PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2009, 36(1): 115~121
www.pibb.ac.cn

基于局部茎搜索的 RNA 二级结构预测算法 *

陈 翔[1,2,4] 卜东波[1,2] 张 法[1,2] 高 文[3]**
([1]中国科学院智能信息处理重点实验室，北京 100190；[2]中国科学院计算技术研究所，北京 100190；
[3]北京大学信息科学技术学院，北京 100080；[4]中国科学院研究生院，北京 100049)

摘要 RNA 的二级结构预测是生物信息学中一个已经有 30 多年历史的经典问题，基于最小自由能模型(MFE)的优化算法是使用最为广泛的方法．但 RNA 结构中假结的存在使 MFE 问题理论上成为一个 NP-hard 问题，即使采用动态规划等优化算法也会面临时间复杂度高的困难，同时研究还发现，由于受 RNA 折叠动力学机制以及环境因素的影响，真实的 RNA 二级结构往往并不处于自由能最小状态．根据 RNA 折叠的特点，提出了一种启发式搜索算法来预测带假结的 RNA 二级结构．该算法以 RNA 的茎为基本单元，采用启发式搜索策略在茎的组合空间中搜索自由能最小并且出现频率最高的 RNA 二级结构，该算法不仅能显著降低搜索 RNA 二级结构的时间复杂度，还有助于弥补单纯依赖能量预测 RNA 二级结构的不足．在多种类型的 RNA 标准数据集上进行了检验，结果表明，该算法在预测的精度上优于目前国际上几个著名的 RNA 二级结构预测算法并且具有较高的运行效率．

关键词 RNA 二级结构预测，假结，NP-hard，启发式算法
学科分类号 TP319，Q7
DOI: 10.3724/SP.J.1206.2008.00329

随着 21 世纪初人类基因组测序的完成，如何破译大量的基因信息，获知生物分子的生物学功能就成为后基因组时代的重要任务．RNA 作为三种最重要的生物大分子之一(另外两种是 DNA 和蛋白质)，担负着重要的生物功能．而 RNA 的这些功能又是通过其结构(包括二级和三级结构)来实现的[1]．因此，获得 RNA 的结构信息将对其功能的发现具有极其重要的意义．RNA 的一级结构用实验的方法容易测定，但是其二级和三级结构目前用实验的方法来测定还十分困难，因此通过计算的方法来预测 RNA 的结构成为生物信息领域一个重要任务和热点问题．

本文主要研究 RNA 的二级结构预测问题．RNA 二级结构的预测算法已经有 30 多年的发展历史，从算法所基于的生物学原理上可分为两大类：一类是基于序列比对的算法[2]；另一类是基于自由能最小的算法．前者必须有一组具有较高序列相似性的 RNA 序列作为比对的模板，因而不适用于对单条 RNA 序列进行预测．而后者则是从单条序列出发，通过计算 RNA 序列自身的碱基配对产生的最小自由能来预测 RNA 的二级结构，因而具有更大的实用性．

目前，基于自由能最小的 RNA 二级结构预测算法主要有两类，它们分别是：基于矩阵的动态规划算法和基于启发式规则的随机(局部)搜索算法．

基于矩阵的动态规划算法是目前应用最为普遍的算法，这种算法可以得到一个 RNA 序列在给定的热力学模型(能量模型)下具有最小自由能的二级结构．但这类算法即使在简化的能量模型下时间复杂度也会达到 $O(n^3)$．若考虑更准确的多分支环能量函数，则时间复杂度将达到 $O(n^4)$[3,4]．更为困难的是不少 RNA 的二级结构中还包含一种特殊而又重要的非嵌套结构——假结．假结的出现破坏了动态规划算法所依赖的 RNA 二级结构的嵌套子结构性质，尽管有一些动态规划算法通过限制假结的类

* 国家重点基础研究发展计划(973)资助项目(2002CB713807)，中国科学院前沿知识创新项目(20076020)和国家自然科学基金资助项目(60503060，90612019，60752001).
** 通讯联系人.
Tel: 010-62758602, E-mail: wgao@pku.edu.cn
收稿日期：2008-05-06，接受日期：2008-06-26

