单链 RNA (DNA) 分子的最大碱基配对折叠结构的预测

乐树云　　江寿平

(中国科学院上海生物化学研究所)

天然 RNA 为单链分子,现有的X-光衍射分析证明: 单链 RNA 分子能通过自身的回折使可以彼此配对的碱基(主要是碱基A和 U, G 和C的配对)以氢键相连而折叠成部分双螺旋结构,而彼此不能配对的碱基则形成部分突环区. 由此单链 RNA 分子可以形成部分螺旋和部分突环相间排列的高级结构. 本文报道了借助于 TRS-80 微型机和拓扑平面图的最大匹配处理方法而预测这些单链 RNA 或 DNA 分子所具有的最大碱基配对数构象最二级结构.

一、方　法

1. 拓扑平面图的数学处理　对于长度为N个核苷酸残基的一已知单链 RNA (或 DNA) 序列, 若按照碱基配对原则即碱基A和U(A 和 T), G 和C的配对, 该核酸分子可以折叠成形形色色部分螺旋,部分突环区排列的二级结构. 但是从生物意义角度来看,我们感兴趣的仅仅是该单链分子自身回折的最可能的折叠构象. 这显然取决于该折叠构象的稳定性, 即取决于形成该稳定结构的自由能 G[1-2]. 为了便于问题的简化, 我们假定在 RNA (或 DNA) 单链分子的自身回折过程中形成的 G—**C** 和 A—**U** (A—T) 碱基配对的稳定性被视为是等同的. 这样,当忽略碱基堆积和突环区形成对结构稳定性的贡献,则寻求单链 RNA (或 DNA)分子自身回折成具有最小自由能结构的问题就可方便地转化成寻求该单链核酸分子自身回折成的具有最大碱基配对数的二级结构的问题.

这里我们引入数学的图论处理方法,按照图论的处理,上述寻求核酸分子的二级结构问题可归结为一拓扑平面图 $G(X, E)$ 的最大匹配问题. 在处理中, 可以视该预测的单链核酸分子的N个核苷酸残基对应于该拓扑平面图 $G(\boldsymbol{X}, E)$ 中的各顶点 $x_1, x_2, \cdots, x_N$; $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$. 而能够按照 Watson-Crick 配对原则彼此相互配对的碱基间所能形成的氢键, 对应于该拓扑平面图中的弧. 我们以 E 表示该拓扑平面图中弧的集合. 因此当该拓扑平面图 $G(X, E)$ 具有最大匹配时,即为我们所求的解.

按照图的匹配理论,可知该拓扑平面图中的一个最大的匹配,可以从该图中某一匹配逐步地沿无公共顶点的交错链的一系列转换而得到. 而匹配问题可归结为寻找图中连一未饱和顶点到另一未饱和顶点的正当路问题. 基于上述思想,我们可以建立一种递归的, 动态的, 快速预测方法. 从而借助于电子计算机对单链核酸分子序列的处理而预测该分子所具有的最大碱基配对数构象的二级结构.

2. 算法的原理　首先应指出的是我们定义的单链核酸分子的结构与其对应的拓扑平面

本文 1983 年 1 月17日收到.

图具有如下的对应关系和性质.

(1) 拓扑平面图中的顶点依次对应着单链核酸分子中相应的核苷酸残基，每个顶点至多有一条弧与其相匹配的顶点相连,顶点的足标对应着该核苷酸残基在核苷酸序列中序号.

(2) 拓扑平面图中的弧取决于 Watson-Crick 碱基配对原则(对于 RNA 且规定碱基G与U也能配对),当顶点 x_i 与 x_j 相匹配时,记 $E_{i-j}=1$，其中足标 i,j 必须满足条件 $j>i+3(i>0)$. 反之,当 $j<=i+3(j>i)$ 或 x_i 与 x_j 不相匹配时,则记 $E_{i-j}=0$.

(3) 图 $G(X,E)$ 中的弧是两两互不相交：应该说明的是性质(2)中,我们引入了当顶点 x_i 与 x_j 相匹配时必须满足 $j>i+3$ 的条件. 这是为了确保核酸序列当自身回折形成的发夹环中至少有三个核苷酸残基，以保证折叠成的发夹环具有生物意义.

