

基于离散 Hopfield 网络求解极大独立集的茎区选择算法以及在 RNA 二级结构预测中的应用

刘 琦¹⁾ 张 引²⁾ 叶修梓²⁾ 俞荣栋¹⁾

¹⁾(浙江大学沃森基因组科学研究院 杭州 310008)

²⁾(浙江大学计算机科学与技术学院 杭州 310027)

摘 要 提出了一种利用离散 Hopfield 网络求解图论极大独立集的启发式算法,并将其应用于 RNA 二级结构的茎区选择和预测当中. 算法通过映射 RNA 序列的茎区为无向图中的节点,将预测 RNA 二级结构的问题转化为求解图的极大独立集的问题. 定义了合理的能量变化函数,利用离散 Hopfield 网络进行迭代,以获得能量最优的预测结构. 文中将算法与传统的最大匹配数算法以及最小自由能算法在运行时间上进行比较,并且选择特定的序列在茎区和碱基对水平上进行精度测试,结果证明该算法在效率和精度上具有一定的优势. 算法的时间复杂性为 $\max\{O(n^2), O(N^2)\}$, 空间复杂度为 $O(N^2)$, 其中 n 为 RNA 序列长度, N 为 RNA 的茎区段个数.

关键词 RNA; 二级结构; 极大独立集; 离散 Hopfield 神经网络; 茎区

中图法分类号 TP181

A Discrete Hopfield Neural Network Based MIS Finding Algorithm for Stems Selecting and Its Application in RNA Secondary Structure Prediction

LIU Qi¹⁾ ZHANG Yin²⁾ YE Xiu-Zi²⁾ YU Rong-Dong¹⁾

¹⁾(James D. Watson Institute of Genomic Sciences, Zhejiang University, Hangzhou 310008)

²⁾(College of Computer Science and Technology, Zhejiang University, Hangzhou 310027)

Abstract Based on the definitions of Discrete Hopfield Neural Network (DHNN) and Maximal Independent Set (MIS) in Graph theory, a heuristic algorithm is presented to select stems in RNA structure as well as its application in RNA secondary structure prediction. The stems are mapped into an adjacent graph and the thermodynamic motion equation is defined to control the iterations of the network to find the optimal structure. Running time of this algorithm is compared with the Maximal Base Pair algorithm and Minimal Energy Algorithm. Also specific experiments are implemented to test the accuracy of the algorithm in stem and base pair levels. The results have shown that the algorithm is efficient in terms of its running time and accuracy. The time complexity of the algorithm is $\max\{O(n^2), O(N^2)\}$ and the space complexity approximates to $O(N^2)$, n is the sequence length of RNA and N is stem segments number of the RNA sequence.

Keywords RNA; secondary structure; Maximal Independent Set (MIS); discrete Hopfield neural network; stem

收稿日期:2005-10-28;最终修改稿收到日期:2007-10-08. 刘 琦,男,1981年生,博士研究生,研究方向为生物信息学、数据挖掘. E-mail: lq19811015@tom.com. 张 引,女,1970年生,副教授,从事图形与图像处理、多媒体信息处理和人工智能的研究. 叶修梓(通信作者),男,1966年生,教授,博士生导师,研究领域为CAD/CAM、计算机图形学、生物信息学. E-mail: xyz@zju.edu.cn. 俞荣栋,男,1981年生,博士研究生,研究方向为生物信息学、生物大分子可视化.

1 引 言

RNA 是一种重要的(单链)生物大分子,其结构可分为一级结构、二级结构和三级结构. RNA 的二级结构是由 RNA 单链自身回折而形成部分碱基配对和单链交替的茎环结构. 其中碱基互补配对形成的连续双螺旋区域称为茎区,不形成互补配对的单链结构称为环,依据形态的不同环又分为发夹环(hairpin loop)、内环(internal loop)、膨胀环(bulge loop)和多分支环(multibranched loop)等^[1]. RNA 的二级结构在由其一级结构推测三级结构的过程中起着桥梁的作用,并且在研究诸如 tRNA 和蛋白质的相互作用、mRNA 的稳定化处理等方面有着重要的应用. 对于 RNA 二级结构的研究具有极其重要的价值.

