Iranian Journal of Mathematical Chemistry, Vol. 3, No.2, September 2012, pp. 185 – 193 **IJMC**

A zero-one programming model for RNA structures with arc-length ≥ 4

G. H. SHIRDEL AND N. KAHKESHANI

Department of Mathematics, Faculty of Basic Sciences, University of Qom, Qom, Iran

(*Received August 1, 2012*)

ABSTRACT

In this paper, we consider RNA structures with arc-length ≥ 4. First, we represent these structures as matrix models and zero-one linearprogramming problems. Then, we obtain an optimal solution for this problemusing an implicit enumeration method. The optimal solution corresponds toan RNA structure with the maximum number of hydrogen bonds.

Keywords: RNA structure,zero-one linear programming problem, additive algorithm.

1. INTRODUCTION

A problem in mathematical biology is enumeration of RNA structures. The RNAhas an important role within cells and also, its functions depend on the structure of theRNA molecules. Hence, understanding of its helical configuration is important. TheRNA molecule is a sequence of four nucleotides A, C, G and U which plays an importantrole in Biological reactions. These nucleotides are connected to each other via hydrogenbonds. The formation of these bonds stabilizes the molecule by lowering its free energy[2]. The RNA structures can be displayed in various ways such as tree, linear encodingof tree, coarse grained representation, homeomorphically irreducible tree and diagram[3, 4]. In this paper, we represent another two models for RNA structures. Accordingto [4], the diagram representation is defined as following:

Let $G_n = (V_{G_n}, E_{G_n})$ be a directed graph such that

$$V_{G_n} = [n] = \{1, \cdots, n\} \quad \text{and} \quad E_{G_n} \subset \{(i, j) \mid 1 \leq i < j \leq n\}.$$

V_{G_n} and E_{G_n} are called the sets of vertices and arcs, respectively. Each directed graphcan be displayed as a diagram in which the vertices $\{1, \ldots, n\}$ are placed on a horizontalline and the arcs (i, j), where $i < j$, can be displayed above the line. Because of linearordering of the vertices, the direction of the arcs is omitted. The vertices andarcs show the nucleotides and

hydrogen bonds, respectively. We attribute two parameters to the diagrams: the minimum arc-length, λ, the minimum stack-length, σ. In diagram representation, the length of an arc (i, j) is $j-i$ and a stack of length σ is a sequence of the parallel arcs like $((i, j),(i+1, j-1),\ldots,(i+(\sigma-1), j-(\sigma-1)))$, see Figure 1. We denote the number of RNA structures with $\lambda = 4$ and $\sigma = 1$ over $[n]$ by $S_{4,1}(n)$.

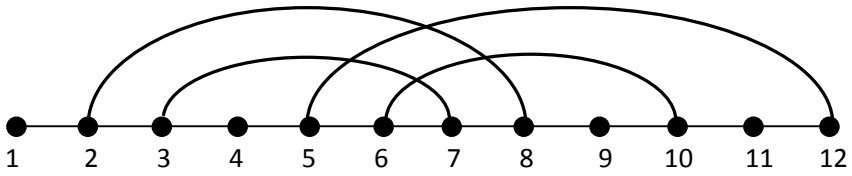

Figure1. RNA Structure with $\lambda = 4, \sigma = 2$.

The remainder of this paper is organized as follows. In sections 2 and 3, we represent RNA structures with arc-length ≥ 4 and stack-length ≥ 1 as matrix models and zero-one linear programming problems, respectively. Also, we express the results forenumeration of RNA structures. In section 4, we use the additive algorithm for solving linear programming problems with binary variables only [1]. The general notion of the additivealgorithm is based on testing a few number of possible solutions, 2^m (in which m is thenumber of variables), of a problem instead of all solutions. In other words, through thismethod, some of the possible solutions of the problem are left unexamined. Also,the zero-one linear programming problems can be solved using each of the generalinteger programming techniques. Finally, conclusions and future works are discussed insection 5.

2. MATRIX MODELS

Each RNA structure over $[n]$, G_n, corresponds to a $n \times n$ matrix. We display this matrix by $M(G_n) = [m_{ij}]$ such that

$$m_{ij} = \begin{cases} 1 & if\ (i, j) \in E_{G_n} \\ 0 & if\ (i, j) \notin E_{G_n} \end{cases}.$$

Theorem 1. Suppose G_n is an RNA structure with arc-length ≥ 4 over $[n]$. Then we have

$$M(G_n) = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \tag{1}$$

where $A_{11} = 0_{(n-4)\times 4}$, $A_{21} = 0_{4\times 4}$, $A_{22} = 0_{4\times(n-4)}$ and A_{12} is the $(n-4)\times(n-4)$ upper triangular matrix.