型(比如说，只允许出现几种比较简单的假结)来预测 RNA 带假结的二级结构．它们的时间复杂度也达到 O(n^5)~O(n^6)[5~7]．这严重制约了算法所能处理问题的规模，从而使带假结的 RNA 二级结构预测成为一个难题．另一方面，由于当前能量函数本身的缺陷以及 RNA 折叠动力学因素的影响，使得当前 RNA 的二级结构往往并不是处于自由能最小的状态．这就使 RNA 二级结构预测存在一个矛盾的问题:一方面我们仍然以自由能最小作为优化目标并用各种优化计算方法力图得到 RNA 的最小自由能状态；另一方面我们耗费了大量时间和空间所得到的最小自由能的结构却往往不是真实的 RNA 结构[8]．而启发式搜索算法则可从一定程度上弥补这个缺陷．

基于启发式规则的随机(局部)搜索算法通过引入一些启发式规则，如遗传算法中的遗传、变异机制、模拟退火算法中的退火机制等等，在 RNA 二级结构的解空间中进行局部搜索直到满足一定的终止条件为止[9~11]．通常这样的搜索方法不能确保得到自由能最小的全局最优解，但却仍然适于用来预测 RNA 的二级结构特别是带假结的 RNA 二级结构，原因在于：a．找到包含假结的 RNA 二级结构的最小自由能已经被证明是 NP 完全问题，不存在合理时间范围内的最优算法．b．越来越多的实验发现，RNA 的真实二级结构实际上并不一定处于自由能最小的状态，而是处于自由能次小状态，因而局部搜索得到的次优解并不一定比最优解差． c．基于局部搜索的算法可以更方便地采用更准确的自由能函数和更强描述能力的假结模型．d．基于局部搜索的算法还可以根据 RNA 折叠的动力学特点灵活地采用启发式规则，以弥补单纯依赖自由能模型的缺陷．

本文提出了一种基于茎的组合算法(StemFind)来预测包含假结的 RNA 的二级结构．该算法以逐步降低能量的方式来不断累积茎，不同于以往基于茎组合算法，我们在这个过程中通过启发式策略来选择合适的茎．通过多次重复这个过程我们得到多个候选解，然后我们以能量和茎出现频率两个指标来评估这些候选解并生成最后的结构．

为了证明 StemFind 的有效性，我们在广泛的标准测试集上进行了测试．测试结果表明，StemFind 不仅在预测包含假结的 RNA 二级结构的准确率上明显高于目前最为著名的几个 RNA 二级结构预测软件，同时在预测不带假结的 RNA 二级结构方面能达到目前常用软件 Mfold 的标准．而且 StemFind 在时间复杂度上要大大低于 PKNOTS 等软件．

1 RNA 二级结构预测的形式化定义

为了更好地理解 RNA 二级结构的特点，我们先用形式化语言来描述 RNA 二级结构及其基本特性．

定义 1：一个长度为 n 的 RNA 序列 $R=r_1r_2r_3\cdots r_n$．($r_i\in\{A, C, G, U\}$，$i=1, 2, 3, \cdots, n$)的二级结构定义为配对碱基的集合 $P=\{(r_i, r_j)\}$，其中(r_i, r_j)满足：

i) $(r_i, r_j)\in\{(A, U), (U, A), (G, C), (C, G), (G, U), (U, G)\}$，$1\leqslant i\leqslant j\leqslant n$ 且 $j-i>3$；

ii) 若$(r_i, r_j)\in P$ 且$(r_i, r_k)\in P$，则 $j=k$；

定义 2：设 R 是一个长度为 n 的 RNA 序列．$R_1=r_ir_{i+1}\cdots r_{i+k-1}$ 和 $R_2=r_jr_{j-1}\cdots r_{j-k+1}$ 是 R 的两个子序列．若 R_1 和 R_2 中的碱基依次互补配对(即$(r_{i+t}r_{j-t})$满足定义 1 中的性质 ii)，$t=0, 1, \cdots, k-1$)，则称 R_1 和 R_2 在 R 的二级结构中构成一个茎(stem)．记为 $s(i, j, k)$，其中 i, j 分别表示茎在序列 R 中的 5′端起始位置和 3′端结束位置；k 表示茎的长度．

