PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2009, 36(9): 1222~1228
www.pibb.ac.cn

基于量子进化算法的 RNA 序列-结构比对

赵英杰* 王正志
(国防科技大学机电工程与自动化学院自动化所，长沙 410073)

摘要 多序列比对是计算分子生物学的经典问题，也是许多生物学研究的重要基础步骤．RNA 作为生物大分子的一种，不同于蛋白质和 DNA，其二级结构在进化过程中比初级序列更保守，因此要求在 RNA 序列比对中不仅要考虑序列信息，更要着重考虑二级结构信息．提出了一种基于量子进化算法的 RNA 多序列 - 结构比对程序，对 RNA 序列进行了量子编码，设计了考虑进结构信息的全交叉算子，提出了适合于进行 RNA 序列 - 结构比对的适应度函数，克服了传统进化算法收敛速度慢和早熟问题．在标准数据库上的测试，证实了方法的有效性．

关键词 RNA 多序列 - 结构比对，二级结构，量子进化算法，全交叉算子，适应度函数
学科分类号 TP301 **DOI:** 10.3724/SP.J.1206.2009.00047

序列比对是生物序列分析的基础，传统的序列比对工具(如 ClustralW[1]、T-Coffee[2])通常只考虑初级序列信息，一般用于数据库搜索、序列保守区域探测、系统进化分析和序列特征探测等方面．RNA 作为生物大分子的一种，其结构在进化过程中比序列具有更强的保守性，特别是众多非编码 RNA 的功能更多地依赖结构来体现．许多 RNA 相关的分析研究也正是基于这一特点，如 RNA 结构分析[3~5]、RNA 同源搜索[6]、非编码 RNA 探测[7,8]和基于 RNA 的系统进化推断[9]等．而这些方法都以准确的序列比对为基础(这里的准确是指序列比对不仅要考虑序列信息，而且要更多地考虑结构信息)，此时，传统的、基于序列信息的比对工具就显得力不从心了．最近的研究[10,11]显示，当平均双序列一致性低于 50%～60%时，只是比对 RNA 的序列信息，并不能得到满意的结果．

1 RNA 序列-结构比对

Sankoff[12]第一个提出同时考虑了序列比对和结构预测的方法，不幸的是，它的计算代价太高：当序列长度为 L、序列数为 N 时，计算的时间复杂度为 $O(L^{3N})$，空间复杂度为 $O(L^{2N})$．同样采用 Sankoff 算法，Corpet 等[13]和 Bafna 等[14]虽然减少了动态规划算法的运行时间，但对大的序列仍然无法使用．同类的已实现程序还有 Foldalign[15]、Dynalign[16]和 Murlet[17]，它们都是执行限制版的 Sankoff 算法．而程序 PMmulti[18]则是通过渐进比较碱基配对概率矩阵来实现多 RNA 序列 - 结构比对，同类程序还有 FoldalignM[19]、Locarna[20]和 Stemloc[21,22]．程序 MARNA[23]的思路是，在双序列比对时考虑进结构信息，而后用 T-Coffee[2]程序执行多序列比对．通过采用了一种新的打分方法，程序 StrAl[24]在考虑序列相似性的同时，也通过配对概率的形式结合了结构信息．基于图论方法的程序 Lara[25]则是将序列结构比对问题转换成一个整数线性规划问题(ILP).

序列比对方法一般可以分为三类：第一类是精确比对方法，通常采用动态规划算法，只能处理小数量的序列；第二类是渐进比对方法,首先进行所有序列的两两比对，得到一个距离矩阵，然后据此推断出一个进化指导树，先比对进化中“最近的”一对序列，而后根据指导树的“指导”，依据进化的“远近”关系依次加入所有序列．第三类是用迭代更新的方法来同时对齐所有的序列．前面提到 RNA 结构 - 序列比对方法多数属于渐进比对方法，这类方法的一个缺点是，序列比对初期引入的错

* 通讯联系人.
Tel: 0731-4574991, E-mail: matriz@163.com
收稿日期：2009-01-19，接受日期：2009-04-22

误，无法在后续的比对过程中得到改正．而迭代方法可以在比对过程中通过不断地更新比对序列来修正这种错误，理论上可以通过反复修正达到最优比对．

迭代更新方法的基本思想是采用自然进化理论，初始一个比对个体种群，然后更新每个个体，并用目标函数来评估每一代个体，直到找到最好的比对．根据改进比对质量的策略，迭代法可以分为确定性迭代法和随机性迭代法，其中又以确定性迭代法最简单．随机迭代方法包括隐马尔可夫模型(Hidden Markov Model，HMM)[26]、模拟退火(simulated annealing)[27]、禁忌搜索(tabu search)[28]、遗传算法(genetic algorithms，GAs)[29~33]和进化设计(evolutionary programming)[34]，这些方法的主要优点是优化过程和评估标准是分离的，目标函数确定了优化过程的目标．