由于实验条件的限制,绝大多数的 RNA 分子的二级结构还不能用实验方法来测定,因此,结合计算机算法预测 RNA 二级结构显得尤为重要. 目前, RNA 二级结构预测的方法主要有如下 3 类:第 1 类是基于矩阵运算的动态规划算法,如 Nussinov 等提出的求最大碱基配对数结构的预测方法(以下简称最大匹配数算法)和 Zuker 等提出的求最小自由能结构的预测方法^[2-3](以下简称为最小自由能算法);第 2 类是最近几年发展起来的同源比对算法,如基于 RNA 碱基共变(covariation)思想的预测算法以及利用随机上下文无关文法进行二级结构建模的方法(Stochastic Context-Free Grammar, SCFG)^[4-6];第 3 类是基于茎区组合的启发式预测算法,如茎区最优堆积及其分布算法等等^[7-9]. 第 1 类算法在近二十年来得到了很大的发展,属于经典方法,但这类算法通过寻找较短子序列上的最优结构来寻找更长子序列上的最优结构,需要保留所有子序列的当前最优值,故算法的复杂度较高,这类算法的时间复杂度为 $O(n^3)$,空间复杂度为 $O(n^2)$,这里 n 为 RNA 分子所包含的碱基的数目^[10-12]. 特别的,当严格考虑到多分支环的能量时,最小自由能算法的时间复杂度为 $O(n^4)$,空间复杂度为 $O(n^3)$ ^[13]. 第 2 类方法近年来发展迅速,通常认为基于共变模型的结构预测更加可靠和有效^[14-15]. 但是由于其需要对于一组序列进行综合比较,故仍具有较高的计算复杂度^[6,16]. 第 3 类方法是当前研究的热点,由于茎区组合的确定意味着 RNA 二级结构的确定,而且,对于茎区组合的研究可以借鉴很多有效的启发式算法,故仍有

很大的讨论空间.

茎区组合算法进行结构预测的关键问题是如何有效地对整个茎区组合空间进行搜索,找到满足最优能量条件的 RNA 结构. 文献[7]中利用动态规划算法来组合茎区以获得 RNA 二级构象空间中的能量最优值,文献[8-9]则利用 RNA 茎区在已知二级结构中的分布概率来进行茎区的选择,二者的共同之处在于都是利用一定的启发式经验,来控制茎区的组合,以获得最优的结构. 本文在结合图论的独立集思想的基础上,将 RNA 的茎区结构映射为无向图中的节点,提出了一种基于离散 Hopfield 网络(Discrete Hopfield Neural Network, DHNN)^[17]求解图的极大独立集的迭代算法来确定 RNA 序列的二级结构,在求解极大独立集的过程中,定义了网络节点的能量变化函数,以保证搜索茎区的能量组合在当前条件下最优. 网络收敛时图所对应的极大独立集即可以作为序列的预测结构. 本文最后通过算法的效率和精度试验,进一步说明了方法的有效性.

2 方 法

2.1 RNA 二级结构

通常用 R 表示一个长度为 n 的 RNA 分子序列,记

$$R=r_1r_2\cdots r_n, \quad r_i \in \{A, U, G, C\}, \quad i=1, 2, \dots, n \quad (1)$$

这里 A, U, G, C 分别表示四种碱基. 在 RNA 分子序列中,碱基分为两类:一类是按照碱基互补配对原则(C-G, A-U 或 G-U)形成碱基对的匹配碱基,一类是没有配对而形成环的未配对碱基. 一个典型的 RNA 二级结构的示意图如图 1 所示. 下面给出 RNA 分子二级结构的定义^[1].

定义 1. 假设 R 是一个如式(1)所示的 RNA 分子,定义 R 的二级结构: $r=\{(i, j), 1 \leq i \leq j \leq n, r_i \text{ 与 } r_j \text{ 配对}\}$, 且满足

- (1) 对 $(i, j) \in r, (i', j') \in r$, 若 $i \leq i' \leq j \leq j'$, 则必有 $i=i', j=j'$;
- (2) 若 $(i, j) \in r$, 则 $j-i > 3$.

定义 2. 两个碱基对 (i, j) 和 (i', j') 相容(compatible), 当且仅当

- (1) $i=i'$ 并且 $j=j'$,
- (2) $i < j < i' < j'$ 或者
- (3) $i < i' < j' < j$.