定义 3：给定两个茎 $s_1(i_1, j_1, k_1)$和 $s_2(i_2, j_2, k_2)$，若 s_1，s_2 不发生重叠，则称 s_1 和 s_2 相容．

性质 1：设 $S=\{s_1, s_2, \cdots, s_m\}$是 RNA 序列 R 中的一个茎的集合，若其中任意两个茎都相容，则 S 可以唯一地确定 R 上的一个二级结构．该性质说明 RNA 的二级结构可由茎的组合唯一确定．

依据以上的定义和性质，我们可把 RNA 的二级结构预测问题形式化为一个 RNA 的相容茎区的组合优化问题．

2 方　　法

RNA 二级结构是经过复杂的折叠过程形成的，这个过程也是 RNA 分子由高能态向低能态转化的过程．尽管碱基配对是 RNA 折叠最基本的形式，但连续的碱基配对所形成的茎却可被看成是 RNA 二级结构形成的单元．从某种程度上说，RNA 的二级结构也可被看成是茎不断累积的结果．自由能越低并且配对的碱基之间间隔越小的茎越有可能首先进行折叠[11, 12]．

基于以上特点，RNA 的二级结构预测可被看成是一个茎的一个组合优化问题．这其中包含两个要素：a．如何衡量一个候选茎是否是 RNA 二级结构中“合适”的茎，即评价一个茎好坏的打分函

数；b．以何种搜索路径或者方式在候选茎池中搜索“合适”的茎，即启发式搜索策略．下面就分别加以说明．

2.1　能量函数

在 RNA 二级结构形成的过程中，如果一个茎的形成能使 RNA 的二级结构更稳定，则更有可能先形成，而衡量结构稳定的这个指标就是自由能．因此 StemFind 采用自由能作为评估和衡量候选茎的标准．于是合适的能量模型是决定预测准确度的一个关键因素．从理论上说，对于能量模型的选择，StemFind 不存在动态规划算法那样的限制，因而可以考虑更复杂准确的能量函数以及任意类型的假结．为了获得更好的预测效果，StemFind 采用了如下的能量模型．

首先，StemFind 的能量模型包括了著名动态规划软件(如 Mfold[3]，RNAfold[4])所采用的标准自由能模型中的全部要素．

$$G^{nest}=G^{hairpin}+G^{stem}+G^{interloop}+G^{bulge}+G^{multiloop} \quad (1)$$

在此基础上，StemFind 的能量模型还在两个方面进一步精确化：一个方面是采用了更准确的多分支环能量函数($G^{multiloop'}$)，另一方面是采用了完整的 Coaxial Stacking 能量($G^{coaxial}$)，而这两种更精确的能量函数都是很难被动态规划算法所使用的[3]．

$$G^{nest'}=G^{hairpin}+G^{interloop}+G^{bulge}+G^{multiloop'}+G^{coaxial} \quad (2)$$