2 算法描述

进化算法(evolutionary algorithms，EA)是建立在生物进化模型基础上的一种随机搜索技术，它起源于遗传算法(genetic algorithms，GA)[35]，理论上已经证明，进化算法能从概率意义上以随机的方式寻求到问题的最优解，但在实际应用中，还存在较多缺点：容易产生早熟、收敛速度慢、局部寻优能力差等．EA 优化个体的过程，首先是采用选择操作选出比较好的个体，使个体逐渐趋于最优，同时采用交叉变异等进化操作，通过它们的破坏性影响保证产生新的个体，从而生成更好的个体，更重要的是可以维持群体的多样性，即 EA总是在追求群体的收敛性和个体的多样性之间的平衡．若选择压力不足，则算法的收敛速度慢，反之，个体的多样性又不够，算法易陷入局部最优．

融合进化计算和量子计算的研究开始于 20 世纪 90 年代后期，一般可以分成两类：一类集中在用自动程序设计技术来开发新的量子算法[36~38]；另一类集中在开发用于数字计算机的量子激励的进化计算，是进化计算研究的分支，它的特点是在进化算法中运用了量子力学的某些原理，如不确定性、叠加、干扰等[39~42]．

量子进化算法(quantum evolutionary algorithm，QEA)是量子计算与进化算法融合的产物，它以量子计算的一些概念和理论为基础，用量子位编码来表示染色体，并利用量子门实现染色体的更新操作，具有种群规模小、全局寻优能力强、收敛速度快和计算时间短的特点，从而实现目标优化过程．

本文将 RNA 序列 - 结构比对视为一个优化问题，采用量子比特编码方式和量子计算中的量子门进行更新，结合遗传算法的全局寻优能力进行优化，设计了一种新的全局交叉算子和变异算子，提出了适合于进行 RNA 序列结构比对的目标函数．在标准数据库 BRAliBase[10]上的测试，显示了我们方法的有效性．

2.1 多序列比对的定义

给定字母表 $\Sigma=\{A, U, G, C\}$和长度分别为 l_j 的 n 条序列 $S=\{S_i\}$，$i\in[1, n]$，$S_i=S_{i1}S_{i2}\cdots S_{il_i}$，$S_{ij}\in\Sigma$，$j\in[1, l_i]$，则 S 的多序列比对可以用一个 $n\times l$ 矩阵 $M=(a_{ij})$表示且满足下列条件：

a．$i\in[1, n]$，$j\in[1, l]$，$l\geqslant\max(l_i)$，$a_{ij}\in\Sigma'=\Sigma\cup\{-\}$，“$-$”表示空格；

b．如果将 M 每一行中 $a_i=a_{i1}a_{i2}\cdots a_{il}$ 的空格剔除，就精确对应一条序列 S_i；

c．M 中不包含全是空格的列．

我们可以通过为每一个比对赋予相应的分值(适应度或目标函数)来估计比对的质量，多序列比对的目标就是找到最大分值的比对．

2.2 RNA 多序列-结构比对的量子编码

在量子计算中，最小的信息单位是一个量子比特(qubits)，其状态可以为 0 或 1，也可以是任一叠加态，可以用一个 Hilbert 空间的二维单位矢量来表示：

$$|\psi\rangle=\alpha|0\rangle+\beta|1\rangle \tag{1}$$

式中，α、β 是两个复数，表示对应状态的概率幅值且满足 $|\alpha|^2+|\beta|^2=1$，$|\alpha|^2$、$|\beta|^2$ 分别表示量子比特处于状态 0 和状态 1 的概率．在量子进化算法中，使用了基于量子比特的编码方式，即用一对复数定义一个量子比特．一个具有 n 个量子比特位的系统可以描述为：

$$\begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_n\\ \beta_1 & \beta_2 & \cdots & \beta_n\end{bmatrix} \tag{2}$$

其中 $|\alpha_i|^2+|\beta_i|^2=1(i=1,\cdots, n)$．采用这种编码方式，就可以同时表示 2^n 个状态．

量子算法包括在量子系统上连续应用一系列量子操作，量子操作通过量子门和量子回路来实现．量子系统的进化，可以通过作用于量子比特上的线性酉算子 U 来实现：

$$U|\psi\rangle=U[\alpha|0\rangle+\beta|1\rangle]=\alpha U|0\rangle+\beta U|1\rangle \tag{3}$$

线性酉算子 U 通常定义为：

$$\begin{pmatrix} \alpha_1' \\ \beta_1' \end{pmatrix} = \begin{pmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} \quad (4)$$