由定义 1 可以看出, RNA 的二级结构的本质

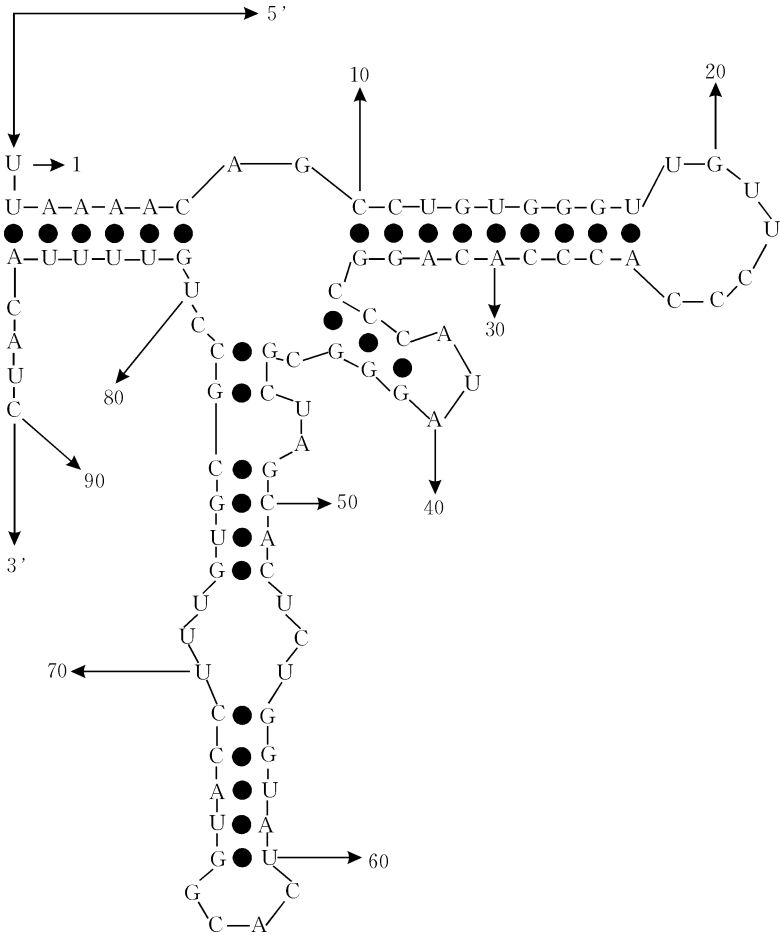


图 1 RNA 二级结构示意图(图中 A,G,C,T 分别代表碱基,‘●’代表碱基的配对,数字表示碱基的序号)

实际上即是碱基的配对,而连续的碱基对即构成了茎区,定义 2 指出我们这儿讨论的结构中碱基对是相容的,即不包括重叠(overlapping)和交叉(pseudo-knots)的情况. 下面给出茎区的严格定义.

定义 3. 假设 R 是长度为 n 的 RNA 分子, $R_f = r_i r_{i+1} \cdots r_{i+k-1}$ 和 $R_l = r_j r_{j-1} \cdots r_{j-k+1}$ 是 R 的两个子序列,如果 R_f 和 R_l 中的碱基依次互补匹配,则称 R_f 和 R_l 形成一个茎,记为 $S(i, j, k)$,并称 R_f 为茎 S 的前段, R_l 为茎 S 的后段, k 为茎 S 的长度,且定义 $L = |j - i|$ 为茎区的环(loop)长度. 通常定义茎区 $k \geq 3$,即连续 3 个以上的碱基对可以称作茎区.

由定义 1~定义 3 可以推出如下结论:
设 (S_1, S_2, \cdots, S_l) 是定义在 R 上的一个茎序列,如果任意两个茎 S_i 和 S_j 既不重叠(overlapping)也不交叉(pseudo-knots),我们称之为相容(compatible)^[18],则 (S_1, S_2, \cdots, S_l) 可以唯一确定 R 上的一个二级结构,记这个二级结构为 $S = (S_1, S_2, \cdots, S_l)$.

由上述结论可知,选择 RNA 的茎区是形成 RNA 二级结构的关键. 目前比较公认的 RNA 折叠方向是向其最低自由能结构折叠,但是并不一定是

完全最低,所以选择标准是使该 RNA 的二级结构稳定,即能量达到最低或者极低. 下文论述的算法即是在这个前提下利用独立集的思想进行茎区的选择.

2.2 基于 DHNN 求解极大独立集的 RNA 二级结构预测算法

2.2.1 基于 DHNN 的极大独立集求解

由上文可知,相容的茎区组合可以确定一个 RNA 的二级结构,构成一个 RNA 二级结构的茎区集合是所有理论茎区的一个子集,如何从这些理论茎区中选择出这样一个子集,是算法的核心所在. 选择的茎区应该符合以下两个条件:(1) 茎区之间互相相容;(2) 选择的茎区可以增加当前 RNA 二级结构的能量稳定性. 对于第一个条件,引入了图论独立集的思想,同时利用 DHNN 进行茎区选择;对于第二个条件,定义适当的网络能量函数来控制网络的迭代. 下面给出算法中相关概念的理论描述和严格定义.

