第1卷2期 生物化学杂志 Vol.1, No.2
1985年4月 Chinese Biochemical Journal Apr., 1985

# 一种预测单链RNA分子二级结构的计算机算法

何东明 陈农安

（中国科学院上海生物化学研究所，上海）

本文给出了一个利用已知能量数据构成具有最小自由能的单链RNA分子二级结构的计算机算法，并给出了此算法的可行性证明和应用实例。

由于RNA分子的高级结构对其生物功能的行使起着极其重要的作用，所以近几年来，人们对它越来越感兴趣。

目前，RNA还不能象很多蛋白质那样采用X光衍射技术来研究它的高级结构，这主要是由于它的晶体难以得到。很多专家分别采用了化学和酶学等各种实验方法来研究RNA的二级结构，取得了一些可喜的结果。但是，实验的方法有它的局限性和人为因素的影响；譬如：有人采用酶学方法把大分子分割成若干个较小部分来研究，据此推测出的三级结构很难说是正确的，因为不能排除分割后分子的某些部分的二级结构发生了改变这种可能性。

借助电子计算机来预测RNA的最稳态二级结构的方法是由Pipas和McMahon首先提出来的[1]。随后，Nussinov等[2]和Studnicka[3]等分别发表了他们各自建立的可用来推测RNA的二级结构的算法。Nussinov的算法对于求得具有最大配对数目的二级结构是很有效的，但是不便于把能量计算结合进去，尤其是对于碱基对之间的堆积能更是无能为力。尽管Studnicka的算法能求出具有最小自由能的二级结构，但方法烦琐、费时。该算法是在列出所有可能的二级结构后，逐个地计算出它们的自由能，然后进行比较，从中选出具有最小能量者。运行此算法所需的时间将是$N^5$数量级的（假设$N$为RNA分子的链长）。

本文介绍的算法是我们在Nussinov算法和Studnicka算法的基础上扬长弃短而成的，既能算出具有最小自由能的结构，又有方法简单实用之优点，运行时间仅是$N^3$数量级。

## 算　法

如Fig.1.所示，我们把具有$N$个碱基的RNA分子看成是平面上的一个圆，每个碱基用圆周上的点来表示，用连接圆周上两点的圆内连线来表示两个碱基之间的配对。图1.中的$i$、

$i$ 分别是分子上某一片段的起点和尾点，$p$ 等于片段链长减 1，$k$ 为片段内某一点，它们满足下列式子：

$$\begin{cases} 8 \leqslant p \leqslant N-1 \\ 1 \leqslant i \leqslant N-p \\ j = i+p \leqslant N \\ i+6 \leqslant k \leqslant j-2. \end{cases}$$

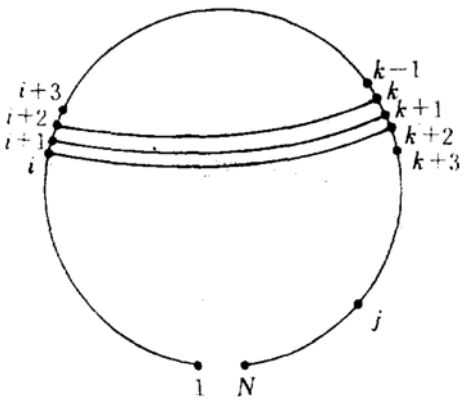


Fig.1　A RNA molecule which has N nucleotides

根据生化知识：

1．碱基之间按A-U、C-G和G·U规则配对。

2．每一碱基至多能同一个碱基配对。

3．不考虑核酸链出现真结（ture knot）和假结（pseudoknot）的情况。

4．一个发夹环（hairpin loop）至少含有三个非配对碱基。

5．少于三对连续的碱基配对区是不稳定的。

按照以上几点，我们可以得到相应的图论规则：

1．每一圆内连线的两端点必须是(A，U)、(C，G)和(G，U)。

2．圆周上每一点至多由一条连线连接。

3．连线不得两两相交。

4．连线两端点在圆周上至少相隔三个点。

5．平行的相邻连线至少要有三条。

根据以上五条规则，并为了便于相邻碱基对之间的堆积能量的计算，我们以连续的三点$\langle i, i+1, i+2\rangle$与三点$\langle k+2, k+1, k\rangle$相对应，看是否都能配对，如能，则计算出这三对碱基之间的自由能$E_{ik}$；所用能量参数是由Salser于1977年提出的[4]。

