

DOI:10.3969/j.issn.1672-5565.201701006

多目标遗传算法的含假结 RNA 二级结构预测

顾侗¹, 蔡磊鑫¹, 王帅¹, 吕强^{2*}

(1. 苏州大学 计算机科学与技术学院, 苏州 215006;

2. 江苏省计算机信息处理技术重点实验室(苏州大学), 苏州 215006)

摘要:假结是 RNA 中一种重要的结构, 由于建模的困难导致它更难被预测。通过碱基之间的配对概率来预测含假结 RNA 二级结构的 ProbKnot 算法具有很高的精度, 但该算法仅用了配对概率作为预测依据, 导致阴性配对大量出现, 因此精度中的特异性较低。实验结合 ProbKnot 算法中碱基配对概率模型, 通过使用多目标遗传算法, 从而提高预测含假结 RNA 二级结构的特异性, 以此促进总体精度的提高。实验过程中, 首先计算出每个碱基成为单链的概率, 作为新增的预测依据, 然后使用遗传算法对 RNA 二级结构进行交叉、变异和迭代, 最后得到 Pareto 最优解, 进一步得出最高的最大期望精度。实验结果表明, 在使用的 RNA 案例中, 采用该方法比现有方法精度平均提高约 4%。

关键词:RNA 二级结构; 假结; 多目标优化; 遗传算法; 最大期望精度

中图分类号:TP391 **文献标志码:**A **文章编号:**1672-5565(2017)03-142-07

Prediction of RNA secondary structure including pseudoknots based on multi-objective genetic algorithm

GU Ti¹, CAI Leixin¹, WANG Shuai¹, LÜ Qiang^{2*}

(1. School of Computer Science & Technology, Soochow University, Suzhou 215006, China;

2. Provincial Key Laboratory for Computer Information Processing Technology of Jiangsu (Soochow University), Suzhou 215006, China)

Abstract: Pseudoknot is a kind of important structure in RNA, which is more difficult to be predicted due to the difficulty in training the prediction model. ProbKnot algorithm has a high accuracy in predicting the secondary structure of pseudoknotted RNA based on base-pair probabilities. However, this algorithm only makes use of the base-pair probabilities as a predictive feature, which leads to a large number of negative pairs and lower specificity. The overall accuracy can be improved which follows the improvement of specificity by combining the probabilistic model of base pairing in ProbKnot algorithm and multi-objective genetic algorithm. In the experiments, we will first calculate the single-stranded probability as a new predictive feature, and then use genetic algorithm to cross, mutate and iterate the secondary structure of RNA. Finally, we can get the Pareto optimal solutions and the highest maximum expected accuracy. The experiment results showed that in the RNA cases, applying this method could achieve an average increase of 4% in the accuracy compared with the current existing methods.

Keywords: RNA secondary structure; Pseudoknot; Multi-objective optimization; Genetic algorithm; Maximum expected accuracy

在最初的生物中心法则中, RNA 被认为在表达遗传信息时作为蛋白质翻译, 发挥了短暂的作用。后来发现 RNA 除了翻译蛋白质, 还具有多种其他功能, 如调控基因表达, 转运 RNA, 催化肽链形成和指

导蛋白质合成等。随着研究的深入, RNA 的形象已经由功能单一的简单线性碱基序列演变成种类多样、结构复杂、功能特异的生命核心物, 同时它在中心法则中取得了与 DNA 和蛋白质同样重要的地位。

收稿日期: 2017-01-20; 修回日期: 2017-04-11.

基金项目: 国家自然科学基金(61170125).

作者简介: 顾侗, 男, 硕士研究生, 研究方向: 生物信息计算; E-mail: 20145227032@stu.suda.edu.cn.

* 通信作者: 吕强, 男, 教授, 研究方向: 生物信息计算、元启发搜索、并行计算; E-mail: qiang@suda.edu.cn.