首先,将 RNA 的所有理论茎区映射为一个无向图中的节点,并且假设茎区与茎区之间不相容则

以一条边连接,那么,搜索一个相容茎区组合即相当于在该无向图中确定一个节点的集合,该集合中任意两节点之间都没有边相连,即互不相邻,这样一个集合称为独立集.

下面给出独立集的严格定义^[19].

定义 4. 给定一个无向图 $G=(V, E)$, 其中 $V=\{V_1, V_2, \dots, V_n\}$ 是图 G 的节点集, $E \subseteq V \times V$ 是图 G 的边集. 称 V 的一个子集 $S \subseteq V$ 为独立集 (Independent Set, IS), 当且仅当它的节点两两互不相邻. 如果一个独立集不是其它任何一个独立集的真子集, 则称该独立集为图 G 的极大独立集.

从能量预测的角度来说, RNA 的预测二级结构可以认为是一个茎区的极大独立集 (当然, 实际的二级结构可能仅仅是一个独立集), 我们引入 DHNN

进行茎区的选择. 在进行 DHNN 迭代之前, 我们需要对于茎区映射图作以下变换:

设由一条 RNA 序列的所有理论茎区所映射成的无向图为 G , V 是图 G 的节点集, 将每一个茎区节点 V_i 分为前段 V_i^f 和后段 V_i^l 两部分, 则所有茎区节点由此分解得到茎区段集合 $V' = \{V_1^f, V_2^f, \dots, V_n^f, V_1^l, V_2^l, \dots, V_n^l\}$, 将茎区段集合 V' 中的所有节点按其在序列 5' 到 3' 段的出现先后次序进行排序, 按圆周分布, 构成圈图 (circle graph) G' , 图中配对的茎区前后段用一条边相连, 如果两个茎区不相容, 则在圈图中其两条边相交, 搜索的相容茎区在圈图中即为互不相交的边的集合. 故求解原无向图 G 中的极大独立集问题可以转化成求解圈图 G' 中的极大平面图 (planer graph) 问题, 如图 2 所示.

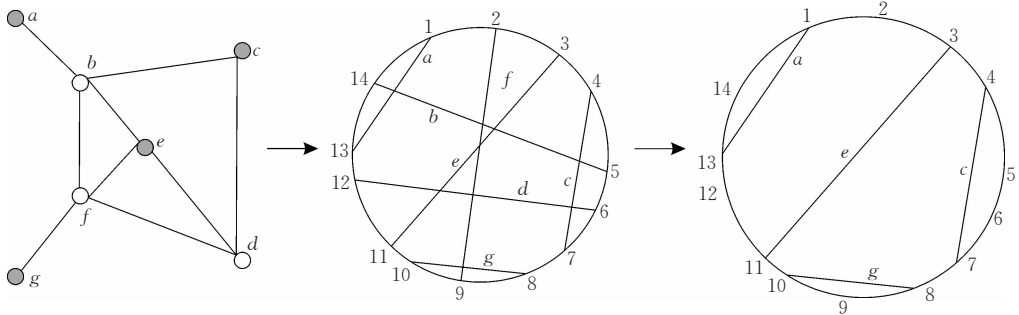


图 2 茎区节点无向图到圈图的变换 (左侧子图为茎区节点无向图 G , 其某一个极大独立集为 $\{a, c, e, g\}$, 右侧子图为转换为圈图 G' 后的极大平面图)

基于以上变换, 我们利用 DHNN 求解圈图中的极大平面图. 圈图中的边对应于 DHNN 的神经元, 每个神经元 V_i 只取二元的离散值 0 和 1. 神经元的状态变化规则如下:

$$V_i = \begin{cases} 1, & U_i \geq 0 \\ 0, & U_i < 0 \end{cases} \quad (2)$$

式中, V_i 为神经元的输出, U_i 为神经元的输入. 定义 $V_i=1$ 表示边 i 不出现在圈图 G' 中, $V_i=0$ 则表示边 i 出现在圈图 G' 中. 同时定义神经元随迭代次数的能量变化率如式 (3):

$$\begin{aligned} \frac{dU_i}{dt} = & A \times \left(\sum_{j=1}^n d_{ij} (1-V_j) (\text{distance}(i))^{-1} \right) \times \\ & (1-V_i) \times p(i)^{-1} - \\ & B \times h \left(\sum_{j=1}^n d_{ij} (1-V_j) \right) \times V_i \times p(i) \end{aligned} \quad (3)$$

式 (3) 中 d_{ij} 取 1 或 0 二值, 分别代表边 i 和边 j 是否相交. $h(x)$ 为激励函数, 当 $x=0$ 时, $h(x)=1$, 其它情况为 0. $\text{distance}(x)$ 为茎区 x 的环长度. 式中前