我们采用递归算法来计算 RNA 分子的最小自由能 $E_{min}$。从5′端数起第 $i$ 个核苷酸到第 $j$ 个核苷酸这一片段上的最小自由能可简单地由公式（1）来表示（请参考Fig.1.）：

公式（1）：
$$E(i,j) = min \begin{cases} E(i+3, k-1) + E(k+3, j) + E_{ik} + E_{des} \\ E_n \quad ; \quad i+6 \leqslant k \leqslant j-2. \end{cases}$$

其中：
$$E_n = \begin{cases} E(i+1, j); & \text{当} E(i+1, j) \neq 0. \\ +\infty \quad ; & \text{当} E(i+1, j) = 0. \end{cases}$$

当 $j-i < 8$ 时，$E(i,j) = 0$。

显然，当 $i=1$，$j=N$ 时，$E(i,j) = E(1,N) = E_{min}$。

公式（1）中的 $E_{des}$ 是片段$(i, j)$中的在二级结构中起不稳定作用的正值的自由能，它的计算方法将在下一节中谈到。下面用第二数学归纳法来证明公式（1）的正确性。

证明：

1，当 $j-i=8$ 时，公式（1）显然成立。

2. 设：当$8<j-i\leqslant l$时，公式（1）成立（$l$：为满足$8<l<N-1$的某一整数）。

3. 现要证明当$j-i=l+1$时公式（1）也成立：

设 $j-i=l+1$，

那么一定有：$(k-1)-(i+3)\leqslant l$，

$$j-(k+3)\leqslant l,$$

$$j-(i+1)\leqslant l;$$

所以，$E(i+3,k-1)$、$E(k+3,j)$和 $E(i+1,j)$分别是相应片段上的最小自由能。

现设，三点$<i,i+1,i+2>$在片段$(i,j)$中存在配对的三点$<k+2,k+1,k>$，由于这时不允许片段$(i+3,k-1)$中的点与片段$(k+3,j)$中的点配对，所以片段$(i,j)$上的能量为 $E_e=E(i+3,k-1)+E(k+3,j)+E_{ik}+E_{des}$。把所有$k$下的$E_e$值都求出，并两两比较，取出最小者，命名取得最小$E_e$值时的$k$为$k_0$。显然在片段$(i,j)$上的任何不含有连接点$i$的连线的结构的能量都大于$E(i+1,j)$，所以$k_0$值下的$E_e$与$E(i+1,j)$两者中的较小者就是片段$(i,j)$上的最小自由能$E(i,j)$。

当三点$<i,i+1,i+2>$在片段$(i,j)$中无相应的配对点时，则有 $E(i,j)=E(i+1,j)$，这时令 $k_0=0$，证毕。

我们在算法中建立了一个矩阵$E$，$E$阵元素的内容是这样确定的：设 $i<j=i+p$，元素$E(i,j)$是与公式（1）中的$E(i,j)$一致的，$E(j,i)$是达到 $E(i,j)$时的$k_0$值。

我们采用一个追溯算法，从$E$阵中找出达到$E_{min}$的结构的所有$k_0$值，这样就等于把具有最小自由能的RNA分子的二级结构找到了。

## 算法的计算机实现

$E_{des}$是指发夹环(hairpin loop)、内环(interior loop)和膨胀环(bulge)这一类结构(Fig. 2.)起不稳定作用的自由能，其值大小与环的大小及封闭环的那对碱基的类型有关。

