structure S in $R_{n+k-2,k-1}$ at last. ∎

Example 2.2 *Following are an RNA secondary structure with 28 bases and 8 pairs and its related bicoloured ordered tree with 21 vertices including 9 even height ones.*

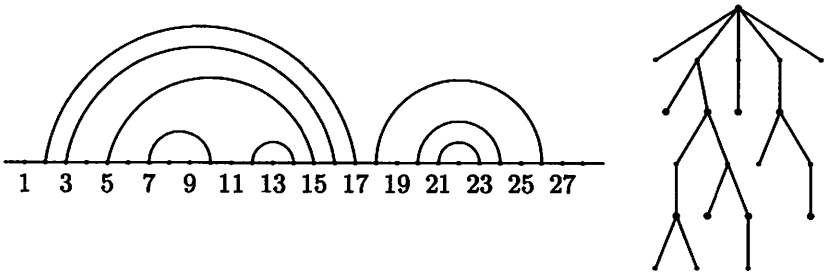

Figure 2

Acknowledgements

I wish to express my sincere thanks to the referees for their very helpful suggestions that have resulted in great improvements that appear here.

References

[1] L. H. Clark, J. E. McCanna and L. A. Székely, A survey of counting bicoloured trees, Bull. Inst. Combin. Appl. 21 (1997), 33-45.

[2] R. C. Penner and M. S. Waterman, Spaces of RNA secondary structures, Advances in Mathematics 101 (1993), 31-49.

[3] W. R. Schmitt and M. S. Waterman, Linear trees and RNA secondary structure, Discrete Appl. Math. 51 (1994), 317-323.