许多 RNA 序列具有准确定义的结构,想要理解这些 RNA 序列如何实现它们的功能,知道它们的结构是非常重要的^[1]。

RNA 中最基本的 4 种碱基为 A(腺嘌呤)、U(尿嘧啶)、G(鸟嘌呤)、C(胞嘧啶)。RNA 一级结构是碱基的有序序列。RNA 二级结构是由碱基之间相互作用形成的 3 种规范碱基对 AU, GC 和 GU 以及未配对碱基的单链组成的结构^[2]。RNA 三级结构是原子的三维排列。因为 RNA 的结构一般是分层次的,所以二级结构可以在不用知道三级结构的情况下得到确定,同样也可以为预测三级结构提供很大的帮助。RNA 二级结构承担着由 RNA 一级结构推测其三级结构的桥梁作用。研究 RNA 二级结构,可以为提高 RNA 及蛋白质结构预测的准确率提供帮助。因此, RNA 结构研究的一个重要切入点是对其二级结构的研究。

鉴于二级结构的重要性,对二级结构的预测研究显得尤为重要。预测 RNA 二级结构的方法主要有序列比对法和最小自由能法。其中,序列比对法需要耗费大量人力,预测效率较低,因而最小自由能法是预测 RNA 二级结构时广泛采用的一种方法。Zuker^[3]提出的 Mfold 算法是早期基于最小自由能方法的二级结构预测算法,该算法有很高的正确率,但不能预测含假结的复杂结构。二级结构预测中,假结作为一种特殊的结构,是实现结构功能的重要因素。但是因为碱基相互重叠的特性,生物信息计算对假结的预测难以奏效。预测包含假结的 RNA 二级结构问题是 NP 难的^[4],但要分析 RNA 的真实结构,对假结的准确预测是必须解决的问题。

通过加入假结约束来预测假结的最小自由能算法是目前使用比较多的预测方法,其中 Probknot 算法^[5-6]在几种 RNA 含假结二级结构预测算法^[7-10]中精准确度较高。Probknot 算法的配分函数(partition function)^[11]利用机器学习法和热力学模型^[12]计算出每个碱基之间的配对概率,各个碱基通过配对概率相互配对,然后解除形成的结构中连续配对碱基数较少的茎区,以确保稳定性^[13],最终得出预测结果。本文运用多目标遗传算法,通过提高特异性从而提高总体精度的方式来改进 Probknot 算法,并与 Probknot 算法以及其他算法的结果进行对比和总结。

1 打分函数与预测性能指标

1.1 打分函数 MaxExpect

相比最小自由方法的最小自由能结构,

MaxExpect 利用 RNA 结构中各个碱基配对情况打分得到的最大期望精度结构比最小自由能结构具有更高的精度^[14]。

MaxExpect 的打分公式如下

$$\text{Expected accuracy}(S) = \sum_{i,j \in BP} \gamma \cdot 2P_{bp}(i,j) + \sum_{k \in SS} P_{ss}(k) \quad (1)$$

式中: γ 取 1; $P_{bp}(ij)$ 为碱基 i 与 j 相互配对的概率; $P_{ss}(k)$ 为碱基 k 是单链的概率。ProbKnot 算法中 $P_{bp}(i,j)$ 是已知的,本文通过 $P_{bp}(i,j)$ 计算出 $P_{ss}(k)$ 为

$$P_{ss}(i) = 1 - \sum_{j=1}^N P_{ij} \quad (2)$$

本文将以 $\sum_{i,j \in BP} P_{bp}(i,j)$ 和 $\sum_{k \in SS} P_{ss}(k)$ 这两个函数为依据进行多目标优化,以得出最高的最大期望精度。

1.2 预测性能指标

预测性能指标是用来评价 RNA 二级结构预测结果好坏的标准。

马休兹相互作用系数(Mathews correlation coefficient, MCC)是通过比对天然结构与预测结构而计算出来的值,最小值为-1,表示预测结构与天然结构碱基配对相似度为 0;最大值为 1,表示预测结构与天然结构相似度为 1。具体衡量公式如下:

$$X = \frac{TP}{TP + FN} \quad (3)$$

$$Y = \frac{TP}{TP + FP} \quad (4)$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

