生物工程学报
journals.im.ac.cn
cjb@im.ac.cn

Chin J Biotech 2008, April 25; 24(4): 659-664
Chinese Journal of Biotechnology ISSN 1000-3061

研究报告

基于堆积协变信息与最小自由能预测含伪结的 RNA 二级结构

杨金伟, 骆志刚, 方小永, 王金华, 唐可成

国防科学技术大学计算机学院, 湖南 410073

摘　要: RNA 伪结预测是 RNA 研究的一个难点问题。文中提出一种基于堆积协变信息与最小自由能的 RNA 伪结预测方法。该方法使用已知结构的 RNA 比对序列(ClustalW 比对和结构比对)测试此方法, 侧重考虑相邻碱基对之间相互作用形成的堆积协变信息, 并结合最小自由能方法对碱基配对综合评分, 通过逐步迭代求得含伪结的 RNA 二级结构。结果表明, 此方法能正确预测伪结, 其平均敏感性和特异性优于参考算法, 并且结构比对的预测性能比 ClustalW 比对的预测性能更加稳定。文中同时讨论了不同协变信息权重因子对预测性能的影响, 发现权重因子比值在λ1: λ2=5:1 时, 预测性能达到最优。

关键词: RNA 二级结构、伪结、堆积协变信息、最小自由能

Predicting RNA Secondary Structures Including Pseudoknots by Covariance with Stacking and Minimum Free Energy

Jinwei Yang, Zhigang Luo, Xiaoyong Fang, Jinhua Wang, and Kecheng Tang

College of Computer Science, National University of Defense Technology, Changsha 410073, *China*

Abstract: Prediction of RNA secondary structures including pseudoknots is a difficult topic in RNA field. Current predicting methods usually have relatively low accuracy and high complexity. Considering that the stacking of adjacent base pairs is a common feature of RNA secondary structure, here we present a method for predicting pseudoknots based on covariance with stacking and minimum free energy. A new score scheme, which combined stacked covariance with free energy, was used to assess the evaluation of base pair in our method. Based on this score scheme, we utilized an iterative procedure to compute the optimized RNA secondary structure with minimum score approximately. In each interaction, helix of high covariance and low free energy was selected until the sequences didn't form helix, so two crossing helixes which were selected from different iterations could form a pseudoknot. We test our method on data sets of ClustalW alignments and structural alignments downloaded from RNA databases. Experimental results show that our method can correctly predict the major portion of pseudoknots. Our method has both higher average sensitivity and specificity than the reference algorithms, and performs much better for structural alignments than for ClustalW alignments. Finally, we discuss the influence on the performance by the factor of covariance weight, and conclude that the best performance is achieved when $\lambda_1 : \lambda_2 = 5 : 1$.

Keywords: RNA secondary structure, pseudoknots, covariance with stacking, minimum free energy

Received: August 15, 2007; **Accepted:** November 5, 2007
Supported by: the Natural Science Foundation of China (No. 60673018) and Natural Science Foundation of HuNan (No. 06JJ4123).
Corresponding author: Zhigang Luo. Tel: 13807311725, E-mail: zgluo@nudt.edu.cn
国家自然科学基金(No. 60673018)和湖南省自然科学基金(No. 06JJ4123)资助。

非编码 RNA 对染色体复制、基因转录调节、蛋白质的合成和运输等生命活动过程都有着重要的作用，其功能和结构是密切相关的。直接利用实验方法获得 RNA 结构相对比较困难，从而制约了非编码 RNA 的研究和发展。研究与开发准确、快速的 RNA 二级结构预测算法是克服这一困难的手段之一。