为了能将量子原理用于序列比对，需要将序列比对转换成量子表示．首先将比对转化为一个二值矩阵 P，其中每一行表示比对 S 中的每一条序列 S_i，P 中元素 1 表示这个位置对应着碱基，0 表示此位置对应着空格．用量子计算中的术语表示，每条序列都用一个量子寄存器表示：

$$S_i=S_{i1}S_{i2}\cdots S_{il}\rightarrow \begin{pmatrix} \alpha_{i1} & \alpha_{i2} & \cdots & \alpha_{il} \\ \beta_{i1} & \beta_{i2} & & \beta_{il} \end{pmatrix}, i\in[1, n] \quad (5)$$

这个寄存器包含了这条序列所有可能比对状态组合的叠加，其中每一列对应该条序列在比对中的元素(空格或是碱基)．概率幅值 α_i 和 β_i 是实数值，满足 $|\alpha_i|^2+|\beta_i|^2=1(i=1, \cdots, n)$，对于每个量子比特来说，对应的二元值可以通过它的概率 $|\alpha_i|^2$ 和 $|\beta_i|^2$ 来计算，$|\alpha_i|^2$ 和 $|\beta_i|^2$ 可以解释为比对 M 中元素为碱基或空格的概率．因此，所有可行的比对可以表示成一个量子矩阵 Q，它包含了比对 M 中所有序列对应的量子寄存器：

$$Q=\begin{bmatrix} \begin{pmatrix} \alpha_{11} & \alpha_{22} & \cdots & \alpha_{1l} \\ \beta_{11} & \beta_{22} & & \beta_{1l} \end{pmatrix} \\ \begin{pmatrix} \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2l} \\ \beta_{21} & \beta_{22} & & \beta_{2l} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nl} \\ \beta_{n1} & \beta_{n2} & & \beta_{nl} \end{pmatrix} \end{bmatrix} = \left\{ \begin{pmatrix} \alpha_{ij} \\ \beta_{ij} \end{pmatrix} \right\}, i=1\cdots n, j=1\cdots l \quad (6)$$

量子矩阵可以看作是所有潜在比对的概率表示，所有比对构象都重叠在这个表示中．当嵌入到进化框架中时，它就起到染色体的作用．

2.3　算法流程

基于量子遗传算法的 RNA 多序列 - 结构比对程序可描述如下．

输入：序列集合 S

1. 用 ClustalW 产生初始比对 S'，用 P'表示对应二元矩阵；
2. 令最优比对 $S_{best}=S'$，对应二元矩阵 $B=P'$，最优比对分值 $C_{best}=C(S')$；
3. 根据 S'的长度，初始化量子矩阵 Q；

Repeat：

4. 对 Q 应用测量操作，得到相应二元矩阵 P 和比对 S，根据 P 和 B 对 Q 应用干扰操作；
5. 对 Q 应用突变操作；
6. 对 Q 应用交叉操作；
7. 评估当前比对 S；
8. 如果 $C(S_{best})<C(S)$, 则 $S_{best}=S$、$C(S_{best})=C(S)$和 $B=P$;

Until：终止条件达到

输出：S_{best} 和 $C(S_{best})$

算法采用 ClustalW 的比对结果作为初始最优比对，以便于实施干扰操作时选择合适的旋转角．量子矩阵的初始化是将所有个体的所有比特位都赋值为 $1/\sqrt{2}$，即$(\alpha_{ij}, \beta_{ij})'=(1/\sqrt{2}, 1/\sqrt{2})'$．对量子矩阵的测量操作是为了得到一组确定解的二元矩阵 $P=\{p_{ij}\}$，方法是：产生一个[0,1]间的随机数 r 作为标准，若 $|\alpha_{ij}|^2>r$ 则 $p_{ij}=1$，否则 $p_{ij}=0$．干扰操作是量子进化算法有别于传统进化算法的重要方面，它根据当前解和历史最优解，由事先确定的调整策略(见 **2.4** 节)更新种群中的个体，使种群始终向着最优解方向进化．量子突变操作采用简单的单量子比特突变操作：随机确定一个突变位点，将该量子比特的几率幅 α_i 和 β_i 位置对调．

交叉是遗传算法的一种寻优的途径，一般有单点交叉、多点交叉、均匀交叉和算术交叉等．这里，我们使用量子的相干特性，并结合 RNA 序列 - 比对的特点，构造了一种新的全干扰配对保守交叉算子，种群中所有个体都参与交叉，并且在交叉过程中保持两个父个体中共有的碱基对(图 1)．交叉流程为：a．构造种群中所有个体的两两组合；b．对任意一对二元矩阵 $P1$ 和 $P2$，将其还原成序列比对 $S1$ 和 $S2$；c．分别对 $S1$ 和 $S2$ 应用 RNAalifold[43]程序预测一致二级结构 $STR1$ 和 $STR2$；d．找出 $STR1$ 和 $STR2$ 中共有的碱基对(对应 $S1$ 和 $S2$ 中的碱基列对必须一致)，作为交叉过程中的保守区，同时找出 $S1$ 和 $S2$ 中相同的保守区域，并用这两类保守区将比对划分成连续的区段；e．两类保守区域直接进入子比对，其他区域的选择由该区段所含碱基对(由程序 RNAalifold[43]预测的结构)数及空格位多少来决定；f．以上操作是在实际比对 S