式中: TP(true positive) 为正确预测碱基对的个数; FN(false negative) 为真实结构中不存在但没有被正确预测出的碱基对个数; FP(false positive) 为真实结构中不存在却被错误预测到的碱基对个数; TN(true negative) 为正确预测的不配对的碱基的个数;敏感性 X (sensitivity) 指真实结构中所有的碱基对被正确预测到的百分比;特异性 Y (specificity) 指在所有预测到的碱基对中正确预测的百分比。

在评价 RNA 二级结构预测的结果时,如果天然结构中 i 与 j 配对,那么预测结构中除了 i 与 j 配对, i 与 $j+1$ 或 $j-1$ 以及 j 与 $i-1$ 或 $i+1$ 配对都被认为是正确的配对,这是考虑到了多种因素,包括确定确切配对的困难性,以及这些预测的配对在质量上与正

确配对的一致性^[15]。

2 多目标遗传算法

2.1 多目标优化问题

多目标优化问题的关键点在于如何使各个目标之间同时达到最优^[16-17]。多目标优化问题的解不仅仅是一个全局最优解,而是一个解集,本文称这个解集为 Pareto 最优解集^[18]。在这个解集中彼此任意解都不相互支配,并且解集外的解都会被解集中至少一个解支配。

遗传算法是一种可用于寻找最优解的搜索算法,它借鉴了生物界的进化规律,通过模拟自然进化过程以此来搜索最优解。由于遗传算法运行一次能找到多目标优化问题的多个 Pareto 最优解,因而它是求解多目标优化问题的一个有效手段^[19]。

NSGA-II 是目前多目标优化领域中最优秀的多目标遗传算法之一,它被国内外学者广泛引用,本文将主要结合该算法的思想来进行实验。

2.2 NSGA-II

NSGA-II(Non-dominated sorting genetic algorithm)是一种新型多目标优化遗传算法,相对于 NSGA 的缺陷,NSGA-II 有如下改进:计算复杂性从 $O(mN^3)$ 降至 $O(mN^2)$,具备最优保留机制及无需确定一个共享参数的优点,从而进一步提高计算效率和算法的鲁棒性。NSGA-II 该算法求得的 Pareto 最优解分布均匀,收敛性和鲁棒性好,具有良好的优化效果,是求解多目标优化问题的一种新思路^[20-23]。

NSGA-II 的遗传算法中,通过适应度函数计算每个个体的适应度并对其进行排序,排序算法分为两个步骤。

Step 1 Non-dominated sorting(非支配排序)。种群基于 Pareto 顺序分为 k 个子集(Q_1, \dots, Q_k),每个 Q_i 中的个体都不会被 Q_j 中的个体支配($i < j$)。

Step 2 Crowding-distance sorting(拥挤距离排序)。子群 Q_i 用拥挤距离排序。在不拥挤局域内的解优先度更高。

2.2.1 非支配排序

快速非支配排序是基于非支配排序的改进算法见表 1。

非支配排序。 m 个个体和种群中的其他个体进行支配关系比较,是否支配其他全部个体,复杂度为 $O(mN)$;循环进行直到等级 1 中的非支配个体全部被搜索到,复杂度为 $O(mN^2)$;最坏的情况下,有 N 个等级,每个等级只存在一个解,复杂度为 $O(mN^3)$ 。

表 1 快速非支配排序

Table 1 Rapid non dominated sorting

算法	快速非支配排序
1	fast-nondominat-sort(P)
2	for each $p \in P$
3	for each $q \in P$
4	if $p < q$ then
5	$S_p = S_p \cup \{q\}$
6	else if ($q < p$) then
7	$n_p = n_p + 1$
8	if ($n_p = 0$) then
9	$F_1 = F_1 \cup \{p\}$
10	//个体之间支配关系的比较,引入 S_p 存放和 n_p 记录,循环得到等级 1
11	$i = 1$
12	while ($F_i \neq \emptyset$)
13	$H = \emptyset$
14	for each $p \in F_i$
15	for each $q \in S_p$
16	$n_q = n_q - 1$
17	if ($n_q = 0$) then ($H = H \cup \{q\}$)
18	$i = i + 1$
19	$F_i = H$
20	//循环得到等级 2,等级 3, ..., 复杂度为 $O(MN^2)$