通常，RNA 二级结构基本单元包括茎环(Stem)、发夹环(Hairpin loop)、内部环(Interior loop)、突起环(Bulge loop)、多分枝环(Multiple loop)。伪结(Pseudoknot)则是通过二级结构单元之间的相互作用形成的一种结构，它是 RNA 研究中最难预测的一种结构。目前，针对 RNA 伪结的预测方法主要有 3 种：基于热力学的动态规划算法、基于比较 RNA 组学方法以及启发式算法。第一类基于热力学的动态规划算法主要是找出 RNA 序列具有最小自由能的折叠，代表性算法有 pknotsRE[1]和 pknotsRG-mfe[2]，处理的伪结类型比较丰富，但是过高的时间复杂度 $O(n^6)$和空间复杂度 $O(n^4)$限制了此类方法的应用；第二类基于比较 RNA 组学的方法，通过计算多条同源 RNA 序列的碱基协变信息，推断其公共的二级结构，代表性算法有 Construct[3]和 Ifold[4]，能够获得比较准确的结构，但是由于需要手工干预，不适合规模化分析，并且其预测精度也需进一步提高；第三类启发式算法，采用添加子结构的方式逐步得到整条 RNA 序列的二级结构，代表性算法有 ILM[5]和 HotKnots[6]，不依赖于特定的数学模型，计算速度比较快，但是预测精度不高。

针对 RNA 伪结预测所存在的问题，本研究在 Ifold[4]的基础上，改进了协变信息计算模型，提出了一种基于堆积协变信息和最小自由能的 RNA 伪结预测方法。数值实验表明，引入堆积协变信息能够有效地提高 RNA 二级结构的预测精度。

1 方法

1.1 堆积协变信息计算模型

通常，协变信息用来度量多条同源序列不同位点上的相关性。常用的一种协变信息是互信息量，但是由于互信息量没有考虑 RNA 碱基配对原则和一致非补偿性突变，Hofacker[7]等人提出了协变信息得分公式：

$$C_{ij} = \sum_{XY,X'Y'} f_{ij}(XY) D_{XY,X'Y'} f_{ij}(X'Y')$$

其中 C_{ij} 表示 N 条同源 RNA 序列 i 和 j 两个位点的协变信息得分，$f_{ij}(XY)$表示 i 和 j 两个位点出现碱基对 XY 的概率，$D_{XY,X'Y'}$表示碱基对 XY 与 $X'Y'$的海明距离矩阵，定义如下：

$$D_{XY,X'Y'} = \begin{cases} 0 \text{ 当 } XY \notin P \text{ 或 } X'Y' \notin P \text{ 或 } XY = X'Y' \\ 1 \text{ 当 } XY, X'Y' \notin P \text{ 并且 } X \text{ 与 } X'、Y \text{ 与 } Y' \\ \quad \text{仅有一者不等} \\ 2 \text{ 当 } XY, X'Y' \notin P \text{ 并且 } X \neq X'、Y \neq Y' \end{cases}$$

其中 P 为碱基配对集合，P={A·U, U·A, C·G, G·C, G·U, U·G}。

对于非一致序列(i 和 j 两个位点不能形成碱基配对或有位点为空位)给予罚分：

$$q_{ij} = 1 - \frac{1}{N}\sum_{\alpha}\left\{ \Pi_{ij}^{\alpha} + \delta(a_i^{\alpha}, \text{gap})\delta(a_j^{\alpha}, \text{gap}) \right\}$$

其中α为 N 条同源 RNA 序列中的任意一条序列。当序列α的 i 和 j 两个位点形成碱基配对 $i \cdot j$ 时，$\Pi_{ij}^{\alpha} = 1$；否则，$\Pi_{ij}^{\alpha} = 0$。当序列α位点 i 的值 $a_i^{\alpha} = \text{gap}$ 时，$\delta(a_i^{\alpha}, \text{gap}) = 1$；否则，$\delta(a_i^{\alpha}, \text{gap}) = 0$。由一致序列的协变信息得分和非一致序列的罚分组合成协变信息评估公式：

$$B_{ij} = C_{ij} - \varphi_1 q_{ij}$$

其中系数 φ_1 调节非一致序列罚分的权重。

在上述协变信息计算模型中，Hofacker[7]等人只是根据多条同源序列的 2 个位点上的碱基计算协变信息，而 RNA 二级结构有一个共同特征是相邻的碱基对之间可以相互作用，形成碱基对堆积[8]，并且在热力学上表现为最近邻能量规则。因此，我们在协变信息中也引入碱基对堆积，以提高协变信息评判的可靠性。以碱基对(i, j)为例，如果 B_{ij} 很大，也就意味着相邻的碱基对$(i-1, j+1)$和$(i+1, j-1)$也应该有较高的协变信息得分。我们定义堆积协变信息计算公式为：

$$B_{ij}^{S} = \frac{\lambda_1 B_{i-1,j+1} + \lambda_2 B_{ij} + \lambda_3 B_{i+1,j-1}}{\lambda_1 + \lambda_2 + \lambda_3}$$