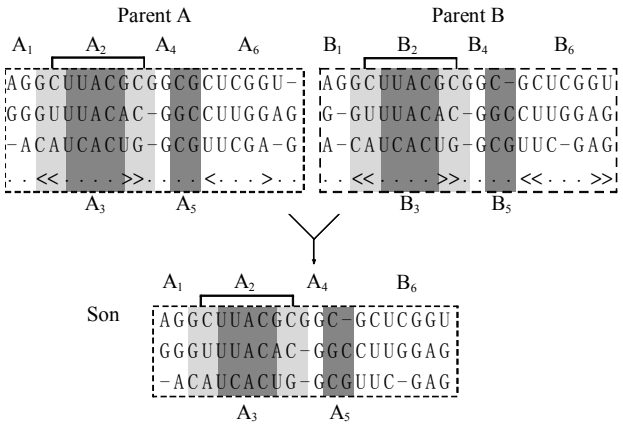

Fig. 1　The crossover operator

上进行，将得到 S 的量子矩阵 Q 作对应操作.

对比对的评估由适应度函数(或目标函数)完成(目标函数的定义见 **2.5** 节). 终止条件为达到最大迭代次数或是连续 50 代适应度函数分值都没有改进为止.

2.4　量子旋转门调整策略

量子干扰操作的目标是增加那些从测量操作中抽取出的、高质量比对的概率. 它主要是将 Q中的每一量子比特向对应最好比特值方向上进行移位. 这个操作可以用量子旋转门在复空间的旋转来完成，如图 2 所示，$|\alpha_{ij}|^2$ 和 $|\beta_{ij}|^2$ 分别给出了量子比特将处在状态 0 和 1 的概率. 在第一、三象限的逆时针旋转将增加 $|\beta_{ij}|^2$，而第二、四象限的逆时针旋转则增加 $|\alpha_{ij}|^2$，它完成一个旋转，旋转角 $\Delta\theta_{ij}$ 是 α_{ij} 和 β_{ij} 以及对应二元值的函数.

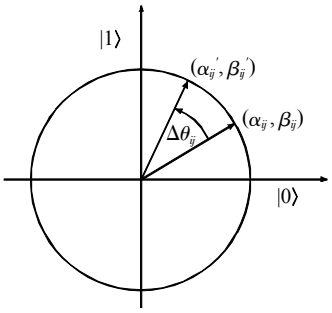

Fig. 2　The quantum interference

旋转角 $\Delta\theta_{ij}$ 的选择要避免过早收敛，方向及大小由 α_{ij} 和 β_{ij} 以及对应二元值的函数来确定，这里采用文献[44]提出的一种通用的、与问题无关的调整策略，如表 1 所示. 旋转角 $\theta_{ij}=s(\alpha_{ij}, \beta_{ij})\Delta\theta_{ij}$，$s(\alpha_{ij}, \beta_{ij})$和 $\Delta\theta_{ij}$ 分别代表旋转的方向和角度，$f(P)$和 $f(B)$分别为当前解和历史最优解的适应度函数值，调整的结果使得几率幅$(\alpha_{ij}, \beta_{ij})$始终向着最优解出现的方向进化. δ 为每次调整的步长，其取之大小影响收敛速度，这里采用类似文献[44]提到的动态调整旋转角策略，即：

$$\delta = 0.5\pi \bullet \exp(-t/t_{\max}) \tag{7}$$

采用动态步长调整策略后，在进化初期，搜索的网格较大，能加速收敛速度；在进化末期，搜索网格较小，从而提高了搜索精度.