2.2.2 拥挤距离排序

拥挤距离排序用于保持解的多样性。每个个体的拥挤距离是通过计算与其相邻的两个个体在每个子目标函数上的距离差之和来求取,如图 1 所示。

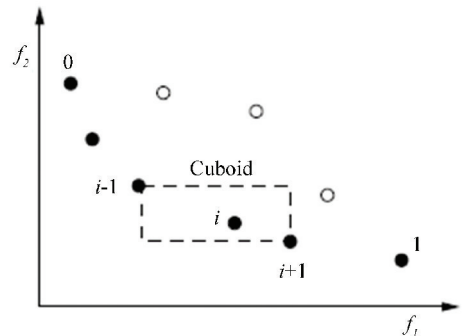


图 1 二目标函数的拥挤距离计算

Fig.1 The crowding distance calculation of two objective functions

图 1 所示个体 i 的拥挤距离为

$$D_i = (f_{i+1,1} - f_{i-1,1}) + (f_{i-1,2} - f_{i+1,2}), \quad (6)$$

即图 1 中虚线四边形的长与宽之和。

排序时当两个个体属于不同等级的非支配解集时,等级较小的优先度高,当两个个体属于同等级

的非支配解集时,拥挤距离较大的优先度高,即

$$\text{if}(i_{\text{rank}} < j_{\text{rank}}) \text{ or } ((i_{\text{rank}} = j_{\text{rank}}) \text{ and } (i_{\text{distance}} > j_{\text{distance}})) \\ \text{then}(i \succ_n j)$$

2.2.3 遗传算法

NSGA-II 中遗传算法流程如图 2 所示。

Step 1 创建一个初始父代种群 P_0 ,使用交叉和变异操作产生子代种群 Q_0 。

Step 2 对 P_0 和 Q_0 组成的整体 R_0 进行非支配排序,构造所有不同等级的非支配解集 Z_1, Z_2, Z_3, \dots 。

Step 3 对分好等级的非支配解集进行拥挤距离排序,根据适应度高低得到前 N 个解,构成下一次迭代的父代种群 P_1 。

Step 4 重复上述 3 个步骤,直到结果收敛。

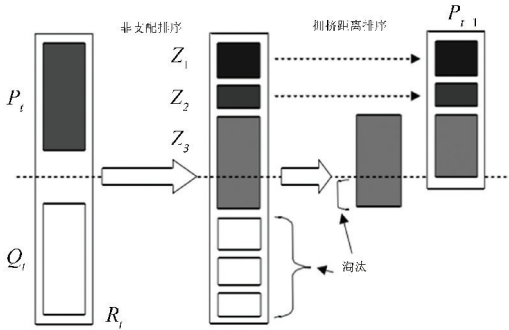


图 2 NSGA-II 步骤
Fig.2 Steps of NSGA-II

3 数据与实验流程

3.1 数据集

本文采用 ASE 以及 CRW 的 RNA 数据集作为实验的研究对象,对 ASE 与 CRW 的 RNA 天然结构进行计算,提取出每个 RNA 结构的一级序列,通过 ProbKnot 算法的配分函数,得到包含碱基配对概率的 pfs 文件,将此文件作为本文输入。经过计算,有效的数据集为 877 个。其中 ASE 案例 450 个,CRW 案例 427 个(见表 2)。ASE 的序列长度大多数在 200~500 之间,CRW 的序列长度大多数在 1 000 以上。

表 2 实验数据集表

Table 2 The experimental dataset

名称	RNA 类型	数据来源	数目
ASE	RNase P	RNase P Database	450
CRW	5S rRNA	Gutell Lab	427

3.2 实验过程

基于多目标遗传算法的含假结 RNA 二级结构预测方法,实验过程如下。

Step1 首先通过对单个 RNA 序列进行变异算法得到不同的 RNA 二级结构作为遗传算法的初始种群,然后通过对初始种群内的个体进行变异交叉算法形成数量相同的新的 RNA 二级结构,再对每个 RNA 二级结构进行适应度计算排序,选择适应度较高的个体组成新的种群。