$E_{des}$的计算过程请见程序的粗框图Fig.3。

追溯算法在数据结构上是由一个先进后出的栈实现的。

程序是用FORTRAN语言编写的，移植性强，已分别在IBM 4341机、宝来B1955机和TRS-80机上成功地运行。


Fig.2 a. Hairpin loop b. Interior loop c. Bulge

## 应 用 实 例

为了验证算法和程序的有效性，我们计算了一些 RNA 分子，得出了它们的最稳态二级结构。由于篇幅所限，在此仅列举其中的几个，以供参考。

Table 1

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| pair | 1→10 | 11→14 | 16→19 | 20→23 | 28→30 | 31→34 | 79→86 |
| | 119←110 | 76←73 | 70←67 | 63←60 | 56←54 | 51←48 | 97←90 |

Table 2

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| pair | 1→7 | 8→10 | 16→21 | 24→26 | 29→32 | 67→72 | 73→75 | 80→82 |
| | 118←112 | 66←64 | 62←57 | 53←51 | 48←45 | 108←103 | 97←95 | 93←91 |

Tab.1和Tab.2分别是大肠杆菌 5Sr RNA分子和蓖麻蚕 (*Philosamia cynthia ricini*) 5S r RNA 分子的自由能最小的二级结构中出现的碱基配对区。这些从理论上推测出的结构与从实验中推测出的结构[5,6] 基本上是一致的。值得一提的是：从实验中推测出的蓖麻蚕 5S rRNA的二级结构模型[6]中存在着U-C 和 U-U 这种非规则配对，在我们的算法中没有考虑这种极少见的情况，所以，我们推出的结构与参考文献[6]中的结构稍有差别。我们推测出的大肠杆菌16S rRNA分子的中间片段（560～912）的自由能最小的二级结构[7]与实验结果[8]基本一致。大肠杆菌 5S rRNA、蓖麻蚕 5S rRNA 和大肠杆菌 16S rRNA 的最小自由能分别为 −60.95kcal/mole、−49.97kcal/mole和−144.36kcal/mole。

## 结　　语

虽然 RNA分子的核苷酸排列顺序是决定其二级结构的最主要的因素[2]，但是除此之外影响RNA二级结构的其它因素还是很多的，例如RNA的环境对它的结构也起着一定的作用，所以为了推出一个更符合实际的 RNA 的 二级结构，我们应该在算法中尽可能多地考虑各种影响 RNA结构的因素。

### 参 考 文 献

[1] Pipas, J.M.and McMahon, J.E.(1975), *Proc.Nat.Acad.Sci.USA*, 72, 2017-2021.

[2] Nussinov, R., Pieczenik, G., Griggs, J.R.and Kleitman, D.J., (1978), *SIAM J.Appl.Math.*, 35, 68-82.

[3] Studnicka, G.M., Rahn, G.M., Cummings, I.W.and Salser, W. A. (1978), *Nucleic Acids Res.*, 5, 3365-3387.