Table 1　The selective strategy of rotation angle

p_{ij}	b_{ij}	$f(P)>f(B)$	$\Delta\theta_{ij}$	$S(\alpha_{ij}, \beta_{ij})$			
				$\alpha_{ij}\beta_{ij}>0$	$\alpha_{ij}\beta_{ij}<0$	$\alpha_{ij}=0$	$\beta_{ij}=0$
0	0	False	0	–	–	–	–
0	0	True	0	–	–	–	–
0	1	False	δ	+1	–1	0	±1
0	1	True	δ	–1	+1	±1	0
1	0	False	δ	–1	+1	±1	0
1	0	True	δ	+1	–1	0	±1
1	1	False	0	–	–	–	–
1	1	True	0	–	–	–	–

2.5　适应度函数

本算法中采用类似文献[45]中定义的代价函数作为适应度函数，它不仅要考虑 RNA 序列信息，而且要更多地考虑结构信息，具体由两部分组成，比对项和结构项，详见附录 1(http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20090047&flag=1)及文献[45].

$$cost(A, S)=K \bullet P(A)+\sum_{(i,j)\in S}(\alpha \bullet cost_P(i, j) +\beta \bullet cost_C(i, j))/\#(i, j) \tag{8}$$

式中，$P(A)$ 表示比对分值，$cost_P(i, j)$ 和 $cost_C(i, j)$分别表示结构分值中的自由能分值和共变分值，参数 α,β 是为了平衡各项在组合分值中的贡献，默认值为 α=1.5，β=0.6，可以通过初始网格搜索参数优化获得. K=1000，$\#(i, j)$表示结构 S 中包含的碱基对数，这两个参数是为了数据的规范化.

3　测试数据及结果

3.1　测试数据集

测试数据来自标准 RNA 比对数据库 BRAliBase[10]中的数据集一(dataset-1)，它包含有 481 组比对，平均双序列一致性 29%～95%，这些比对分别来自不同的非编码 RNA 家族，包括 Group Ⅱ Intron、5S rRNA、SRP、tRNAs 和 U5 spliceosomal RNA，每组比对包含 5 条序列.

3.2　评价标准

我们采用 BRAliBase[10]数据库提供的两个独立但互补的分值来评估比对的质量，一个是广泛采用

的 SPS 分值(sum-of-pairs score)[46]，另一个是结构保守指数(structure conservation index，SCI)[47]. SPS 定义为所有既在预测比对也在参考比对中出现的字符对的比例，值域为[0, 1]，预测比对的质量越好，分值越接近 1. 与 SPS 不同，SCI 分值不涉及参考比对，它给出了比对中所包含的保守的二级结构信息，由 RNAalifold[43]程序计算的分值推出. 如果 RNAalifold[43]没有在比对中识别出共同的 RNA 结构，则 SCI 为 0，反之，如果有一组非常保守的结构，则 SCI 接近 1. SCI 的值可以大于 1，此时暗示着保守的二级结构是由互补突变或是一致突变来维系的.

作为比较程序，我们在标准数据集上也运行了 MASTR [45] 和 SAGA [48] (ClustalW [1] 的结果来自 BRAliBase[10]数据库). ClustalW[1]是经典的多序列比对程序，采用了渐进比对的方法，MASTR[45]程序采用了模拟退火方法进行 RNA 序列结构比对，SAGA[48]是一种采用传统遗传算法进行多序列比对的程序，后两者属于迭代更新方法.

3.3 参数设置

对照程序均采用默认设置，SAGA[48]程序所需参考比对序列由 ClustalW[1]产生，修改替代矩阵为 dna_idmat(默认设置为 pam250mt，适用于蛋白质序列). 我们的算法称为 QEA-MRNA，最大进化代数设为 200，种群个数为 10.

3.4 结果及讨论

按照 BRAliBase[10]中的标准，将各程序运行结果统计如表 2 所示，可以看到，QEA-MRNA 得到的比对质量和 MASTR[45]的结果相当，但优于只考虑序列信息的 SAGA[48]和 ClustalW[1]，特别是在较低的平均双序列一致性情况下，说明了该算法的有效性.

Table 2 The statistical results of SCI and SPS score of various programs

Methods	The range of average pair sequence identity					
	High (ID ⩾ 75%), Num=173		Medium (75% > ID⩾55%), Num=218		Low (ID < 55%), Num=90	
	SCI	SPS	SCI	SPS	SCI	SPS
QEA-MRNA	0.8997	0.9276	0.7539	0.8625	0.7944	0.7687
MASTR	0.8949	0.9201	0.7569	0.8204	0.7788	0.7524
SAGA	0.8891	0.9561	0.6813	0.8669	0.5583	0.7023
ClustalW	0.8843	0.8385	0.6840	0.6826	0.5266	0.6046

SCI and SPS score are computed by *RNAz*[47] and *bali_score*[46], respectively.The sequence identity is computed by *alistat* in SQUID[49].

表 3 分别给出了 QEA-MRNA 和 SAGA[48]达到最优值时的进化代数、运行时间及相应的比对长度和平均序列一致性. 可以看出 QEA-MRNA 迭代次数和运行时间都明显少于采用传统遗传算法的 SAGA[48]，充分说明了 QEA-MRNA 算法在运算速度上的优势.

