DNA 计算中的模板框优化方法研究

刘文斌[1],陈丽春[1],白宝钢[1],朱翔鸥[1],张 强[2],马润年[3]
(1. 温州大学计算机科学与工程学院,浙江温州 325027; 2. 大连大学信息科学与工程重点实验室,
辽宁大连 116622; 3. 空军工程大学电讯工程学院,陕西西安 710077)

摘 要: 编码问题是目前 DNA 计算中的重点和难点之一,编码问题的难点就是当这些编码以某种方式线性连接起来表示一个特定的信息(如图的一个路径或一个最大团等),如何确保其中的每个编码能被唯一的识别. 因此,如何有效使用编码是编码研究中要解决的另一个问题. 本文在模板编码的基础上,提出了模板框的概念,并对其移位距离性质进行了研究. 在此基础上,考察了词标长度、单词标及多词标等因素对模板框性能的影响. 计算结果表明:多词标方法能够明显改善模板框的移位距离性质. 最后,指出了模板框优化的进一步的研究方向.

关键词: DNA 计算; 编码问题; 模板编码方法; 模板框
中图分类号: TN18 **文献标识码**: A **文章编号**: 0372-2112 (2007) 08-1490-05

Research on Optimizing the Template Frame in DNA Computing

LIU Wen-bing[1], CHEN Li-chun[1], BAI Bao-gang[1], ZHU Xiang-ou[1], ZHANG Qiang[2], MA Run-nian[3]
(1. *College of Computer Science and Engineering, Wenzhou University, Wenzhou, Zhejiang* 325027, *China*;
2. *University Key Laboratory of Information Science and Engineering, Dalian University, Dalian, Liaoning* 116622, *China*;
3. *Telecommunication Engineering Institute, Air Force Engineering University, Xi'an, Shaanxi* 710077, *China*)

Abstract: The encoding problem is a most fundamental issue in DNA based computing. Its difficulty lies in how can we assure that each code could accurately identify itself in linear DNA sequences. Therefore, how to use those codes effectively becomes an urgent problem. In this paper, we introduce the concept of template frame and its shift distance property based on the template strategy. Then, we study the influence of the length of labels, single labels and multiple labels on the shift distance. The result shows that the multiple label method can improve the shift distance property dramatically. Finally, we point out some possible directions for further studying.

Key words: DNA computation; encoding problem; template method; template frame

1 引言

DNA 计算是近年来计算机研究领域的一个热点方向,其标志是 Adleman 1994 年在 Science 上发表的文章-Molecular Computation of Solution to Combinatorial problems[1]. 在这种新型计算方式中,信息是通过 DNA 分子的四种碱基来编码的,并通过 DNA 分子间的特异性杂交来实现的.

由于 DNA 计算中的核心操作—杂交反应在不完全互补的情况下也有可能发生,从而形成各种不希望的二级结构(如图 1 的 2、3、4、5),从而导致错误的计算结果. 在 PCR 扩增过程中,引物与引物之间同样会出现上述不希望的二级结构,以致扩增失败. 因此,如何通过有效的编码来提高 DNA 计算过程中的"信噪比",是 DNA 计算研究中的一个重点和难点问题.

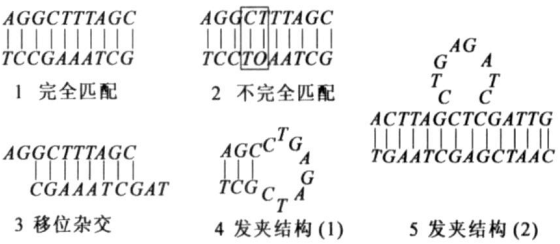

图 1 DNA 的几种可能的杂交形式

编码研究的目的就是希望能够在实际的生化反应过程中,编码每一个信息元的 DNA 序列能够被最大限度地唯一识别,从而使得计算过程能够按照计算模型所设计的方向进行. 目前有关编码的研究主要集中在如何降低编码之间的相似度. Garzon 给出了 DNA 计算中的

收稿日期:2006-08-18;修回日期:2007-05-21
基金项目:国家自然科学基金(No. 60403002, 60403001, 30670486);中国博士后科学基金(No. 2004036130);浙江省自然科学基金(No. Y106654, Y405553)

编码问题定义[2],他还借鉴二进制超立方体的理论对编码进行研究[3]. Baum 提出降低 DNA 序列间的相似度假设[4]. Feldkamp 等给出了另一种定义序列间相似度的方法[5]. Suyama 等在基于 DNA 计算的基因表达分析的 DNA 编码数概念[6]. 有的学者还提出采用三字母表的编码策略[7],来降低 DNA 分子生产二级结构. Braich 等提出了 DNA 序列编码的约束条件,用来解决可满足性问题的编码问题,并在实验中取得了良好的效果[8]. Frutos 等提出的模板编码方法[9],我们对优化模板编码方法做了进一步的研究[10].