其中，B_{ij}^{S} 为 N 条同源 RNA 序列 i 和 j 两个位点的堆积协变信息评估值，$B_{i-1,j+1}$、B_{ij} 和 $B_{i+1,j-1}$ 分别为碱基对$(i-1, j+1)$、(i, j)和$(i+1, j-1)$的协变信息评估值，λ_1、λ_2 和 λ_3 分别为三个评估值的权重因子。

通常情况下，我们认为碱基对$(i-1, j+1)$和$(i+1, j-1)$对(i, j)的影响是相同的，所以我们取$\lambda_1 = \lambda_3$。故

上式化简为:

$$B_{ij}^{S}=\frac{\lambda_1 B_{i-1,j+1}+\lambda_2 B_{ij}+\lambda_1 B_{i+1,j-1}}{2\lambda_1+\lambda_2}$$

权重因子λ_1和λ_2的大小体现了相邻碱基对对参考位置碱基对的影响。随着λ_1: λ_2的增大，相邻碱基对对参考位置碱基对的影响越大，反之，影响越小。当λ_1=0, λ_2=1 时，相邻碱基对(i−1, j+1)和(i+1, j−1)对参考位置碱基对(i, j)没有影响，堆积协变信息计算公式就成为上面提到的协变信息评估公式。

为了根据堆积协变信息判断两个位点是否形成碱基配对，我们设定一个阈值 B^*。当堆积协变信息 B_{ij}^{S} 大于B^*时，同源序列i和j两个位点形成配对；否则，不形成配对，即:

$$\Pi_{ij}=\begin{cases}0 & \text{当} B_{ij}^{S}<B^{*}\text{时}\\ 1 & \text{当} B_{ij}^{S}\geqslant B^{*}\text{时}\end{cases}。$$

由随机序列测定[7]，随机序列的 C_{ij} 和 q_{ij} 期望值分别为 0.203、0.833，则设定阈值为:

$$B^{*}=0.203-\varphi_1\times 0.833$$

1.2 基于堆积协变信息和最小自由能的动态规划算法

我们同时考虑堆积协变信息和最小自由能，对同源序列的公共折叠进行打分，使用动态规划算法寻找多条同源序列的不含伪结结构的公共最优折叠。

为了同时考虑堆积协变信息和最小自由能，我们结合堆积协变信息 B_{ij}^{S} 和自由能 $e(a_i^{\alpha},a_j^{\alpha})$ 形成碱基对的综合得分:

$$\beta_{ij}=\frac{1}{N}\sum_{\alpha}e(a_i^{\alpha},a_j^{\alpha})-\varphi_2 B_{ij}^{S}$$

其中，$e(a_i^{\alpha},a_j^{\alpha})$ 表示序列α的位点 i 和 j 形成碱基对时的自由能，$\frac{1}{N}\sum_{\alpha}e(a_i^{\alpha},a_j^{\alpha})$ 则表示 N 条同源序列在位点 i 和 j 形成碱基对时的平均自由能，φ_2 调节堆积协变信息的权重。设 E_{ij} 为 N 条同源序列位点 i 到 j 具有最优折叠的综合得分，那么 E_{ij} 满足下列迭代公式:

$$E_{ij}=\min\left\{E_{i,j-1};\ \min_{\substack{k:i+m<k\leqslant j\\ \Pi_{ik}=1}}(E_{i+1,k-1}+E_{k+1,j}+\beta_{ik})\right\}$$

其迭代和回溯过程与 Mfold[9]相同，在此不再详细阐述。$E_{1,n}$ 为 N 条长度为 n 的同源序列形成最优折叠时的综合得分。

1.3 算法流程

基于堆积协变信息和最小自由能预测同源 RNA 含伪结的公共二级结构算法流程如下:

(1) 计算 N 条同源 RNA 序列的堆积协变信息矩阵 $B^{S}=\left(B_{ij}^{S}\right)_{n\times n}$;

(2) 由基于堆积协变信息和最小自由能的动态规划算法计算出同源 RNA 序列的公共结构(不含伪结)，从中得出所有碱基配对的螺旋区列表 *HelixList*;