Table 3 The number of generations and runing times when programs find the optimal solution

RNA name	SAGA				QEA-MRNA			
	t/s		generation		*t*/s		generation	
	max	ave	max	ave	max	ave	max	ave
g2intron	1008	315.1	200	80.8	98	20.4	25	13.2
rRNA	768	160.4	150	33.5	65	13.2	18	10.5
SRP	6843	2095.5	240	85.9	861	189.6	40	17.4
tRNA	814	114.1	215	34.8	73	15.7	28	14.2
U5	1112	343.7	240	75.3	110	32.8	39	16.3

4 结　论

量子遗传算法充分利用量子编码的叠加性，提高了种群的多样性，通过量子旋转门实施进化，充分发挥了量子计算的并行性，结合传统遗传算法的突变和交叉操作，使得进化速度和对局部最优沦陷的控制达到了较理想的结果．本文将 RNA 多序列 - 结构比对问题采用量子遗传算法进行求解，并设计了既考虑了序列信息，又兼顾了结构信息的全干扰交叉算子和适应度函数，在 RNA 标准比对数据库上的结果验证了方法的有效性．

采用迭代方法进行 RNA 多序列 - 结构比对，计算代价相对渐进方法没有优势，但比对精度要明显优于渐进算法．迭代方法适用于得到初步比对结果后，需要进一步提高比对精度的情况．迭代方法的另一个明显的优点是，优化过程和目标函数分离，从而可以对不同的目标函数进行测试，以便找出最适合研究问题的目标函数．

参 考 文 献

1 Thompson J D, Higgins D G, Gibson T J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 1994, **22**(22): 4673～4680

2 Notredame C, Higgins D G, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol, 2000, **302**(1): 205～217

3 Bindewald E, Shapiro B A. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA, 2006, **12**(3): 342～352

4 Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res, 2003, **31**(13): 3423～3428

5 Pedersen J S, Meyer I M, Forsberg R, *et al*. A comparative method for finding and folding RNA secondary structures within protein-coding regions. Nucleic Acids Res, 2004, **32**(16): 4925～4936

6 Griffiths-Jones S, Baterman A, Marshall M, *et al*. Rfam: an RNA family database. Nucleic Acids Res, 2003, **31**(1): 439～441

7 Rivas E, Eddy S. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2001, **2**(1): 8

8 Washietl S, Hofacker I, Stadler P. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA, 2005, **102**(7): 2454～2459

9 Hudelot C, Gowri-Shankar V, Jow H, *et al*. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. Mol Phylogenet Evol, 2003, **28**(2): 241～252

10 Gardner P P, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res, 2005, **33**(8): 2433～2439

11 Washietl S, Hofacker I L. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J Mol Biol, 2004, **342**(1): 19～30

12 Sankoff D. Simultaneous solution of the RNA folding,alignment, and proto-sequence problems. SIAM Journal on Applied Mathematics, 1985, **45**(5): 810～825

13 Corpet F, Michot B. RNAlign program: alignment of RNA sequences using both primary and secondary structures. Comput Appl Biosci, 1994, **10**(4):389～399

14 Bafna V, Muthukrishnan S, Ravi R. Computing similarity between RNA strings. Proceedings of the 6th Symposium on Combinatorial Pattern Matching, 1995

15 Gorodkin J, Heyer L J, Stormo G D. Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Res, 1997, **25**(18): 3724～3732

16 Mathews D H, Turner D H. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol, 2002, **317**(2):191～203

17 Kiryu H, Tabei Y, Kin T, *et al*. Murlet: a practical multiple alignment tool for structural RNA sequences. Bioinformatics, 2007, **23**(13): 1588～1598

18 Hofacker I L, Bernhart S H, Stadler P F. Alignment of RNA base pairing probability matrices. Bioinformatics, 2004, **20**(14): 2222～2227

19 Torarinsson E, Havgaard J H, Gorodkin J. Multiple structural alignment and clustering of RNA sequences. Bioinformatics, 2007, **23**(8): 926～932

20 Will S, Reiche K, Hofacker I L, *et al*. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol, 2007, **3**(4): e65

21 Holm L. A probabilistic model for the evolution of RNA structure. BMC Bioinformatics, 2004, **5**(1): 166

22 Holm L, Rubin G M. Pairwise RNA structure comparison with stochastic context-free grammars. Pacific Symposium on Biocomputing, 2002

23 Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics, 2005, **21**(16): 3352～3359

24 Dalli D, Wilm A, Mainz I, *et al*. STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. Bioinformatics, 2006, **22**(13): 1593～1599

25 Bauer M, Klau G W, Reinert K. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. BMC Bioinformatics, 2007, **8**(1): 271

26 Eddy S R. Profile hidden Markov models. Bioinformatics, 1998, **14**(9): 755～763

27 Kim J, Pramanik S, Chung M J. Multiple sequence alignment using simulated annealing. Comput Applic Biosci, 1994, **10**(4): 419～426