项为惩罚项, 后项为激励项, 参数 A 和 B 分别为前后项的权值, 一般 $A=B=1$ (后续试验均取该值). 当边 i 出现在圈图中且与其它边相交时, 前项取正值, 将会使 V_i 在下一次迭代时向 0 发展, 即不出现在圈图中. 而如果边 i 未出现在圈图中但与其它边均不相交, 则后项为正值, 将使 V_i 在下一次迭代时向 1 发展, 即出现在圈图中. $p(i)$ 为茎区的能量函数, 用来控制惩罚项和激励项的变化速度. 对于给定茎区 $i=S(i, j, k)$, $p(S(i, j, k)) = |e(r_{i+1} r_{j-1}) + \dots + e(r_{i+k-1} r_{j-k+1})|$, 其中每个分项 $e(x)$ 表示配对碱基的能量, 在这里我们采用文献 [10] 提出的一种国际上比较通用的自由能记分规则——最近-邻居规则 (nearest-neighbor rule) 来计算茎区能量 (如表 1 所示), 具体数值取表中数值的绝对值 (原值均为负值). 从能量角度来说, $p(i)$ 越大表示茎区能量越低越稳定, 这样的茎区出现在 RNA 二级结构中的概率越大.

表 1 最近-邻居规则的碱基对能量(kcal/mole at 37℃)^[10]

	碱基对能量					
	A/U	C/G	G/C	U/A	G/U	U/G
A/U	−0.9	−1.8	−2.3	−1.1	−1.1	−0.8
C/G	−1.7	−2.9	−3.4	−2.3	−2.1	−1.4
G/C	−2.1	−2.0	−2.9	−1.8	−1.9	−1.2
U/A	−0.9	−1.7	−2.1	−0.9	−1.0	−0.5
G/U	−0.5	−1.2	−1.4	−0.8	−0.4	−0.2
U/G	−1.0	−1.9	−2.1	−1.1	−1.5	−0.4

注:该表表示两个邻近碱基对所形成的茎区结构的能量值.举例说明,第1行第2列的能量值为−1.8,表示由当前碱基对C/G和其邻接碱基对A/U形成的茎区结构能量值为−1.8.

文献[20]采用了类似于式(3)的能量变化函数求解图的极大独立集,并将其应用于RNA二级结构的预测中,但是文献[20]是从碱基对的角度出发,并且忽略了碱基对的能量权值,而本文方法从茎区角度进行宏观考虑,同时引入了茎区能量函数的权值控制方法,具有更加合理和明确的生物学意义.

基于式(2)和(3),适当选取初始参数,对DHNN进行迭代,最后收敛于一个能量最优点,作为RNA二级结构的预测结果.需要说明的是,利用DHNN求解极大独立集,其收敛性是可以得到保证的^[21-22],在此不再赘述.

2.2.2 算法流程

下面给出利用DHNN进行茎区选择的算法流程:

1. 求解RNA序列的所有理论茎区(算法见文献[18]).
2. 基于DHNN求解极大独立集的RNA二级结构预测.
- 步2的具体描述如下:

输入:以RNA序列所有理论茎区段为节点组成的圈图 $G'(V,E)$,DHNN迭代的最大次数 max_t 以及DHNN网络神经元的随机初始值 $U_i(0), i=1,2,\cdots,n$, n 为圈图中的节点个数

输出:RNA二级结构序列 $S=(S_1^f, S_2^f, \cdots, S_p^f, S_1^b, S_2^b, \cdots, S_p^b)$,其中 S_i^f 为茎区前段, S_i^b 为茎区后段

- (1) 设开始迭代次数 $t=0$;
- (2) 利用式(2)计算 $V_i(t)$,其中 $i=1,2,\cdots,n$;
- (3) 利用式(3)计算 $\Delta U_i(t)$,其中 $i=1,2,\cdots,n$;
- (4) $U_i(t+1)=U_i(t)+\Delta U_i(t)\Delta t, t=t+1$,如果 $U_i(t)=U_i(t-1)(i=1,2,\cdots,n)$ 或者 $t=max_t$,则终止算法,返回茎区段集合;否则转到步(3).

整个算法的复杂度分析如下:步1求解序列的所有理论茎区算法其时间复杂度不低于 $O(n^2)$,这里的 n 为序列长度;步2的时间复杂度依赖于DHNN的具体迭代,如果采用并行机进行迭代,则可以提高计算效率,其时间复杂度约为 $O(N^2)$,这里的 N 为所有理论茎区段个数;故该算法的总时

间复杂性为 $\max\{O(n^2), O(N^2)\}$,空间复杂度为 $O(N^2)$,相对于前面所述的传统算法来说有一定程度的提高.