Step2 重复该步骤直到结果收敛,将种群中的最优个体作为结果输出,得到这个 RNA 序列的二级结构预测结果。

Step3 输入其他每个 RNA 序列来完成上述两个步骤,得到所有 RNA 序列的二级结构预测结果,作为实验结果。其中,为了适应遗传算法中的交叉与变异,算法对 RNA 二级结构采用如图 3 所示的精英保留策略。图 3 中,位于上方的结构(a)表示 1 号和 4 号碱基配对、2 号和 3 号碱基为单链的 RNA 二级结构,它通过变异算法可能形成结构(b)与结构(c)。同样的,位于下方的结构(a)与结构(b)通过交叉算法可能形成结构(c)与结构(d)。

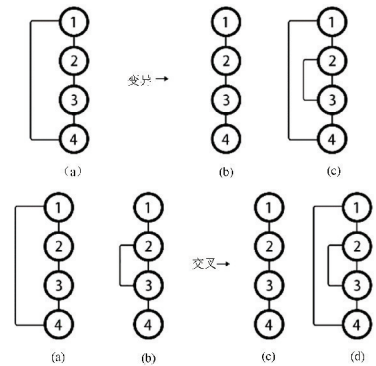


图 3 RNA 二级结构的交叉和变异

Fig.3 Crossing and mutating of RNA secondary structure

3.3 实验参数设置

实验的参数设置会对实验结果产生影响。本次实验中有 4 个参数对实验结果影响较大,它们是 populations, iterations, minloop 与 α , 其中: Populations 是初始种群数, iterations 是收敛后迭代次数。这两个参数值设置的越大,遗传算法搜索到最优解的可能性越高,但运行时间也会越长。根据 RNA 序列的长度大小,本次实验 populations 的取值为 1 000~2 000。由于 RNA 序列长度与种群数的区别,不同的结构使用遗传算法的迭代次数相差较大,所以 iterations 表示收敛后的迭代次数而不是总迭代次数,它的取值为 500~1 000。

Minloop 是最小螺旋长度, α 是用于比较配对概率与单链概率的一个数值。为了提高交叉变异算法的效率,算法对形成的结构进行了两方面的约束:一是在每次交叉变异形成 RNA 二级结构后,算法会去

除长度小于 minloop 的螺旋来保证 RNA 结构的稳定性,因为天然结构中螺旋长度小于 3 的结构较为少见,所以 minloop 取值为 3;二是在交叉变异时,只进行概率大于 α 的配对与解除配对,这样可以减少结构中配对概率与单链概率为 0 或很小的碱基对与单链,避免采用完全随机交叉变异会消耗大量不必要资源的情况, α 取值为 0.001~0.1,这个值越大,算法收敛速度越快,但会降低结果的多样性。

4 结果与分析

4.1 多目标遗传算法与 ProbKnot 算法结果对比

在 877 个实验结果中,有 9 个案例(其中 ASE 案例 2 个,CRW 案例 7 个)的敏感性和特异性在两种算法的计算下都为 0,因此之后的数据统计中删除了这 9 个案例。

图 4、5 比较了实验案例的敏感性与特异性,多目标遗传算法的结果用黑色的点表示,Probknot 算法的结果用灰色的点表示,点所在的位置越是接近右上,结果越好。图 4、5 中表明,单从敏感性与特异性的结果上看,多目标遗传算法的优势并不明显。

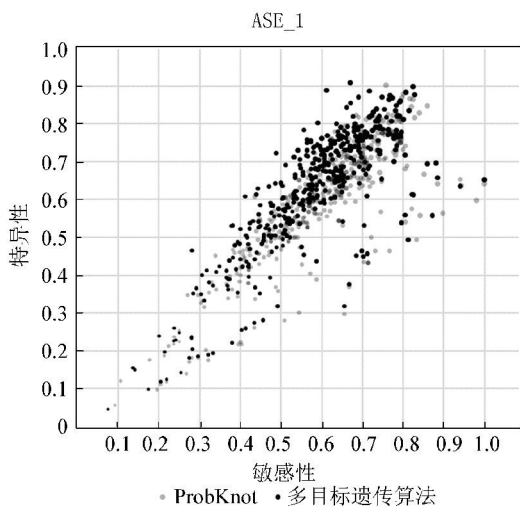


图 4 448 个 ASE 案例敏感性与特异性值统计
Fig.4 The statistic of sensitivity and specificity about 448 cases of ASE