28 Riaz T, Wang Y, Li K B. Multiple sequence alignment using tabu search. Procceedings of the 2nd Conference on Asia-Pacific Bioinformatics, Australian Computer Society, New Zealand, 2004

29 Notredame C, Higgins D G. SAGA: sequence alignment by genetic algorithm. Nucleic Acids Res, 1996, **24**(8): 1515～1524

30 Zhang C, Wang A K. A genetic algorithm for multiple molecular sequence alignment. Comput Appl Biosci, 1997, **13**(6): 565～581

31 Anbarasu L A, Narayanasamy P, Sundararajan V. Multiple molecular

sequence alignment by island parallel genetic algorithm. Curr Sci, 2000, **78**(7): 858～863
32 Nguyen H D, Yoshihara I, Yamamori K, *et al*. Aligning multiple protein sequences by parallel hybrid genetic algorithm. Genome Inform, 2002, **13**(1): 123～132
33 Shyu C, Sheneman L, Foster J A. Multiple sequence alignment with evolutionary computation. Genet Prog Evol Mach, 2004, **5** (2): 121～144
34 Chellapilla K, Fogel G B. Multiple sequence alignment using evolutionary programming. Proceedings of the First Congress of Evolutionary Computation, Washington DC, 1999
35 Holland J H. Genetic algorithms and classifier systems: foundations and their applications. Proceedings of the Second International Conference on Genetic Algorithms, 1987
36 Spector L, Barnum H, Bernstein H J, *et al*. Finding a better-than-classical quantum AND/OR algorithm using genetic programming. Proceedings of the First Congress of Evolutionary Computation, Washington DC, 1999
37 Rubinstein B I P. Evolving quantum circuits using genetic programming. Proc. 2001 Congr. Evolutionary Computation, Seoul, Korea, 2001
38 Lukac M, Perkowski M. Evolving quantum circuits using genetic algorithm. Proc. 2002 NASA/DOD Conf. Evolvable Hardware, 2002
39 Han K-H, Kim J-H. Quantum-inspired evolutionary algorithm for a class of combinatorial optimization. IEEE Trans Evol Comput, 2002, **6**(6): 580～593
40 Narayanan A, Moore M. Quantum-inspired genetic algorithms. Proc. IEEE Int. Conf. Evolutionary Computation, Nagoya, Japan, 1996
41 Han K. A graphical tool for parametric simulation of the RNA structure formation. Mol Cells, 2000, **10**(3): 348～355
42 Han K-H, Kim J-H. On setting the parameters of quantum-inspired evolutionary algorithm for practical applications. Proc. 2003 Congr. Evolutionary Computation, Canberra, Australia, 2003
43 Hofacker I L, Fekete M, Stadler P F. Secondary structure prediction for aligned RNA sequences. J Mol Biol, 2002, **319**(5): 1059～1066
44 Junan Y, Bin L, Zhengquan Z. Research of quantum genetic algorithm and its application in blind source separation. Journal of Electronics (China), 2003, **20**(1): 62～68
45 Lindgreen S, Gardner P P, Krogh A. MASTR: Multiple alignment and structure prediction of non-coding RNAs using simulated annealing. Bioinformatics, 2007, **23**(24): 3304～3311
46 Thompson J D, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. Nucl Acid Res, 1999, **27** (13): 2682～2690
47 Washietl S, Hofacker I L, Stadler P F. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA, 2005, **102**(7): 2454～2459
48 Notredame C, Higgins D G. SAGA: sequence alignment by genetic algorithm. Nucl Acid Res, 1996, **24**(8): 1515～1524
49 Eddy S. SQUID-C function library for sequence analysis. http://www.genetics.wustl.edu/eddy/software/#squid, 2004

RNA Sequence-structural Alignment Based on Quantum Evolutionary Algorithm

ZHAO Ying-Jie*, WANG Zheng-Zhi
(*College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha* 410073, China)

Abstract As a classical problem of computational molecular biology, the multiple sequences alignment is also important foundational process. RNA is one of biological polymer, and is different from protein and DNA that the secondary structure of RNA is more conservative than its primary sequence. Therefore, RNA multiple sequences alignment require not only information of sequences, but also information of secondary structures which those sequences will form. Here, a program——QEA-MRNA, which based on quantum evolutionary algorithm(QEA) to align RNA sequences, is proposed. The program introduce a full crossover operator and a fitness function which considering the information of RNA premary sequence and secondary structure, and improving on prematurity controling and the convergent speed. The effectiveness and performance of QEA-MRNA are demonstrated by testing cases in BRAliBase.