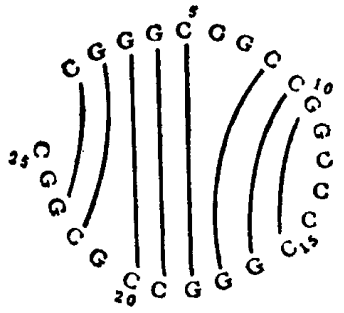

图1 25 核苷酸残基长的核酸序列具有最大碱基配对数二级结构

在图1中表示了与一假想的长为25个核苷酸残基的核酸序列 $^{5'}C-G-G-G-C-G-G-C-C-G-G-C-C-C-C-G-G-G-C-C-G-C-G-G-C_{-OH}$,相对应的满足上述性质的具有最大匹配数的拓扑平面图. 根据上述定义的性质,对于一长度为N个核苷酸残基的单链核酸分子,它所具有的最大碱基配对数的二级结构可由如下算法公式决定. 首先我们以记号 M. M. S. (i,j) 表示子图 $G_{ij}(X_{ij},E_{ij})$ 的最大匹配数,其中 $X_{ij}=\{x_i,x_{i+1},\cdots,x_j\}$. 则依照最大匹配原理和 Nussinov 等人的处理[3] 我们有算法公式

$$\text{M.M. S. }(1,N)=\text{MAX}\begin{cases}\text{M. M. S. }(2,N),\\ \text{M. M. S. }(1,N-1),\\ \text{M. M. S. }(1,k-1)+\text{M. M. S. }(k+1,N-1)+1,\\ \qquad\text{(当顶点 } x_k \text{ 与顶点 } x_N \text{ 相匹配)}\\ k:\ 1<k<N-3.\end{cases}\tag{I}$$

公式I中各子图的最大匹配数有公式

$$\text{M. M. S. }(i,j)=\text{MAX}\begin{cases}\text{M. M. S. }(i+1;j),\\ \text{M. M. S. }(i,j-1),\\ \text{M. M. S. }(i,k-1)+\text{M. M. S. }(k+1,j-1)+1,\\ \qquad\text{(当顶点 } x_k \text{ 与顶点 } x_j \text{ 相匹配时)}\\ i:\ 1\leqslant i\leqslant N-1;\quad j:\ 5\leqslant j\leqslant N,\\ k:\ i<k<j-3.\end{cases}$$

和公式 M. M. S. $(i,j)=0$，当 $i\geqslant j-3$ 且 $i<j$ 时.

这里我们已经考虑了更一般的情况，而对 Nussinov 等人提出的算法公式加以一定的修正. 例如，对于求解子图 $G_{ij}(X_{ij},E_{ij})$ 的最大匹配数 M. M. S. (i,j) 时，若该子图中顶点 x_i 与 x_j 互不匹配时,则 M. M. S. (i,j)应等于子图 $G_{ij-1}(X_{ij-1},E_{ij-1})$ 和子图 $G_{i+1j}(X_{i+1j},E_{i+1j})$ 中二最大匹配数中较大者. 因此在上算法公式I、II中引入了量 M. M.S. $(i+1,j)$ 和 M. M. S.$(2,N)$.

依照上述的算法公式,我们可以依次地增大 j，减小 i 而可以逐次递归地求解各子图的最

大匹配数，最终可求得该拓扑平面图 $G(X, E)$ 的最大匹配数．另外在求解各子图的最大匹配数的同时，将各子图当具有最大匹配数时的弧 $\overset{\frown}{x_k x_j}$ 进行存储．因此一旦求得了该拓扑平面图 $G(X,E)$ 的最大匹配后，就可按照 Nussinov 提出的后退算法[3] 求得该具有最大匹配的拓扑平面图 $G(X,E)$ 的各弧．