Proof. For all (i, j), where $i \geq j, (i, j) \notin E_{G_n}$. Therefore, $M(G_n)$ is an upper triangular matrix. Since G_n is a directed graph with arc length ≥ 4, we have $(i, j) \notin E_{G_n}$ and $m_{ij} = 0$ for all (i, j), where $j - i \leq 3$. Therefore, three diagonals above theprincipal diagonal are zero. Then, the matrix $M(G_n)$ is of the form (1). □

We now write the upper triangular matrix A_{12} as follows:

$$\begin{pmatrix} m_{15} & m_{16} & \cdots & m_{1n} \\ 0 & m_{26} & \cdots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_{(n-4)n} \end{pmatrix}$$

Theorem 2. $S_{4,1}(n)$ is equal to the number of situations in which the entries aboveand on the principal diagonal in A_{12} can be equal to 1 such that for each $1 \leq i \leq n-4$ and $5 \leq j \leq n$, at most one of the entries

$$\{m_{i(i+4)}, \ldots, m_{in}, m_{j(j+4)}, \ldots, m_{jn}, m_{1i}, \ldots, m_{(n-4)i}, m_{1j}, \ldots, m_{(n-4)j}\}$$

be 1.

Proof. Let G_n is an RNA structure with arc-length ≥ 4 over $[n]$. The degree of each vertex G_n is at most 1. Therefore, each row and column A_{12} is0 or e_k, where $1 \leq k \leq n-4$. Suppose that there exist $1 \leq i \leq n-4$ and $5 \leq j \leq n$ so that $m_{ij} = 1$. Then $(i, j) \in E_{G_n}$. Since the degree of each vertex G_n is at most 1,we have $(k,i),(i,h) \notin E_{G_n}$ for each $k \leq i$ and $h > i$, where $h \neq j$. Similarly, we have $(j,h),(k, j) \notin E_{G_n}$ for each $h \geq j$ and $k < j$, where $k \neq i$. This completes our argument. □

3. ZERO-ONE LINEAR PROGRAMMING PROBLEMS

Here, we present a zero-one linear programming problem for displaying of RNA structures with arc-length ≥ 4 and stack-length ≥ 1.Since each row of the matrix A_{12} is 0 or e_k, where $1 \leq k \leq n-4$, we have the following constraints:

$$(a)\begin{cases} x_{15}+x_{16}+x_{17}+\cdots+x_{1n}\le 1\\ x_{26}+x_{27}+\cdots+x_{2n}\le 1\\ x_{37}+\cdots+x_{3n}\le 1\\ x_{48}+\cdots+x_{4n}\le 1 \end{cases}$$

$$(b)\begin{cases} x_{59}+\cdots+x_{5n}\le 1\\ \vdots\\ x_{(n-7)(n-3)}+\cdots+x_{(n-7)n}\le 1\\ x_{(n-6)(n-2)}+x_{(n-6)(n-1)}+x_{(n-6)n}\le 1\\ x_{(n-5)(n-1)}+x_{(n-5)n}\le 1\\ x_{(n-4)n}\le 1 \end{cases}$$

Similarly, since each column of matrix A_{12} is 0 or e_k, where $1\le k\le n-4$, we have the following constraints:

$$(c)\begin{cases} x_{15}\le 1\\ x_{16}+x_{26}\le 1\\ x_{17}+x_{27}+x_{37}\le 1\\ x_{18}+x_{28}+x_{38}+x_{48}\le 1\\ x_{19}+x_{29}+x_{39}+x_{49}+x_{59}\le 1\\ \vdots\\ x_{1(n-4)}+x_{2(n-4)}+\cdots+x_{(n-8)(n-4)}\le 1 \end{cases}$$

$$(d)\begin{cases} x_{1(n-3)}+x_{2(n-3)}+\cdots+x_{(n-7)(n-3)}\le 1\\ x_{1(n-2)}+x_{2(n-2)}+\cdots+x_{(n-6)(n-2)}\le 1\\ x_{1(n-1)}+x_{2(n-1)}+\cdots+x_{(n-5)(n-1)}\le 1\\ x_{1n}+x_{2n}+\cdots+x_{(n-4)n}\le 1 \end{cases}$$

We also know that the degree of each vertex is at most 1. Therefore, we have thefollowing constraints:

$$(e)\begin{cases} x_{15}+x_{59}+x_{5(10)}+\cdots+x_{5n}\le 1 \\ x_{16}+x_{26}+x_{6(10)}+x_{6(11)}+\cdots+x_{6n}\le 1 \\ x_{17}+x_{27}+x_{37}+x_{7(11)}+x_{7(12)}+\cdots+x_{7n}\le 1 \\ x_{18}+x_{28}+x_{38}+x_{48}+x_{8(12)}+x_{8(13)}+\cdots+x_{8n}\le 1 \\ \vdots \\ x_{1(n-6)}+x_{2(n-6)}+\cdots+x_{(n-10)(n-6)}+x_{(n-6)(n-2)}+x_{(n-6)(n-1)}+x_{(n-6)n}\le 1 \\ x_{1(n-5)}+x_{2(n-5)}+\cdots+x_{(n-9)(n-5)}+x_{(n-5)(n-1)}+x_{(n-5)n}\le 1 \\ x_{1(n-4)}+x_{2(n-4)}+\cdots+x_{(n-8)(n-4)}+x_{(n-4)n}\le 1 \end{cases}$$

Based on the following lemma, some of the constraints are extra and they can beomitted.