MCC 值是综合敏感性与特异性的评价指标,如图 6、7 所示。多目标遗传算法的结果用黑色的点表示,Probknot 算法的结果用灰色的点表示,点所在的位置越是接近上方,结果越好。从图 6、7 中可以看出,大多数黑色的点在灰色点的上方,这意味着在大多数实验案例上,多目标遗传算法的结果优于 Probknot 算法。

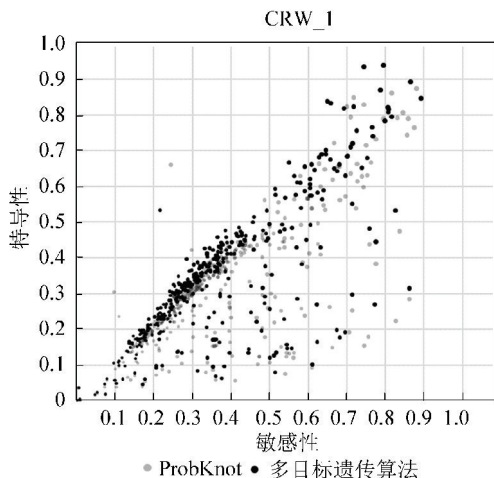


图 5 420 个 CRW 案例敏感性与特异性值统计
Fig.5 The statistic of sensitivity and specificity about 420 cases of CRW

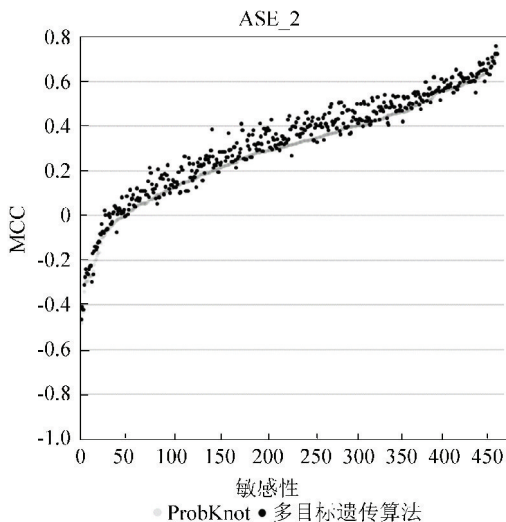


图 6 448 个 ASE 案例 MCC 值统计
Fig.6 The statistic of MCC about 448 cases of ASE

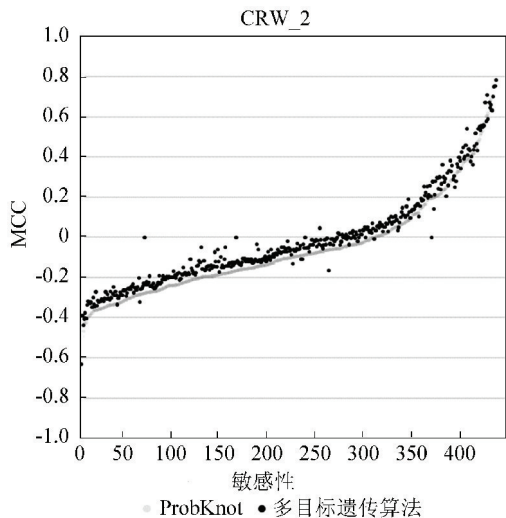


图 7 420 个 CRW 案例 MCC 值统计
Fig.7 The statistic of MCC about 420 cases of CRW

表3整理了 ProbKnot 算法和多目标遗传算法的实验数据的平均值。表3中数据表明,多目标遗传算法在 ASE 和 CRW 数据集上有更高的特异性与 MCC 值。

表3 RNA 二级结构预测方法实验对比表

Table 3 Comparison of methods for predicting RNA secondary structure

RNA	ProbKnot			多目标遗传算法		
	敏感性	特异性	MCC	敏感性	特异性	MCC
ASE	0.597 6	0.592 2	0.294 0	0.584 4	0.623 7	0.331 9
CRW	0.372 6	0.322 4	-0.052 6	0.356 9	0.336 4	-0.011 4