3. 计算技术　根据上述的算法，我们在 TRS-80 的磁盘操作系统下以磁盘 BASIC II 语言编制了预测单链 RNA（或 DNA）分子的具有最大碱基配对数的二级结构的程序．在程序设计中，我们采取了模块与嵌套相结合的程序设计方法使该程序与我们的核酸序列资料管理系统有机地结合起来．由于算法需要建立存储各子图的最大匹配数 M. M. S. (i, j) 的矩阵和存储各子图当具有最大匹配时弧 $\overset{\frown}{x_k x_j}$ 的信息矩阵．它们都是$N \times N$的二维矩阵因此算法需要大量的内存空间．由于我们考虑的拓扑平面图是无向图，这二个矩阵是对称矩阵，为了节约内存空间我们以一个矩阵 $M(i, j)$ 加以存储．其中矩阵 $M(i, j)$ 的下三角块存储各子图的最大匹配数，而 $M(i,j)$ 的上三角块存储弧 $\overset{\frown}{x_k x_j}$ 的信息（这里只将 x_k 的足标 k 的序数存入矩阵）．因此该程序至少需占用$N \times N$的用户内存空间．由于 TRS-80 微型机的字长为 8 位，为节约内存在程序中我们尽最大可能将变量定义为整型量．

二、结果和讨论

为了说明该算法的预测能力，我们已对真核生物的 C. F. 5S RNA 和 B. M. 5S RNA[4]，鼠线粒体 DNA 中 12S rRNA 基因序列 3′ 末端区域[5]，丝状单链噬菌体 M 13 (f1, fd) 基因组中基因 IX 等序列[6]在 TRS-80 微型机上进行了处理，求得了它们各自的具有最大碱基配对的二级结构．

(1) 真核生物 C. F. 5S RNA 和 B. M. 5S RNA 的最大碱基配对的二级结构．

真核生物 C. F. 5S RNA 和 B. M. 5S RNA 序列全长 120 核苷酸，它们自身回折的形成的具有最大碱基配对数的折叠结构分别如图 2、3 所示．由于 TRS-80I 系统缺乏绘图设备，所以我们根据计算机的输出结果由人工进行作图．

(2) 鼠线粒体 DNA 中 12SrRNA 基因 3′ 末端片段序列[5] (560—439) 长 122 核苷酸，它自身回折形成的具有最大碱基配对的折叠结构如图 4 所示．

(3) 组状单链噬菌体 M 13、f1、fd 基因组中基因 IX 序列的最大碱基配对的二级结构．

组状单链噬菌体 M 13, f1, fd 基因组序列是十分相似的且它们基因组中基因 IX 序列是完全重合的 (1206—1304)．基因 IX 全长 99 个核苷酸，它是合成外壳蛋白质的结构基因．该结构基因自身回折而折叠成具有最大碱基配对的二级结构如图 5 所示．由图 2—5 可知，上述的各二级结构中都包含着一定数目的发夹结构．

总而言之借助于计算机处理，单链核酸分子最大碱基配对的折叠结构的预测算法可为我们提供 1 单链核酸分子二级结构的有关信息，这无疑是十分有益的．但是我们也应该注意到的是仅仅考虑单链核酸分子自身回折形成的最大碱基配对数而预测该分子的二级结构这仍然是不完全的．因此预测单链分子的二级结构算法还必须进一步完善．关于寻求单链核酸分子具有最小自由能的稳定结构，并使该结构能兼容其分子生物学有关信息的算法及其计算机处理，我们正在研究中．