为了更进一步地预测假结，StemFind 的能量模型也包括了复杂假结模型[5, 6]，该模型用多种不同的参数来从各个侧面描述了各种复杂的假结(图 1)．

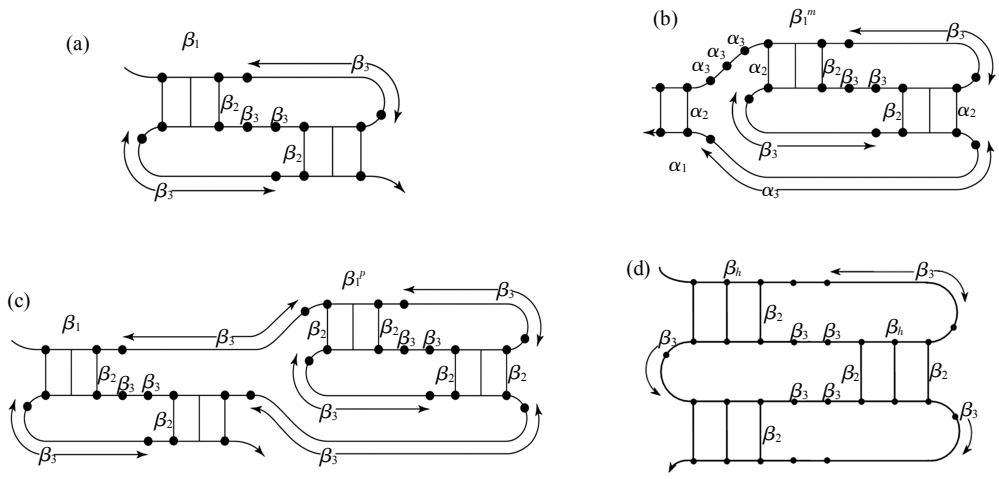

Fig. 1　Illustration of the pseudoknots

(a) A Simple pseudoknot (H-type pseudoknot). (b) A pseudoknot inside a multiloop. (c) A pseudoknot within a pseudoknot. (d) Overlapping pseudoknots. β_1 is the penalty of generating a simple pseudoknot, β_1^m is the penalty of the pair in a multiloop, β_1^p is the penalty of the pseudoknot within a pseudoknot. and β_2 is the penalty of paired base in a pseudoknot and β_3 is the penalty of non-paired base in a pseudoknot.

假结的能量由下面的公式得到：

$$G^{pseudo}=\beta_1+\beta_2 B^p+\beta_3 U^p+C_w \quad (3)$$

其中 β_1 是产生一个新假结的罚分；B^p 是假结内部边界上的碱基对个数；U^p 是假结内部未配对的碱基个数.如果这个新假结是多分枝环内部的假结，则 β_1 由 β_1^m 代替；如果这个新假结在另一个假结内部，则 β_1 由 β_1^p 所代替；如果这个新假结和另一个假结重叠，则 β_1 由 β_h 代替．C_w 是在三种不同类型的假结中的同轴堆积能(coaxial stacking)罚分．

最后，一个 RNA 二级结构的总能量可用下面的公式来表示：

$$G = G^{nest} + G^{pseudo} \quad (4)$$

2.2　启发式搜索算法

StemFind 启发式搜索算法的思路是在茎池中选择候选茎并且逐步“组装”成最后的 RNA 二级结构，并一定程度上体现了 RNA 折叠的两个特点[11, 12]：

1) RNA 的二级结构的形成可被看成候选茎的叠加过程，并且越能减少当前结构能量的候选茎越可能先形成．

2) RNA 的折叠过程也是一个从高能态向低能态转化的过程，在高能态时 RNA 的折叠具有多种

可能的状态，而在低能态时 RNA 的折叠逐步趋向一个贪心过程(类似模拟退火).

因而 StemFind 采用了如下的算法：StemFind 在每次选择候选茎时，先按降低当前结构能量的能力给当前茎池中的所有茎排名，能力越强的候选茎赋予更高的优先级(这符合特点 1)，同时在折叠的初始阶段(此时处于高能态)考虑更多的“次优茎”(即同时考虑多个能够降低当前结构能量的候选茎)，考虑的范围随着茎的不断加入而逐步减少(此时自由能也逐步降低，RNA 结构由高能态向低能态逐步转化，这符合特点 2)，最后阶段只选择最优茎，直至再没有茎可以减少当前结构的能量为止．此时一次迭代结束并得到一个由多个茎“组装”成的 RNA 二级结构．这个过程反复多次，得到多个 RNA 二级结构，最后我们通过打分函数来决定最优结构．下面先定义这个过程中需要用到的公式.

首先，我们定义候选茎选择概率来决定当前我们应该选择哪个候选茎：

$$P\{S=r\} = \Delta G(r) / \sum_{s \in D(t)} \Delta G(s), \quad r \in D(t), t = 1, 2, \cdots \quad (5)$$

其中 $P\{S=r\}$表示在当前状态选择茎 s_r 的概率；$\Delta G(r)$表示加入茎 s_r 后能使当前 RNA 结构的能量降低的数值；$D(t)$是在当前状态需要被考察的茎的有序集合(这个次序是按照茎减少当前结构能量的能力来排列，t 是 RNA 的折叠次数，即当前结构中加入的茎区数量)并且 $D(t)$的范围随着 t 的增大而逐渐缩小，最后变为 1，即最后变为一个贪心选择过程．以上的过程反复迭代进行 m 次，每次都会得到一个“组装”的 RNA 二级结构 $T_r(1 \leqslant r \leqslant m)$.