表4 4种 RNA 二级结构预测方法实验对比

Table 4 Comparison of four methods for predicting RNA secondary structure

RNA	ProbKnot			多目标遗传算法			MaxExpect			Fold		
	敏感性	特异性	MCC	敏感性	特异性	MCC	敏感性	特异性	MCC	敏感性	特异性	MCC
ASE	0.589 6	0.564 2	0.293 4	0.579 9	0.594 3	0.329 6	0.531 7	0.548 2	0.165 3	0.505 5	0.495 5	0.085 8
CRW	0.644 4	0.663 2	0.397 2	0.694 5	0.708 7	0.479 1	0.690 5	0.687 5	0.380 5	0.694 8	0.681 0	0.365 9

4.3 实验小结

两次对比实验中,多目标遗传算法在 ASE 和 CRW 数据集上的特异性与 MCC 的平均值更高,大多数案例在改进算法后 MCC 值得到提高,敏感性提升不明显。因此本文可以得出结论,多目标遗传算法的精度更高,得到的结果与天然结构更接近。

5 结论

1) 本文基于多目标遗传算法,使用概率模型进行 RNA 二级结构打分,在多目标优化部分计算出碱基的单链概率,以此完善概率模型。在遗传算法部分根据 RNA 二级结构的特征编写适合的交叉变异算法,以此提高程序运行效率。

2) 实验结果表明,在 ASE 和 CRW 数据集上的对比试验中,多目标遗传算法的精度更高,相比 ProbKnot 算法,MCC 值平均提高 4%。

3) 本文运用多目标优化算法对假结 RNA 二级结构进行了实验预测,证明了多目标遗传算法可以有效地增加含假结 RNA 二级结构预测的精度,但仍然具有发展和改进的空间。

参考文献 (References)

[1] WAN Yue, QU Kun, ZHANG Qiangfeng, et al. Landscape and variation of RNA secondary structure across the human transcriptome [J]. Nature, 2014, 505 (7485): 706-709. DOI:10.1038/nature12946.

4.2 多目标遗传算法与其他 RNA 二级结构预测算法结果对比

本文选取了 263 个案例(其中 ASE 案例 158 个, CRW 案例 105 个),采用 ProbKnot 算法、多目标遗传算法、MaxExpect 算法、Fold 算法^[24]进行对比实验。

表4整理了 ProbKnot 算法、多目标遗传算法、MaxExpect 算法、Fold 算法的实验数据的平均值。表4中数据表明,4种 RNA 二级结构预测方法中,多目标遗传算法在 ASE 和 CRW 数据集上有更高的特异性与 MCC 值,在 CRW 数据集上的特异性也较高。

由于 CRW 数据集中 RNA 序列长度相差较大,第2次对比实验选取了一些长度较短的案例,表3中 CRW 数据集的总体精度比表2要高。

[2] PENALVA L O F. RNA secondary structure [M]. New York: Encyclopedia of Systems Biology, 2013: 1864-1864. DOI:10.1007/978-1-4419-9863-7_319.

[3] ZUKER M. Mfold web server for nucleic acid folding and hybridization prediction [J]. Nucleic acids research, 2003, 31 (13): 3406-3415. DOI: 10.1093/nar/gkg595.

[4] RUAN Jianhua, STORMO G D, ZHANG Weixiong. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots [J]. Bioinformatics, 2004, 20(1): 58-66. DOI: 10.1093/bioinformatics/btg373.

[5] MATHEWS D H, DISNEY M D, CHILDS J L, et al. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 (19): 7287-7292. DOI: 10.1073/pnas.0401799101.

[6] BELLAOUSOV S, MATHEWS D H. ProbKnot: fast prediction of RNA secondary structure including pseudoknots [J]. Rna, 2010, 16 (10): 1870-1880. DOI: 10.1261/ma.2125310.