Lemma 1. Let $S=\{x \mid Ax\le b, x\ge 0\}$, where A is a $m\times n$ matrix with rank m and $b\in \mathbf{R}^m$. Let two constraints

$$x_1+\cdots+x_k\le b_i \quad \text{and} \quad x_1+\cdots+x_k+x_{k+1}+\cdots+x_{k+j}\le b_i$$

belong to the set of constraints $Ax\le b$. Then the constraint $x_1+\cdots+x_k\le b_i$ is extra.

Proof. Suppose that S' be the feasible region after deleting the constraint $x_1+\cdots+x_k\le b_i$. Let the constraint $x_1+\cdots+x_k\le b_i$ isn't extra. Then $S\subset S'$. Let $(x_1',\ldots,x_n')\in S'$ and $(x_1',\ldots,x_n')\notin S$.
Therefore,

$$x_1'+\cdots+x_k'>b_i \quad \text{and} \quad x_1'+\cdots+x_k'+x_{k+1}'+\cdots+x_{k+j}'\le b_i.$$

Introducing the slack variables $y\ge 0$ and $z>0$, we have the following constraints in standard form:

$$x_1'+\cdots+x_k'-z=b_i \quad \text{and} \quad x_1'+\cdots+x_k'+x_{k+1}'+\cdots+x_{k+j}'+y=b_i.$$

Therefore,

$$x_{k+1}'+\cdots+x_{k+j}'+y+z=0. \tag{3}$$

On the other hand,

$$x_{k+1}'+\cdots+x_{k+j}'+y+z>0, \tag{4}$$

Since $x'_{k+1},\ldots,x'_{k+j}, y \geq 0$ and $z > 0$. But, this is a contradiction. Then $S = S'$. □

According to Lemma 1, the constraints (b) and (c) are extra. Therefore, we can delete them. Now, we define the Problem A as follows:

Problem A:

$$
\begin{aligned}
&Max \quad \sum_{(i,j)\in E} x_{ij} \\
&s.t. \quad \sum_{j=i+4}^{n} x_{ij} \leq 1 && i \in \{1,2,3,4\} \\
&\quad \sum_{i=1}^{j-4} x_{ij} \leq 1 && j \in \{n-3, n-2, n-1, n\} \\
&\quad \sum_{i=1}^{j-4} x_{ij} + \sum_{i=j+4}^{n} x_{ji} \leq 1 && j \in \{5,6,\ldots,n-4\} \\
&\quad x_{ij} = 0,1 && \forall (i,j) \in E
\end{aligned}
$$

where $E = \{(i,j) \mid j - i \geq 4\}$.

The number of variables and constraints of the Problem A has been presented inTable 1.

Theorem 3. $S_{4,1}(n)$ is equal to the number of feasible solutions of the Problem A. Also, thenumber of optimal solutions of the Problem A is equal to the number of RNA structures with arc length ≥ 4 and maximum number of arcs over $[n]$.

Proof. The Problem A is written on the basis of the matrix model (1). Therefore, Theorem 2 guarantees that $S_{4,1}(n)$ is equal to the number of feasible solutions. Sincethe objective function is equal to the sum of the variables and the Problem A is themaximization problem, then among the feasible solutions, the optimal solution belongs to the one in which the maximum number of x_{ij} variables would be equal to 1. So, the optimal solution has maximum number of arcs over $[n]$. □

4. USING ADDITIVE ALGORITHM FOR SOLVING THE PROBLEM A

There are different methods for solving a zero-one linear programming problem such as the additive algorithm. For using additive algorithm, a problem must possess three following conditions:

1. Its objective function should be in the form of minimization.
2. The coefficients of the objective function should be nonnegative.
3. All the constraints must be of the $\leq$ type.