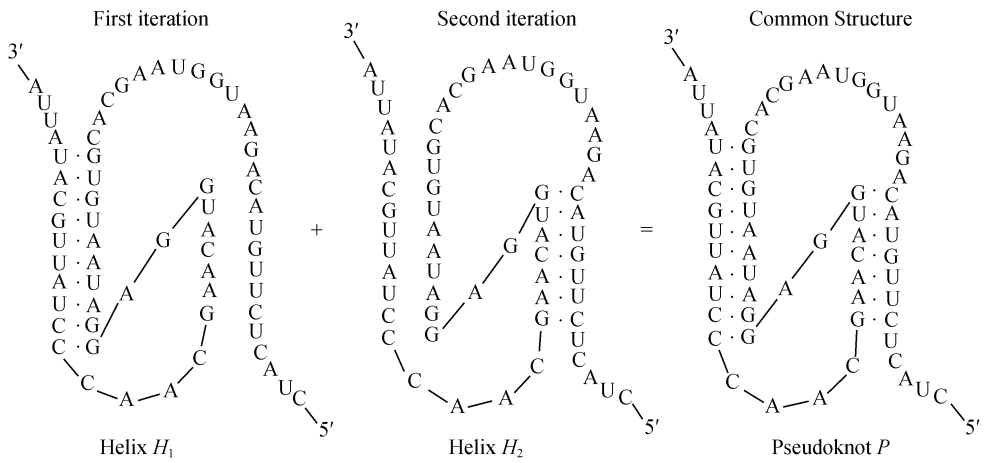

图 1 相互交叉的螺旋区形成伪结

Fig. 1 Two cross-cutting helixes form a pseudoknot

Take predicting common structure of Corona virus family as an example. In the first iteration, helix H_1 get the best combined score and is added to the list of common structure. In the second iteration, the bases belonging helix H_1 (the gray part)are removed from the sequences. Another helix H_2 that cross with H_1 get the best combined score and is also added to the list, whereupon H_1 and H_2 form a pseudoknot

(3) 合并螺旋区列表 *HelixList* 中被小内环或突起环隔离的螺旋区，形成更大的螺旋区。如果没有发现螺旋区，转到(6);

(4) 计算螺旋区列表 *HelixList* 中每个螺旋区的综合得分，选择具有最低综合得分的螺旋区，保存到公共结构列表 *ComStructureList* 中;

(5) 从初始序列中删除公共结构列表 *ComStructureList* 中的碱基，更新相应的得分矩阵，清空螺旋区列表 *HelixList*，转到(2);

(6) 整理公共结构列表 *ComStructureList*，删除 *ComStructureList* 中非标准的碱基配对，并根据 *ComStructureList* 输出碱基配对列表。

在此算法中，每次迭代时选择综合得分最低的螺旋区，而最终的螺旋区列表中可能出现相互交叉的螺旋区，交叉的螺旋区能形成伪结(如图 1 所示)，故此算法可以预测伪结。

2 实验结果

我们将上述算法实现为程序 CSfold，并做了大量的数值实验测试。一方面预测多个家族同源序列的公共二级结构，并和一些典型算法(Pknots、ILM、Ifold)进行预测精度比较；另一方面分析不同的同源序列比对质量和堆积协变信息权重因子对预测精度的影响。

我们从文献[10]中选取了 10 种不同 RNA 家族比对序列作为测试数据集，包括 Coronavirus、Enterovirus、Tombusvirus、α-operon mRNA 和 HDV 家族的 ClustalW 比对序列以及 5S rRNA、Telomerase RNA、SRP RNA、RNase P RNA 和 tmRNA 家族的结构比对序列。

我们使用同一家族的多条同源序列比对作为 CSfold 的输入，输出是同源序列的公共预测结构，然后再去掉参考序列在比对时产生的空位，结合公共预测结构，得出参考序列的预测结构。最后对比参考序列的预测结构和参考结构评估预测精度。预测精度使用敏感度和特异性两个指标评估。设 *RP* 为参考结构的碱基配对数，*TP* 为预测正确的碱基配对数(真阳性)，*FP* 为参考结构未包含的预测碱基配对数(假阳性)，则敏感度定义为 $SS=100\times TP/RP$，特异性定义为 $SP=100\times TP/(TP+FP)$[11]。而伪结的预测正确性使用 *PK* 评估，设 *TK* 为正确预测的伪结个数，*RK* 为参考结构中的伪结个数，则 $PK=TK/RK$。