然后我们对这些二级结构采用下面的打分函数计算其分值：

$$Score\ (T_r) = G(T_r) * a/n + P(T_r) * b, \quad r \in [1, m] \quad (6)$$

其中 T_r 表示第 r 次迭代得到的 RNA 二级结构；$G(T_r)$表示 T_r 的自由能；$P(T_r)$表示 T_r 中的茎在所有的 RNA 二级结构中出现频率的总和；n 是 RNA 序列的长度；a 和 b 是常数．公式 6 表明在选择最终的二级结构时 StemFind 既考虑自由能最小的因素，同时也考虑茎的出现频率．换句话说，若一个茎在多个局部搜索结果中出现，我们则认为这个茎也很有可能也会出现在最终的 RNA 二级结构中(目前的参数设置为 a=5，b=1.5，m=50).

StemFind 算法的形式化定义如下.

StemFind Algorithm:

1: 找出序列所有可能存在的茎区并把它们放入初始茎池 $S_0 \leftarrow \{s_1, s_2, \cdots, s_n\}$
2: 初始化结构(不包含任何茎区) $T_0 \leftarrow \varphi$
3: **for** m from 0 to M **do** (局部搜索过程)
4: $S \leftarrow S_0$, $T \leftarrow T_0$, $find \leftarrow 1$ (S 代表当前茎池；T 代表当前结构；$find$ 是个标志位, 代表是否找到能够减少当前结构能量的茎)
5: **while** ($find$ =1) **do**
6: **for** each stem s_i in S **do**
7: **If** s_i 与 T 中的所有茎都相容 **do**
8: 把 s_i 加入 T: $T_i \leftarrow T \cup \{s_i\}$
9: 计算加入 s_i 后的结构 T_i 的自由能: $G_i \leftarrow G(T_i)$
10: 计算加入 s_i 后结构 T 减少的自由能:$\Delta G_i \leftarrow G_i - G(T)$
11: **end if**
12: **end for**
13: 根据公式 5 计算每个茎 s_i 出现的概率,并假设 s_k 是依据这个概率被选择的茎
14: **if** ($G_k < G(T)$) **then**
15: $S \leftarrow S - \{s_i\}$
16: $T \leftarrow T \cup \{s_i\}$
17: $find$ = 1
18: **else** $find$ = 0
19: **end while**
20: $T_m \leftarrow T$
21: **end for**
22: 依据公式 6, 输出具有最高分值的那个结构(T_{find})

3 实验结果

为了验证和比较 StemFind 算法预测 RNA 二级结构的有效性．我们与几个国际上比较著名的 RNA 二级结构预测软件 Mfold[3]，Pknots[5]，ILM[7] 和 HotKnots[13]进行了比较.

通常衡量一个算法预测 RNA 二级结构的准确性有两个指标 Sensitivity(称为查全率，简写为 SE) 和 Specificity(称为查准率，简写为 SP)．设 RP(real pair)是真实 RNA 二级结构中碱基对的数目，TP (true positive)是正确预测出的碱基对数目，FP(false positive) 是错误预测的碱基对数目．则 SE = TP/RP，SP = TP/(TP+FP).

为了更好地说明 StemFind 的局部搜索算法及其所采用的复杂能量模型的有效性．实验被分为两个部分进行．首先，在和 Mfold 等动态规划算法同

样的非假结能量模型下进行比较(对非假结的 RNA 数据)；然后，我们采用更为复杂和准确的能量模型再次和这些算法进行比较(包括非假结和假结 RNA 数据).

StemFind 的时间效率用 StemFind 在这些实验数据上的实际运行时间来衡量.

实验数据集来自 tRNAs, 5SrRNA, PseudoBase 以及其他常作为指标的包含假结的 RNA 序列(表 1). 其中对于 tRNA 和 5SrRNA 数据库我们从中随机挑选了 100 条序列进行测试.