Table 1.The Number of Variables and Constraints of the Problem A.

n	6	7	8	9	10	11	$n \geq 12$
The number of variables	3	6	10	15	21	28	$8n-60$
The number of constraints	2	4	6	9	10	11	n

Bysetting $x'_{ij} = 1 - x_{ij}$, the Problem A is converted into the following problem:

Problem B:

$$Min \quad \sum_{(i,j)\in E} x'_{ij}$$

$$s.t. \quad -\sum_{j=i+4}^{n} x'_{ij} \leq i+4-n \qquad i \in \{1,2,3,4\}$$

$$-\sum_{i=1}^{j-4} x'_{ij} \leq 5-j \qquad j \in \{n-3, n-2, n-1, n\}$$

$$-\sum_{i=1}^{j-4} x'_{ij} - \sum_{i=j+4}^{n} x'_{ji} \leq 8-n \qquad j \in \{5,6,\ldots,n-4\}$$

$$x'_{ij} = 0,1 \qquad \forall (i,j) \in E$$

Now, we can apply the additive algorithm for solving the Problem B. In Table 2, we list the optimal solutions for $n = 6,\ldots,10$.

Example 1. For *n* = 7, the problems A is as follows:

$$Max \quad z = x_{15} + x_{16} + x_{17} + x_{26} + x_{27} + x_{37}$$

$$s.t. \quad x_{15} + x_{16} + x_{17} \leq 1$$

$$x_{26} + x_{27} \leq 1$$

$$x_{16} + x_{26} \leq 1$$

$$x_{17} + x_{27} + x_{37} \leq 1$$

$$x_{15}, x_{16}, x_{17}, x_{26}, x_{27}, x_{37} \geq 0$$

Using of additive algorithm, the optimal solution of the Problem A is equal to

$$x_{16} = x_{17} = x_{27} = 0,\ x_{15} = x_{26} = x_{37} = 1,\ z^* = 3.$$

This solution is corresponding to the RNA structure which is shown in Figure 2.

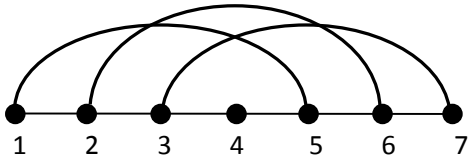

Figure 2. Optimal Structure for $n = 7$.

The matrix model of this optimal structure is as follows:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Table2. The Optimal Solutions of the Problem A, for $n = 6,\ldots,10$.

n	Optimal solution	Optimal solution value
6	$x_{15} = x_{26} = 1,\ x_{16} = 0$	2
7	$x_{15} = x_{26} = x_{37} = 1,\ x_{16} = x_{17} = x_{27} = 0$	3
8	$x_{15} = x_{26} = x_{37} = x_{48} = 1,$ $x_{16} = x_{17} = x_{18} = x_{27} = x_{28} = x_{38} = 0$	4
9	$x_{16} = x_{37} = x_{48} = x_{59} = 1,$ $x_{15} = x_{17} = x_{18} = x_{19} = x_{26} = x_{27} = 0,$ $x_{28} = x_{29} = x_{38} = x_{39} = x_{49} = 0$	4
10	$x_{15} = x_{27} = x_{38} = x_{49} = x_{6(10)} = 1,$ $x_{16} = x_{17} = x_{18} = x_{19} = x_{26} = x_{28} = 0,$ $x_{29} = x_{37} = x_{39} = x_{48} = x_{59} = x_{1(10)} = 0,$ $x_{2(10)} = x_{3(10)} = x_{4(10)} = x_{5(10)} = 0$	5

5. CONCLUSION

Poolsap et al. in [5] represented an integer programming problem for the RNA structures. Their model is complicated with many variables. But, in here, we represented another programming problem for the RNAstructures such that the number of variables is lessthan the number of variables in [5]. The optimal RNA structure obtained by solving this problem is of arc-length ≥ 4 and has the maximum number of hydrogen bonds. Inother words, formation of these bonds stabilizes the structure by lowering its free energy over $[n]$. If the enumeration of the feasible solutions of the zero-one linear programming problem is possible, then we are able to enumerate the RNA structures with arc-length ≥ 4 over $[n]$.also, in this case, the number of optimal solutions will be equal to the number of optimal RNA structures. Therefore, a future work can be the enumeration of the RNA structures using matrix and linear programming models.

REFERENCES

1. E. Balas, *An additive algorithm for solving linear programs with zero-one variables*, Operations Research, Vol. **13**, No. 4 (1965) 517–546.
2. R. T. Batey, R. P. Rambo, J. A. Doudna, *Tertiary motifs in RNA structure and folding*, Angew. Chem. Int. Ed. **38** (1999) 2326–2343.
3. I. L. Hofacker, P. schuster, P. F. Stadler, *Combinatorics of RNA secondary structures*, Discrete Appl. Math. **88** (1998) 207–237.
4. E. Y. Jin, C. M. Reidys, *Combinatorial design of pseudo knot RNA*, Adv. Appl. Math. **42** (2009) 135–151.
5. U. Poolsap, Y. Kato, T. Akutsu, *Prediction of RNA secondary structure with pseudo knots using integer programming*, MBC Bioinformatics. **10** (Suppl. I):S38 (2009).