[7] BINDEWALD E, KLUTH T, SHAPIRO B A. CyloFold: secondary structure prediction including pseudoknots [J]. Nucleic Acids Research, 2010, 38 (suppl_2): W368-W372. DOI:10.1093/nar/gkq432.

[8] SATO K, KATO Y, HAMADA M, et al. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming [J]. Bioinformatics, 2011, 27(13): i85-i93. DOI:10.1093/bioinformatics/btr215.

[9] JABBARI H, CONDON A. A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary struc-

- tures[J]. *BMC Bioinformatics*, 2014, 15: 147. DOI: 10.1186/1471-2105-15-147.
- [10] DAWSON W K, TAKAI T, ITO N, et al. A new entropy model for RNA: part III. Is the folding free energy landscape of RNA funnel shaped? [J]. *Journal of Nucleic Acids Investigation*, 2014, 5: 2652. DOI: 10.4081/jnai.2014.2652.
- [11] MATHEWS D H. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization [J]. *Rna*, 2004, 10(8): 1178-1190. DOI:10.1261/rna.7650904.
- [12] SEETIN M G, MATHEWS D H. TurboKnot: rapid prediction of conserved RNA secondary structures including pseudoknots[J]. *Bioinformatics*, 2012, 28(6): 792-798. DOI: 10.1093/bioinformatics/bts044.
- [13] YONEMOTO H, ASAI K, HAMADA M. A semi-supervised learning approach for RNA secondary structure prediction [J]. *Computational Biology and Chemistry*, 2015, 57: 72-79. DOI:10.1016/j.compbiolchem.2015.02.002.
- [14] LU Z J, GLOOR J W, MATHEWS D H. Improved RNA secondary structure prediction by maximizing expected pair accuracy[J]. *Rna*, 2009, 15(10): 1805-1813. DOI:10.1261/rna.1643609.
- [15] DEIGAN K E, LI T W, MATHEWS D H, et al. Accurate SHAPE-directed RNA structure determination[J]. *Proceedings of the National Academy of Sciences*, 2009, 106(1): 97-102. DOI:10.1073/pnas.0806929106.
- [16] DEB K. Multi-objective optimization [M]. Springer US: Search Methodologies, 2014: 403-449.
- [17] DEB K. Multi-objective optimization using evolutionary algorithms[M]. New York, NY: John Wiley & Sons, 2001: 3-34. DOI: 10.1007/978-0-85729-652-8_1.
- [18] ARNOLD B C. Pareto distribution [M]. New York, NY: John Wiley & Sons, Ltd, 2015. DOI: 10.1002/9781118445112.stat01100.pub2.
- [19] LI Hui, ZHANG Qingfu. Multiobjective optimization problems with complicated Pareto sets, MOEA/D and NSGA-II [J]. *IEEE Transactions on Evolutionary Computation*, 2009, 13(2): 284-302. DOI: 10.1109/TEVC.2008.925798.
- [20] DEB K, PRATAP A, AGARWAL S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II [J]. *IEEE Transactions on Evolutionary Computation*, 2002, 6(2): 182-197. DOI: 10.1109/4235.996017.
- [21] DEB K, AGRAWAL S, PRATAP A, et al. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II [C]//International Conference on Parallel Problem Solving From Nature. Berlin Heidelberg: Springer, 2000: 849-858. DOI: 10.1007/3-540-45356-3_83.
- [22] DEB K, JAIN H. Handling many-objective problems using an improved NSGA-II procedure [C]//2012 IEEE Congress on Evolutionary Computation. Brisbane, QLD, Australia: IEEE, 2012: 1-8. DOI: 10.1109/CEC.2012.6256519.
- [23] SEADA H, DEB K. U-nsga-III: A unified evolutionary optimization procedure for single, multiple, and many objectives: Proof-of-principle results [C]//International Conference on Evolutionary Multi-Criterion Optimization. Cham: Springer, 2015: 34-49. DOI: 10.1007/978-3-319-15892-1_3.
- [24] BELLAOUSOV S, REUTER J S, SEETIN M G, et al. RNAstructure: Web servers for RNA secondary structure prediction and analysis [J]. *Nucleic acids research*, 2013, 41: W471-W474. DOI:10.1093/nar/gkt290.