3 实验

我们在Windows平台下实现了该算法,下面从算法的效率和精度两个方面对其进行实验测试.本算法的迭代次数 max_t 均为500次.

3.1 算法效率测试

由于RNA分子的实际结构很难得到,我们将算法的性能与经典的最大匹配数算法^[2]以及目前较权威的最小自由能算法^[3](其实现的Windows版本为RNAstructure(Version4.2))做了比较,从EMBL数据库内随机选取长度为1000~5000不等的RNA序列数据进行测试,在P4 2.4GHz/512MB内存的PC上的实验结果如表2所示.

表 2 算法平均耗时比较

序列长度	最大匹配数 算法平均耗时/s	最小自由能 算法平均耗时/s	本文算法 平均耗时/s
1000	55	34	0.9
2000	560	265	1.8
3000	∞	840	5.0
4000	∞	∞	10.0
5000	∞	∞	20.0

由表2可见,当序列长度为1000时,本文算法效率分别为最大匹配数算法的61倍、最小自由能算法的37倍;当序列长度为2000时,本文算法效率分别为最大匹配数算法的311倍、最小自由能算法的147倍.本文算法效率随序列长度的增加而显著,而当序列长度大于4000时,上述两种算法均已失效,而本文的算法耗时增加缓慢,其主要原因在于DHNN的并行迭代算法可以大大提高空间搜索的效率.

3.2 算法精度测试

由于算法是从茎区组合的角度来考虑RNA二级结构,故首先从茎区角度衡量算法的精度.从EMBL中随机选择长度介于50~200之间不等的30个tRNA,测试用本文算法预测得到的茎区和用RNAstructure预测得到的茎区的相同程度.试验结果如图3所示,图中横坐标表示tRNA的序号,共30个,纵坐标表示预测得到的茎区个数,以“*”构成的折线为用RNAstructure预测的30个tRNA二级结构茎区的个数曲线,以“.”构成的折线为同时出现在我们算法预测结果和RNAstrucure预测

结果中的茎区个数曲线,从图中可以看出,我们的预测结果基本上与 RNAstructure 预测的结果吻合。

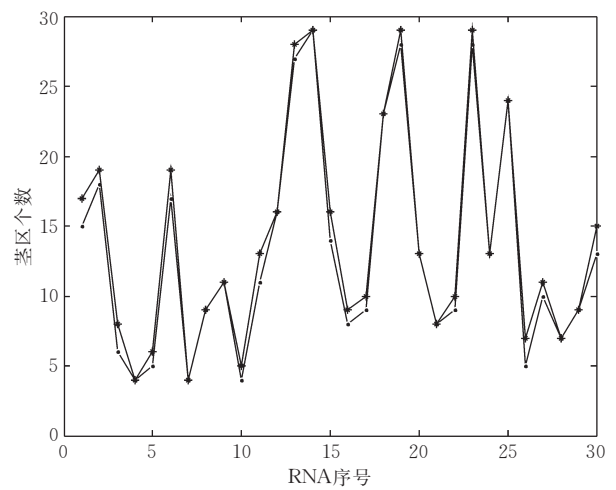


图 3 茎区水平的精度测试

其次,从碱基对的水平,将本文的算法与文献[20]以及 Zuker 的最小自由能算法进行比较.采用文献[23]中已试验测定二级结构的 RNA 序列,衡量特定位置碱基在预测和实测中的差别,并且引入 Matthews 系数^[24]来定量描述预测精度. Matthews 系数(CC)定义为 $CC \approx \sqrt{\frac{TP}{TP+FN} \cdot \frac{TP}{TP+FP}}$, $0 \leq CC \leq 1$,式中 TP, FP 和 FN 分别为预测的真阳率、假阳率和假阴率. CC 可以作为综合衡量预测灵敏性(sensitivity)和特异性(specificity)的参数指标,其值越接近 1 表示预测的精度越高. 预测的结果如表 3 所示,其中本文的算法相对于文献[20]的算法在精度上有所提高,相对于 Zuker 的最小自由能算法在精度上还具有一定的差距,但是也处于可以接受的范围.