Key words RNA multiple sequences alignment, secondary structure, quantum evolutionary algorithm, full crossover operator, fitness function
DOI: 10.3724/SP.J.1206.2009.00047

*Corresponding author.
Tel: 86-731-4574991, E-mail: matriz@163.com
Received: January 19, 2009 Accepted: April 22, 2009

附录 1 适应度函数中比对分值和结构分值的确定

比对分值：比对分值项以全概率形式给出，假设比对中各位点是独立的．给定一个长为 L 包含N条序列的比对，用 x_j^i 表示序列 i 中第 j 个字符，用$P(x_j^i)$表示在这个位点观察到字符x_j^i 的概率，则多序列比对 A 的概率为：

$$P(A)=\prod_{j=1}^{L}\prod_{i=1}^{N}P(x_j^i) \tag{1}$$

$P(x_j^i)$为单个字符的概率(空格也要考虑进来)．如果 x_j^i 是来自字母表 $\Sigma=\{A, C, G, U\}$的碱基，概率 $P(x_j^i)$的计算要考虑这一列的碱基组成：

$$P(x_j^i)=\begin{cases} P_{GO} & x_j^i=-,\ x^i{}_{j=1}\in\Sigma \\ P_{GE} & x_j^i=-,\ x^i{}_{j=1}\in - \\ (1-P_{GO})\dfrac{c_j(a)}{\sum_{b\in\Sigma}c_j(b)} & x_j^i=a\in\Sigma,\ x^i{}_{j=1}\in\Sigma \\ (1-P_{GE})\dfrac{c_j(a)}{\sum_{b\in\Sigma}c_j(b)} & x_j^i=a,\ x^i{}_{j=1}\in - \end{cases} \tag{2}$$

P_{GO} 表示开空格的概率，P_{GE} 表示扩展空格的概率，这两种概率都可以从已知结构比对中估计出来，本文中 $P_{GO}=0.5$, $P_{GE}=0.74$.

结构分值：结构分值定义为单个碱基对分值的和．用 S 表示由碱基对(i, j)组成的结构，则：

$$\mathrm{cost}(S)=\sum_{(i,j)\in S}\mathrm{cost}(i, j) \tag{3}$$

用序列的平均自由能和共变分值来对结构打分，分别用 McCaskill[1]碱基配对概率$P(bp_{i,j})$和 RNAalifold[2]中采用的扩展共变分值 $C(bp_{i,j})$来表示：

$$P(bp_{i,j})=\frac{1}{N}\sum_{s=1}^{N}M^s(i^s, j^s) \tag{4}$$

式中，矩阵 M^s, $s=1, \cdots, N$ 为剔除空格后单条序列的概率矩阵，这些矩阵在整个算法中是不变的．对于比对中一对位点(i, j)，(i^s, j^s)表示空格校正索引，$i^s=i-m^s$，其中 m^s 是序列 s中位置 i前的空格数，类似的处理索引 j．如果序列 s 中位置 i 或 j有一个是空格，则 $M^s(i^s, j^s)=0$.

最终的自由能分值为：

$$\mathrm{cost}_f(i, j)=P(bp_{i,j})/P_{\mathrm{null}} \tag{5}$$

式中，$P_{\mathrm{null}}=0.25$ 是反映了概率矩阵中任意碱基配对的背景概率(其值可以通过参数优化得到).

对互补突变有不同的度量方法，这里采用 RNAalifold[2] 的方法：

$$B_{i,j}=\left(\frac{1}{\binom{N}{2}}\sum_{\alpha<\beta}\delta(x_i^\alpha x_j^\alpha, x_i^\beta x_j^\beta)\Pi_{i,j}^{\alpha}\Pi_{i,j}^{\beta}\right)-q_{i,j} \tag{6}$$

式中，序列 α 的矩阵 Π^α定义为：如果序列 α 能在位置 i和 j形成一个碱基对，则 $\Pi_{i,j}^{\alpha}=1$，否则 $\Pi_{i,j}^{\alpha}=0$．函数 $\delta(x_i^\alpha x_j^\alpha, x_i^\beta x_j^\beta)$表示序列 α 和 β 在位置 i、j 间的 Hamming 距离．罚分项，$q_{i,j}$ 度量了比对中不一致碱基对的比例.

为了考虑碱基对堆积的影响，临近的两个碱基对也被包括进来，但权重要小：

$$\mathrm{cost}_c(i, j)=\frac{B_{i-1,j+1}+2B_{i,j}+B_{i+1,j-1}}{4} \tag{7}$$

参 考 文 献

1 McCaskill J S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers, 1990, **29**(6～7): 1105～1119

2 Hofacker I L, Fekete M, Stadler P F. Secondary structure prediction for aligned RNA sequences. J Mol Biol, 2002, **319**(5): 1059～1066