首先我们使用 ClustalW 比对序列(包括 Coronavirus、Enterovirus、Tombusvirus、α-operon mRNA 和 HDV)测试 CSfold，其堆积协变信息的权重因子取值为 $\lambda_1=5$，$\lambda_2=1$，其结果与 Pknots、ILM、Ifold 三种算法的结果进行比较，性能结果如表 1 所示。结果表明，大多数的测试结果要优于三种算法，而且 CSfold 的平均敏感性略优于其他三种算法，平均特异性达到了最优，即预测错误的碱基配对比较少。在伪结预测方面，除少预测了α-operon mRNA 的一个伪结之外，CSfold 正确预测了其他伪结结构，伪结预测能力和其他三种算法相当。

然后我们使用比对质量更高的结构比对序列(包括 Telomerase RNA、SRP RNA、RNase P RNA、5S rRNA 和 tmRNA)测试 CSfold，其堆积协变信息的权重因子取值为 $\lambda_1=5$，$\lambda_2=1$，其结果与 ILM、Ifold 两种算法的结果进行比较(因为 Pknots 所处理的序列长度限制在 150 bp 之内，故其不能预测此类较长序列)，性能结果如表 2 所示。结果表明，CSfold 的平均敏感性和特异性分别为 80%和 94%，优于 ILM、Ifold 这两种算法。在伪结预测方面，除 RNase P RNA 的一个伪结没有被正确预测之外，其他序列的伪结结构都被正确预测，预测结果略优于 ILM、Ifold。

另外，比较表 1 和表 2 中的 CSfold 的结果，我们发现由于结构比对序列的比对质量比 ClustalW 比对序列更高，计算得到的堆积协变信息更加可靠，CSfold 预测结果的敏感性和特异性的均方差都有所降低，说明 CSfold 使用结构比对序列预测 RNA 二级结构性能更加稳定。

最后，我们分析了不同的堆积协变信息权重因子对预测性能的影响。我们采用不同的权重因子 λ_1 和 λ_2，使用 CSfold 算法预测了 4 种 RNA 家族比对序列(Tombusvirus、Enterovirus、SRP RNA 和 Telomerase RNA)的二级结构，其结果如表 3 所示。结果表明，权重因子比值为 $\lambda_1:\lambda_2=5:1$ 时，预测性能达到最优；当比值小于此值时，预测性能小于最优值；当比值大于此值并逐渐增大时，预测性能趋于稳定，但是结果不是最优的。这说明相邻碱基对之间的协变信息是相互影响的，从而验证了我们使用堆积协变信息计算模型的正确性。

表 1 使用 **ClustalW** 比对序列比较 **CSfold** 与 **Pknots**、**ILM**、**Ifold** 的性能

Table 1 Performance comparison of Pknots, ILM and Ifold using ClustalW alignments

	Pknots			ILM			Ifold			CSfold		
	SS	*SP*	*PK*	*SS*	*SP*	*PK*	*SS*	*SP*	*PK*	*SS*	*SP*	*PK*
Coronavirus	83	79	1/1	94	77	1/1	94	100	1/1	94	100	1/1
Enterovirus	76	97	0/1	68	93	1/1	68	90	1/1	87	99	1/1
Tombusvirus	79	70	1/1	58	58	1/1	52	48	1/1	92	100	1/1
α-operon mRNA	77	40	1/2	50	31	1/2	68	58	1/2	82	64	1/2
HDV	85	77	1/1	59	55	1/1	63	74	1/1	63	81	1/1
Average	80	73	—	66	63	—	69	74	—	84	89	—
Standard Deviation	4	21	—	17	23	—	15	21	—	12	16	—

Note: SS=100×TP/RP; SP=100×TP/(TP+FP); RP=number of base pairs in the reference structure; TP=number of true positive predicted base pairs; FP=number of false positive predicted base pairs; PK= (number of correctly predicted pseudoknots)/(number of pseudoknots in the reference structure).