表 3 碱基对水平的精度测试

	平均 CC 值		
	10 α operon mRNA leader sequences	11 S15 mRNA sequences	18 viral 3'-UTR sequences
本文算法	0.26	0.40	0.59
文献[20]算法	0.22	0.30	0.50
Zuker 最小自由能算法	0.35	0.49	0.72

4 讨 论

本文提出了一种基于 DHNN 求解极大独立集的茎区选择算法,并将其应用于 RNA 二级结构预测中,这种方法既考虑到了整体结构中茎区的稳定性,又兼顾了热力学中能量越小越稳定的原理. 具有

预测速度快、能处理的问题规模大等优点. 通过与其它算法的比较以及实际数据的测试,结果证明该算法用于 RNA 二级结构预测具有一定的时间和精度的优越性. 特别的,算法对于长度长的 RNA 序列(序列长大于 3000)十分有效,有利于处理长 RNA 序列. 由于本算法以对基于 RNA 茎区的计算代替对序列的整个长度计算,而 RNA 茎区的个数远小于 RNA 的序列长度,从而显著降低了时间和空间的复杂度. 其时间复杂性为 $\max\{O(n^2), O(N^2)\}$, 空间复杂度为 $O(N^2)$,其中 n 为 RNA 序列长度, N 为 RNA 的茎区段个数.

算法中引入了茎区能量函数 $p(i)$ 来定义茎区的能量权值,并且采用常用的最近-邻居规则来计算茎区能量,相对来说明确简洁. 为了提高预测的精度,我们可以结合 Zuker 关于茎区能量的一些更复杂精确的定义,例如:记原始一级序列的能量值为 $E_0=0$,生成第 i 个茎区时的能量为 $\Delta E(S_i)$,其中形成茎区的每个配对碱基配对反应时释放的能量的总和为 $E_{\text{stack}}(S_i)$,为负;由于形成 S_i 而破坏的原有的环的能量的总和为 $E_{\text{destroy}}(S_i)$,为正;由于形成 S_i 而新生成的环的能量总和为 $E_{\text{construct}}(S_i)$,为正,则有 $\Delta E(S_i) = E_{\text{stack}}(S_i) - E_{\text{destroy}}(S_i) + E_{\text{construct}}(S_i)$,该式相对于我们的能量函数对于茎区能量的计算将会更加准确,但是所要花费的时间也更多,效率会有所降低.

另外,从本文算法的效率和精度上综合考虑,虽然在碱基对的精度上相对于传统的 Zuker 自由能算法还具有一定差距,但仍属于可以接受的范围,同时由于采用了 DHNN 进行解空间的搜索,算法的效率方面得到了很大提高,相信通过更加合理的网络能量函数的设计,精度水平还可以得到进一步提升,这也是我们下一步研究工作的重点.

参 考 文 献

[1] Sankoff D, Kruskal J, Mainville S, Cedergren R. Fast algorithms to determine RNA secondary structures containing multiple loops//Sankoff D, Kruskal J. Time Warps, String Edits, and Macro-Molecules: The Theory and Practice of Sequence Comparison. Chapter 3. Reading, MA: Addison-Wesley, 1983

[2] Nussinov R, Jacobson A B. Fast algorithm for predicting the secondary structure of single strand RNA. Proceedings National Academy of Sciences, 1980, 77(11): 6309-6313

[3] Zuker M. Optimal computer folding of large RNA sequence using thermodynamics and auxiliary information. Nucleic Acids Research, 1981, 9(1): 133-148