Table 1 Testing sequences and features

Sequence	Sequence type	Length	BP	Reference
100 tRNAs	tRNA (pseudoknot-free)	70～82	18～20	[14]
100 5SrRNAs	5SrRNA (pseudoknot-free)	116～129	35～37	[15]
PseudoBase	Pseudoknot	23～137	7～44	[16]
BBMV1	Virus RNA	116	38	[5]
BBMV2	Virus RNA	114	39	[5]
Bt-PrP	mRNA	45	12	[13]
Biotin	miRNA	61	12	[5]
CCMV1	Virus RNA	134	46	[5]
CCMV34	Virus RNA	134	46	[5]
Ec_S15	mRNA	67	17	[13]
TMVup	Viral RNA	84	25	[5]
Tt-LSU-P3	mRNA	65	20	[13]

3.1 在非假结能量模型下的比较

为了测试 StemFind 所采用的启发式搜索算法的有效性，我们先让 StemFind 采用与 Mfold 完全一样的能量模型和参数(参见公式 1)，在此情况下进行比较. 测试集来自国际上著名的一个公用数据库(Sprinzl tRNA database)，我们从中随机抽取了 100 个 tRNA 标准数据.

Table 2 Comparison of the testing results of Mfold and StemFind on 100 tRNA sequences and 100 5SrRNA sequences

	tRNA		5SrRNA	
	Average sensitivity	Average specificity	Average sensitivity	Average specificity
Mfold	64	60	62	61
StemFind	68	67	62	60

StemFind adopts the same energy model and parameters that Mfold uses.

从表 2 可以看出，在同样的能量模型下，StemFind 的预测结果(查全率和查准率)并不低于 Mfold (甚至在 tRNAs 上要优于 Mfold). 这个结果说明了 StemFind 搜索策略的有效性(甚至要好于最小自由能策略).

3.2 在假结能量模型下的比较

为了衡量 StemFind 在预测包含假结的 RNA 二级结构的能力，我们把 StemFind 的能量模型扩充到包含假结的能量函数(见公式 3 和 4). 实验数据集有两个，一个是 10 条常被用来作为测试数据的 10 条包含假结的序列(长度从 45 到 134 不等)，在该实验集上的测试结果如表 3 所示.

Table 3 Detail result on 10 sequences with pseudoknots

No	RNA name	Length	HotKnots			ILM			PKNOTS			StemFind		
			SE	SP	K	SE	SP	K	SE	SP	K	SE	SP	K
1	BBMV1	116	68	68	0/2	79	81	0/2	74	72	0/2	89	77	2/2
2	BBMV2	114	79	82	0/2	82	82	0/2	77	77	0/2	97	86	2/2
3	Biotin	61	83	50	2/2	83	53	2/2	58	33	0/2	92	58	2/2
4	Bt-PrP	45	42	38	0/2	42	33	0/2	50	46	0/2	100	80	2/2
5	CCMV1	134	80	84	0/2	80	84	0/2	80	80	0/2	82	79	2/2
6	CCMV34	134	82	86	0/2	82	86	0/2	69	70	0/2	84	85	2/2
7	EMV	80	73	59	1/2	50	50	0/2	73	62	1/2	73	64	1/2
8	Ec_S15	67	100	74	2/2	59	62	0/2	100	74	2/2	100	74	2/2
9	TMVup	84	52	62	0/6	52	65	0/6	52	68	0/6	84	78	4/6
10	Tt-SU-P3	65	95	100	1/2	80	80	0/2	55	61	0/2	95	100	1/2
	Average		75.4	70.3		68.9	67.6		68.8	64.3		89.6	78.1	

Here “K”=(number of correctly predicted pseudoknots)/(expected number of pseudoknots).

从表 3 可以看出，StemFind 在这个测试集上的预测查全率和查准率(89.6，78.1)明显高于其他三种软件. 同时在对假结茎的预测结果上，StemFind 的性能也要明显优于其他三种软件.