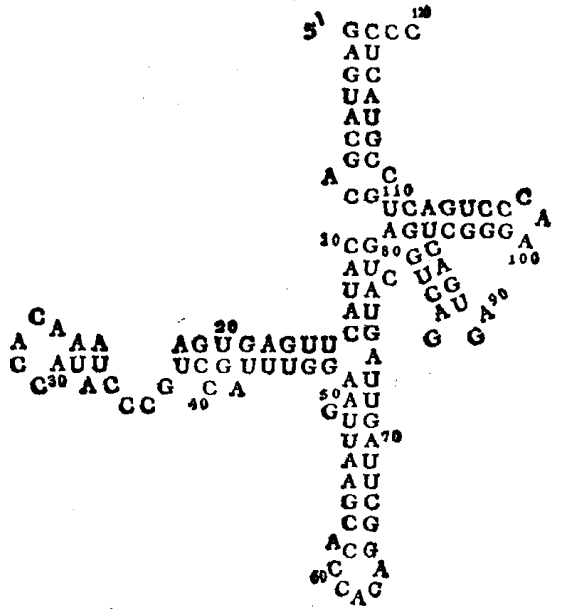

图 2 真核生物 C. F. 5S RNA 最大碱基配对的二级结构

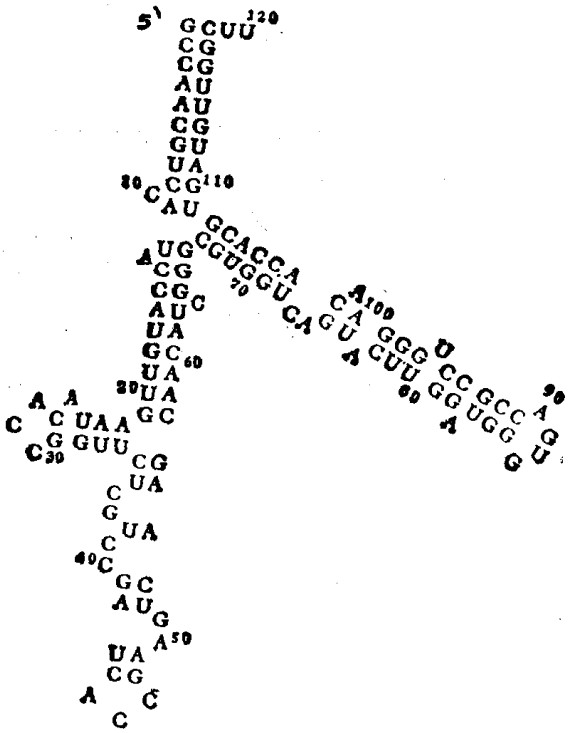

图 3 真核生物 B. M. 5S RNA 最大碱基配对的二级结构

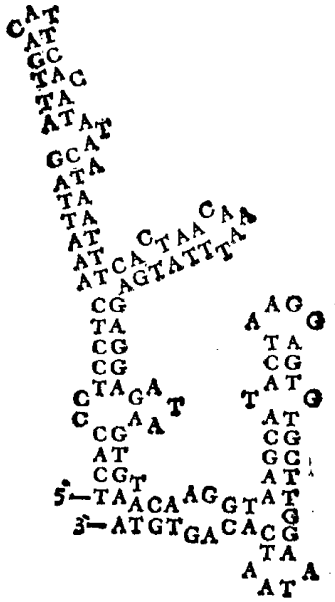

图 4 鼠线粒体 DNA 中 12 SrRNA 基因 3′ 末端 (560-439)区域折叠的最大碱基配对的二级结构

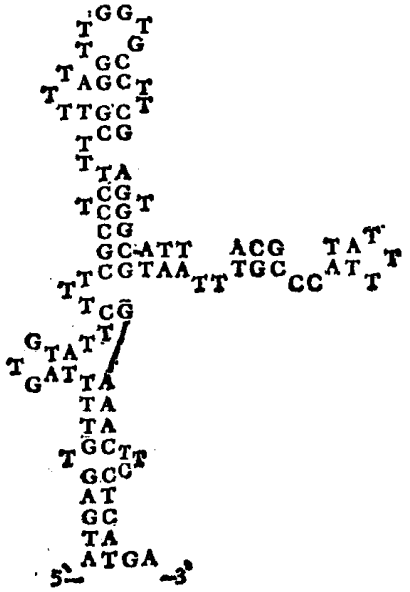

图 5 噬菌体 M13, f1, fd 基因组中基因 IX 序列的最大碱基配对的二级结构

参 考 文 献

[1] Tinoco, I., Uhlenbeck, O. C. & Levine, M. D., *Nature*, **230**(1971), 362—367.

[2] Salser, W., *Cold Spring Harbour Symp. Quant. Biol.*, **42**(1977), 985—1002.

[3] Nussinov, R., Pieczenik, G., Griggs, J. R. & Kleitman, D. J., *SIAM J. Appl. Math.*, **35** (1978), 1: 68—82.

[4] Erdmann, V. A., *Nucleic Acids Research*, **10**(1982), 2: r93—r115.

[5] Kobayashi, M., Seki, T., Yaginuma, K. & Koike, K., *Gene*, **16**(1981), 297—307.

[6] Beck, E. & Zink, B., *Gene*, **16**(1981), 35—38.