表 2 使用结构比对序列比较 **CSfold** 与 **ILM**、**Ifold** 的性能

Table 2 Performance comparison of ILM and Ifold using structural alignments

	ILM			Ifold			CSfold		
	SS	*SP*	*PK*	*SS*	*SP*	*PK*	*SS*	*SP*	*PK*
5S rRNA	83	100	0/0	80	100	0/0	80	100	0/0
SRP RNA	86	67	1/1	91	83	1/1	92	91	1/1
RNase P RNA	76	76	1/2	73	82	1/2	70	92	1/2
Telomerase RNA	57	39	0/1	61	51	0/1	77	92	1/1
tmRNA	90	72	4/4	74	80	3/4	83	95	4/4
Average	78	71	—	76	79	—	80	94	—
Standard Deviation	13	22	—	11	18	—	8	4	—

Note: *SS*、*SP* and *PK* are defined in Table 1.

表 3 不同堆积协变信息权重因子对预测性能的影响

Table 3 The performance influenced by different factor of covariance with stacking

	Tombusvirus			Enterovirus			SRP RNA			Telomerase RNA		
	SS	*SP*	*PK*	*SS*	*SP*	*PK*	*SS*	*SP*	*PK*	*SS*	*SP*	*PK*
λ_1: λ_2=1:1	21	36	0/1	84	90	1/1	86	93	1/1	51	80	1/1
λ_1: λ_2=2:1	21	36	0/1	84	90	1/1	79	82	1/1	57	83	1/1
λ_1: λ_2=3:1	75	86	1/1	84	90	1/1	84	85	1/1	74	88	1/1
λ_1: λ_2=4:1	58	70	1/1	87	99	1/1	90	87	1/1	73	87	1/1
λ_1: λ_2=5:1	91	100	1/1	87	99	1/1	92	91	1/1	77	92	1/1
λ_1: λ_2=6:1	56	51	1/1	87	99	1/1	87	87	1/1	77	92	1/1
λ_1: λ_2=8:1	56	51	1/1	87	99	1/1	86	87	1/1	77	92	1/1
λ_1: λ_2=10:1	56	47	1/1	82	98	1/1	86	86	1/1	77	92	1/1
λ_1: λ_2=12:1	54	45	1/1	82	98	1/1	86	86	1/1	77	89	1/1

Note: *SS*、*SP* and *PK* are defined in Table 1.

3 结论

在本文中，我们论述了一种基于堆积协变信息与最小自由能的 RNA 伪结预测方法。我们采用了更加可靠的堆积协变信息计算模型，并且结合最小自由能来预测 RNA 的二级结构。数值实验结果表明，此方法的平均敏感性和特异性优于其他参考算法。我们还讨论了不同的协变信息权重因子对预测性能的影响，发现权重因子比值在λ_1: λ_2=5:1 时，预测性能达到最优。为了降低问题的复杂性，我们把堆积

协变信息的计算简化为线性模型，而碱基对之间的协变信息影响可能更加复杂，在此方面，有待进一步研究。

REFERENCES

[1] Rivas E, Eddy S. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol*, 1999, **285**: 2053–2068.

[2] Reeder J, Giegerich R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 2004, **5**: 104.

[3] Luck R, Graf S, Steger G. ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Research*, 1999, **27**(21): 4208–4217.

[4] Wang JH, Luo ZG, Guan NY, *et al*. An iterative method for prediction of RNA secondary structures including pseudoknots based on minimum of free energy and covariance. *Hereditas*, 2007, **29**(7): 889–897.
王金华，骆志刚，管乃洋，等. 基于最小自由能和协变信息预测带伪结 RNA 二级结构的迭代化方法. 遗传, 2007, **29**(7): 889–897.

[5] Ruan J, Stormo GD, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 2004, **20**: 58–66.

[6] Ren J, Rastegari B, Condon A, *et al*. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 2005, **11**(10): 1494–1504.

[7] Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction of aligned RNA sequences. *J Mol Biol*, 2002, **319**: 1059–1066.

[8] Lindgreen S, Gardnerand PP, Krogh A. Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics*, 2006, **22**(24): 2988–2995.

[9] Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 1981, **9**(1): 133–148.

[10] Witwer C, Hofacker IL, Stadler PF. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 2004, **1**(2): 66–77.

[11] Baldi P, Brunak S, Chauvin Y, Andersen C, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 2000, **16**: 412–424.