- [4] Searls D. The linguistics of DNA. *American Scientist*, 1992, 80(4): 579-591
- [5] Searls D. The computational linguistics of biological sequences//Hunter L. *Artificial Intelligence and Molecular Biology*. Menlo Park, California: AAAI Press, 1993: 47-120
- [6] Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 1999, 15(6): 446-454
- [7] Gao Qiong, Mo Zhong-Xi, Zheng Zhou. An energy based dynamic partitioning algorithm for RNA secondary structure prediction. *Journal of Mathematics*, 2003, 23(1): 43-48(in Chinese)
(高琼,莫忠息,郑卓. 一种基于能量的 RNA 二级结构预测的动态划分算法. *数学杂志*, 2003, 23(1): 43-48)
- [8] Li Wu-Ju, Wu Jia-Jin. Prediction of RNA secondary structure based on random stacking of helical regions. *Acta Biophysica Sinica*, 1996, 12(2): 214-218(in Chinese)
(李伍举,吴加金. 基于螺旋区随机堆积的 RNA 二级结构预测. *生物物理学报*, 1996, 12(2): 214-218)
- [9] Li Wu-Ju, Wu Jia-Jin. Prediction of RNA secondary structure based on helical regions distribution. *Bioinformatics*, 1998, 14(8): 700-706
- [10] Turner D H, Sugimoto N. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry*, 1988, 17: 167-192
- [11] David H. Expanded sequence dependence of Thermodynamic Parameter Improves Prediction of RNA secondary Structure. *Journal of Molecular Biology*, 1999, 288(5): 911-940
- [12] Setubal, Meidanis. *Introduction to Computational Molecular Biology*. Boston: PWS Publishing Company, 1997
- [13] Waterman. *Introduction to Computational Biology*. New York: Chapman & Hall, 1995
- [14] Holmes I, Rubin G M. Pairwise RNA structure comparison with stochastic context-free grammars//*Proceedings of the Pacific Symposium on Biocomputing*, 2002: 163-174
- [15] Cai L, Malmberg R L, Wu Y. Stochastic modeling of RNA pseudoknotted structures: A grammatical approach. *Bioinformatics*, 2003, 19(Suppl. 1): I66-I73
- [16] Eddy S R, Durbin R. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 1994, 22(11): 2079-2088
- [17] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings-National Academy Of Sciences USA*, 1982, 79(8): 2554-2558
- [18] Jiang Tao, Xu Ying, Zhang Michael Q. *Current Topics in Computational Molecular Biology*. Cambridge, Massachusetts, London, England: The MIT Press, 2002
- [19] Li Jing, Liu Chang-Lin. On the theory of maximal independent sets and a generating method in a graph. *Acta Electronica Sinica*, 1995, 23(8): 78-79(in Chinese)
(李兢,刘长林. 关于图的极大独立集的理论及生成方法. *电子学报*, 1995, 23(8): 78-79)
- [20] Takefuji I, Chen L-L, Lee K-C, Huffman J. Parallel algorithms for finding a near-maximum independent set of a circle graph. *IEEE Transactions on Neural Networks*, 1990, 1(3): 263-267
- [21] Li You-Mei, Xu Zong-Ben, Miao Duo-Qian. A heuristic neural network algorithm for MIS problem. *Pattern Recognition and Artificial Intelligence*, 2003, 16(1): 76-80(in Chinese)
(李有梅,徐宗本,苗夺谦. 用神经网络启发式算法求解最大独立集问题. *模式识别与人工智能*, 2003, 16(1): 76-80)
- [22] Li You-Mei, Xu Zong-Ben, Sun Jian-Yong. A hybrid algorithm for MIS problem. *Chinese Journal of Computers*, 2003, 26(11): 1538-1545(in Chinese)
(李有梅,徐宗本,孙建永. 一类求解最大独立集问题的混合神经演化算法. *计算机学报*, 2003, 26(11): 1538-1545)
- [23] Ji Y, Xu X, Stormo G D. A graph theoretical approach to predict common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, 2004, 20(10): 1591-1602
- [24] Gorodkin J, Stricklin S L, Stormo G D. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Research*, 2001, 29(10): 2135-2144



LIU Qi, born in 1981, Ph. D. candidate. His research interests include bioinformatics and data mining.

research interests include graphic and image processing, multi-media informational processing and AI.

YE Xiu-Zi, born in 1966, professor, Ph. D. supervisor. His research interests include CAD/CAM, CG and bioinformatics.

YU Rong-Dong, born in 1981, Ph. D. candidate. His research interests include bioinformatics and visualization of biological molecular.

ZHANG Yin, born in 1970, associate professor. Her

Background

Ribonucleic acid (RNA) is an important class of molecules which performs a wide range of biological and chemical

functions. Traditionally, most RNA molecules were regarded as involving in the process of translation, including transfer

RNA (tRNA) and ribosomal RNA (rRNA). Since the late 1990s, it has been widely acknowledged that there substantially exists other types of RNA molecules (such as small non-coding RNA) presented in organisms ranging from bacteria to mammals, which affect a large variety of processes including drugs targeting, riboswitches, plasmid replication, phage development, bacterial virulence, RNA modification and others. RNA has recently become the center of much attention because of its diversity functions, leading to a substantially increased interest in obtaining their structural information.

The problem of RNA secondary structure prediction has been an interesting and challenging one for several decades, due to its importance to determination of the RNA tertiary structures and RNA functions as well as the difficulty in solving the problem. Numerous prediction methods have been

developed, which include the thermodynamic energy minimization method, the phylogeny-based comparison method and the stochastic context-free grammar method. However, most of these methods are not suitable for the prediction of long RNA sequences and large-scale data analysis due to their high computational consuming. In this paper, the authors try to address this issue from a graph perspective and utilize Discrete Hopfield Neural Network (DHNN) to improve the processing speed. The results have shown that the algorithm is efficient in terms of its running time and prediction accuracy.

This paper is a primary summarize of their early stage researches of parallelized RNA secondary structure prediction. The ultimate goal is to utilize parallelized clustering system to perform RNA secondary structure prediction, which could dramatically improve the prediction speed and make the more accurate prediction feasible.