为了进行更加客观和公平的比较，我们使用了第二个假结测试数据集. 它来自于 PseudoBase. 这是一个被广泛使用的假结数据库. PseudoBase 包含 16 种类型的 RNA(包括 rRNA, mRNA, tmRNA,

snRNA, snoRNA, hnRNA, Viral ribosomal frameshifting signals, Ribozymes, Aptamers, Artificail moleculars 等). 目前 PseudoBase 中共有 238 个假结序列，基本上近年来所有用于评测的假结测试数据都在这个数据库中[6, 7, 13, 17]. 在去除了一些不适合作为测试集的假结序列后(序列相似性大于 0.85 的冗余序列，非自然的 SELEX 假结序列以及假结的 LOOP1 和 LOOP3 区大于 800 的假结序列)我们的测试集中包含 168 个假结序列(我们称之为 PK168 数据集，这个测试集也被近年的另一个软件 PLMM 作为测试集[18]). 表 4 所示即为在 PK168 数据集上的测试结果.

从表 4 可以看出 StemFind 在查全率(0.78)上显著超过了其他三种软件，而 Pknots 在查准率(0.73)这个指标上最好. 同时我们可以从表 4 和表 5 看出在这两个假结测试集上 StemFind 和 Pknots 相比另外两种算法有更高的准确度.

Table 4 Summary of testing results on pk168 set

	Average SE (sensitivity) %	Average SP (specificity) %
PKNOTS	73	73
HotKnots	70	70
ILM	65	60
StemFind	78	70

3.3 运行效率

StemFind 在 Linux 平台(Red Hat FC6)用 C++语言开发，在主频为 2G Hz，Cache 为 2M 的 CPU 上进行，测试数据是用随机方式生成的 RNA 序列(长度范围从 50～400)，运行效率包括运行时耗费的时间(以秒为单位)和空间(以 MB 为单位). 实验结果见表 5.

Table 5 Detail results on sequences with different length

Length	HotKnots		ILM		PKNOTS		StemFind	
	Time(s)	Memory(MB)	Time(s)	Memory(MB)	Time(s)	Memory(MB)	Time(s)	Memory(MB)
50	0.08	1.1	0.01	0.5	8.4	45.2	0.5	0.5
100	8.1	1.9	0.03	0.7	745	146.5	3.4	0.6
200	79.1	47.8	0.08	1.3	120 563	1 045	36.4	1.1
300	207	78.5	3.4	2.1	*	*	124	1.4
400	697	107	8.5	3.0	*	*	432	2.9

从表 5 可以看出，StemFind 的速度要显著高于 Pknots 和 Hotknots. 同时 StemFind 占用的内存空间也是最小的.

4 讨 论

本文我们提出了一种基于自由能最小原则以及茎出现频率信息的局部搜索算法来预测带假结的 RNA 二级结构. 实验结果表明该算法对于假结和非假结的 RNA 序列都有很好的预测准确性.

从算法角度来看，StemFind 所采用的局部搜索算法只搜索了少量的茎组合方式，但却能够得到比其他方法更接近于真实结构的预测结果. 我们认为这可归因于 StemFind 的两个特点：其一是 StemFind 的局部搜索算法中采用的基于自由能最小和茎出现频率信息的启发式选择策略；其二是 StemFind 所采用的复杂能量模型. 启发式选择策略既在一定程度上体现了 RNA 折叠的特点，又显著地降低 StemFind 所需要搜索的空间. 这不仅可以提高 StemFind 的速度，还可以提高预测精度. 另一方面，StemFind 所采用的复杂能量模型不仅包含了非假结 RNA 二级结构的所有自由能细节而且还扩充到包含复杂假结的很多自由能细节. 因而这个能量模型可以从不同角度来刻画，表达影响一个 RNA 二级结构自由能的各个侧面. 尽管当前的能量函数仍然是不完善的，实验结果却表明 StemFind 所采用的启发式策略能从一定程度上弥补这个缺陷.

致谢 在此感谢 Mfold, PKNOTS, ILM 和 HotKnots 的作者提供的软件和测试集. 感谢中国科学院计算技术研究所的付岩博士，袁作飞博士，孙瑞祥博士和美国堪萨斯大学计算机系的浣军博士所提供的宝贵意见和帮助.

参 考 文 献

1 Walter A E, Turner D H, Kim J, *et al*. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proc Natl Acad Sci USA, 1994, **91**(20): 9218～9222

2 Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.