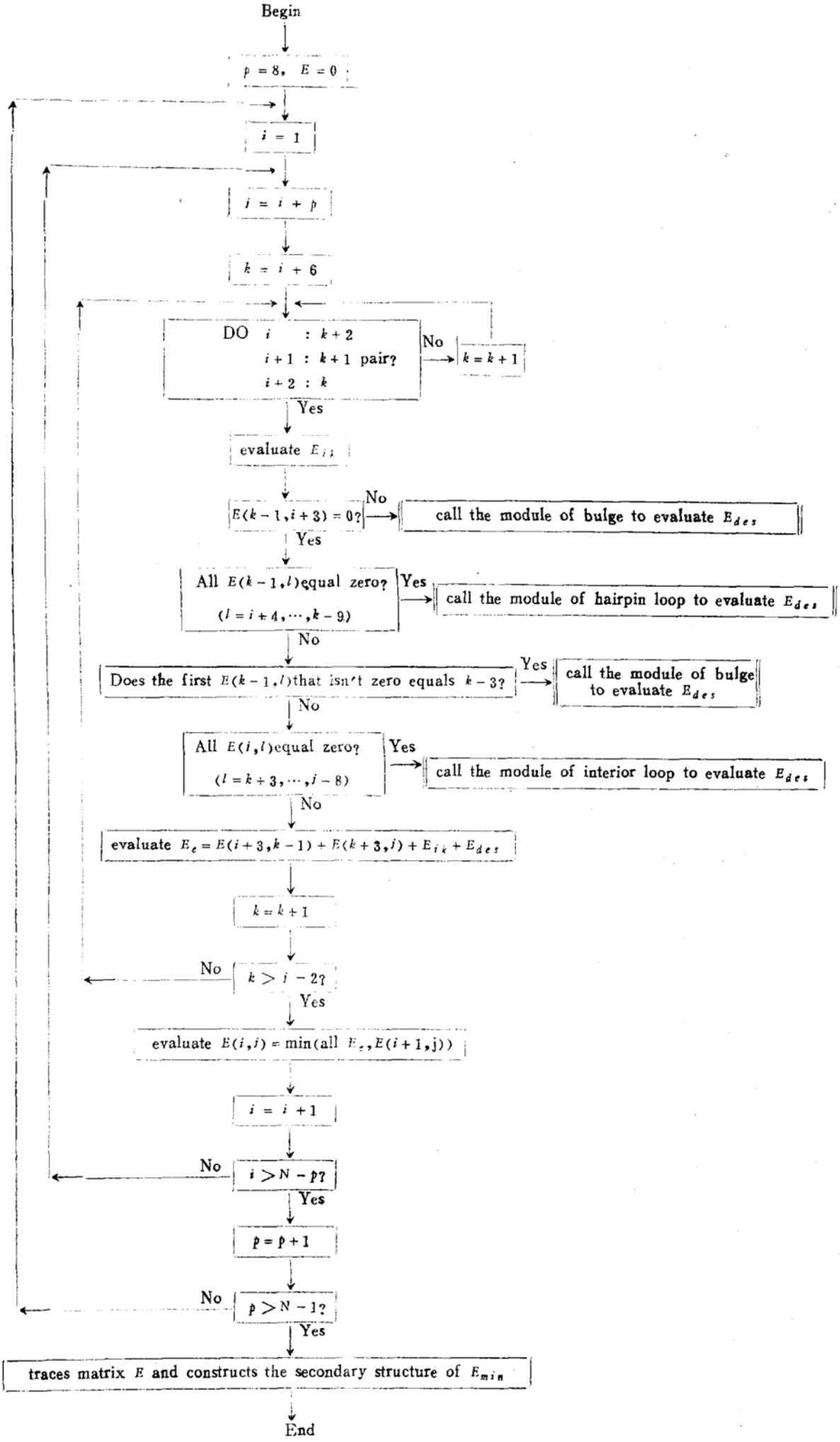


Fig. 3. Program flowchart

[4] Salser, W.(1977), *Cold Spring Harbor Symp.Quant.Biol.*, 42, 985-1002.
[5] Garrett, R.A., Douthwaite, S.and Noller, H.F.(1981), *Trends in Biochemical Sciences*, 6, No.5:137.
[6] Gu, X.R., Nicoghosian, K.and Cedergren, R.J.(1982), *Nucleic Acids Res.*, 10, 5711.
[7] 何东明，陈农安(1985)，生物化学与生物物理学报，vol.17, No.2, (待发表)，
[8] Noller, H.F.and Woese, C.R.(1981), *Science*, 212, 403.

# A COMPUTER ALGORITHM FOR PREDICTING THE SECONDARY STRUCTURE OF RNA

He Dongming Chen Nongan

(Shanghai Institute of Biochemistry, Academia Sinica)

We present a computer method for folding an RNA molecule that finds a conformation of minimum free energy using published values of base pairing energies and destabilizing energies. It is based on a dynamic programming algorithm from applied mathematics.

We have shown that how the algorithm works, and sketched a proof of the validity of the algorithm.

Two simple half matrices are constructed and the best secondary structure can be chosen directly from the second matrix by a back-tracking procedure.

We have included the results from our solutions of secondary structures for the *E.coli* 5S rRNA, the *Philosamia cynthia ricini* 5S rRNA and the central domain of *E.coli* 16S rRNA as an examples of our program in application.

**Keywords:** RNA Structure, Computer Algorithm