Bioinformatics, 1999, **15**(6): 446～454
3 Zuker M, Mathews D H, Turner D H. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide In RNA Biochemistry and Biotechnology. In: Barciszewski J, Clark B F C, eds. NATO ASI Series. Dordrecht, NL: Kluwer Academic Publishers, 1999. 11~43
4 Hofacker I L. Vienna RNA secondary structure server. Nucleic Acids Research, 2003, **31**(13): 3429～3431
5 Rivas E, Eddy S R. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol, 1999, **285**(5): 2053～2068
6 Dirks R M, Pierce N A. A partition function algorithm for nucleic acid secondary structure including pseudoknots. J Comput Chem, 2003, **24**(13): 1664～1677
7 Ruan J, Stormo G D, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics, 2004, **20**(1): 58～66
8 Mathews D H, Turner D H. Prediction of RNA secondary structure by free energy minimization. Curr Opin Struct Biol, 2006, **16**(3): 270～278
9 Gultyaev A P, van Batenburg F H, Pleij C W. The computer simulation of RNA folding pathways using a genetic algorithm. J Mol Biol, 1995, **250**(1): 37～51
10 Gultyaev A P. The computer simulation of RNA folding involving pseudoknot formation. Nucleic Acids Res, 1991, **19**(9): 2489～2494
11 Abrahams J P, van den Berg M, van Batenburg E, *et al*. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. Nucleic Acids Res, 1990, **18**(10): 3035～3044
12 Saenger W. Principles of Nucleic Acid Structure. New York: Springer-Verlag, 1984. 126～131
13 Ren J, Rastegari B, Condon A, *et al*. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. RNA, 2005, **11**(10): 1494～1504
14 Sprinzl M, Horn C, Brown M, *et al*. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res, 1998, **26**(1): 148～153
15 Szymanski M, Barciszewska M Z, Erdmann V A, *et al*. 5S ribosomal RNA database. Nucleic Acids Res, 2002, **30**(1): 176～178
16 van Batenburg F H, Gultyaev A P, Pleij C W, *et al*. PseudoBase: a database with RNA pseudoknots. Nucleic Acids Res, 2000, **28**(1): 201～204
17 Eddy S R, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Res, 1994, **22**(11): 2079～2088
18 Huang X, Ali H. High sensitivity RNA pseudoknot prediction. Nucleic Acids Res, 2007, **35**(2): 656～663

A Local-stem-search Algorithm to Predict The RNA Secondary Structure*

CHEN Xiang[1,2,4], BU Dong-Bo[1,2], ZHANG Fa[1,2,3], GAO Wen[3]**

([1] *Key Laboratory of Intelligent Information Processing, The Chinese Academy of Sciences, Beijing* 100190, China;
[2] *Institute of Computing Technology, The Chinese Academy of Sciences, Beijing* 100190, China;
[3] *School of Electronic Engineering and Computer Science, Peking University, Beijing* 100080, China;
[4] *Graduate University of The Chinese Academy of Sciences, Beijing* 100049, China)

Abstract RNA secondary structure predicting is a classical problem in bioinformatics and the optimal algorithms based on minimal free energy (MFE) criterion are the widely used methods. However, pseudoknots render the problem of computing the RNA MFE structure with pseudoknot becomes a NP-hard problem. A heuristic algorithm——StemFind to predict RNA secondary structure with pseudoknot was presented. The algorithm regard stem as the basic search unit, adopting heuristic search strategy, and search the most possible RNA secondary structure in stem combination space. The StemFind algorithm to a large number of test sets was applied. Performance evaluation demonstrates that StemFind not only outperforms the well-known optimal and heuristic algorithms in overall sensitivity and specificity but also requires significantly less time than the optimal algorithm.

Key words RNA secondary structure prediction, pseudoknot，NP-hard, heuristic algorithm

---

*This work was supported by grants from National Basic Research Program of China (2002CB713807), Frontier Project of Knowledge Innovation Program of The Chinese Academy of Sciences(20076020) and The National Natural Science Foundation of China(60503060, 90612019, 60752001).
**Corresponding author. Tel: 86-10-62758602, E-mail: wgao@pku.edu.cn
Received: May 6, 2008 Accepted: June 26, 2008