得到编码只是编码问题的第一步,如何有效使用编码则是编码研究中要解决的另一个问题. 目前,现有的编码方法都能基本保证编码间有足够的差异. 编码问题的难点就是当这些编码以某种方式线性连接起来表示一个特定的信息(如图中的一个路径或一个最大团等),如何确保其中的每个编码能被唯一识别. 模板编码方法是近年来出现的一种较好的编码方法,在这种编码方法中,模板序列对编码的性质影响最大,同时模板的数量相对编码数量来说又很少. 最近,我们提出了模板框的概念,并发现通过加一个特定的词标可以提高模板框的移位距离性质[11],本文在此基础上系统的对模板框的优化方法进行了研究.

2　编码问题及其约束条件

2.1　编码问题[2]

DNA 计算中的编码问题可以表述为:以构成 DNA 分子的四个碱基为字母表 $\sum = \{A, T, G, C\}$,存在一个长度为 l 的 DNA 分子的基础编码集合 Z, $Z = \sum^l = \{\langle b_1, b_2, \cdots, b_l\rangle \mid b_i \in \sum, i = 1, 2, \cdots, l\}$,显然 $|Z| = 4^l$. 求 Z 的一个子集 W 使得

$$\forall x_i, x_j \in W \quad \tau(x_i, x_j) \geq k \tag{1}$$

其中 k 为正整数,τ 是评价编码的期望准则,如汉明距离、GC 含量、最大相同子序列长度等. 另外,在编码问题中的还要考虑集合 W 的大小 $|W|$. 因为 $|W|$ 越大,可供选择的满足条件的编码就越多. 显然评价准 τ 越严格,可供选择的编码数量 $|W|$ 就越小.

2.2　约束条件

由于 DNA 计算的特殊性,对编码有物理、化学及数学等多方面的约束条件. 为了衡量编码间“相似度”,目前主要以汉明距离及其扩展形式来定量编码间的距离,显然,编码间的距离越大其“相似度”越小.

(1) 汉明距离 $H(x_i, x_j)$:序列 x_i 和 x_j 上所有对应位置上字符不同的总和.

(2) H 测度 $H_G(x_i, x_j)$:是 Garzon 在汉明距离的基础上提出的[12],其定义为

$$H_G(x_i, x_j) = \min_{-n<a<n} H(x_i, \rho^a(x_j)) = n - c_{ij} \tag{2}$$

其中 ρ_a 表示偏移 a 个位置,c_{ij} 为序列 x_j 偏移 a 个位置后与 x^i 的最大相同字符之和. 为了衡量一个序列自身的 H 测度,我们定义当 $x_i = x_j$ 时,$a \neq 0$.

(3) 移位距离 $H_M(x, y)$:是 Arita 在 H 测度的基础上提出的一种定义长度不同的序列 x 和 $y(|x| < |y|)$ 的距离的一种方法,其定义为[13]

$$H_M(x, y) = \min_{1 \leq a \leq |y| - |x| + 1} H(x, y_a) \tag{3}$$

其中,y_a 为 y 中长度为 $|x|$ 的子序列,当 $|x| = |y|$ 时,$H_s(x, y)$ 退化为二进制串之间的汉明距离. 平局移位距离 $\overline{H}_M(x, y)$ 则为

$$\overline{H}_M(x, y) = \sum_{1 \leq a \leq |y| - |x| + 1} H(x, y_a) \tag{4}$$

由于 DNA 序列间容易发生各种不希望的移位杂交,H 测度 H_G 和移位距离 H_S 可以更为准确的描述二个序列间的相似度. H 测度 H_G 和移位距离 H_S 越小,表示两个序列之间的相似度越大,移位杂交发生的可能性越大;反之,两个序列之间的相似度越小,移位杂交发生的可能性越小.

2.3　模板编码方法[9]

模板方法(Template Method)最初是由 Wisconsin 大学的 Frutos 等提出的,其主要思想是:通过一个二进制序列的模板(Template)集合 T 和另一个二进制序列的映射(Map)集合 M 生成满足要求的 DNA 序列集合 S

$$T \times M \rightarrow S$$

其生成规则为 $1 \times 1 \rightarrow T, 1 \times 0 \rightarrow A, 0 \times 1 \rightarrow G, 0 \times 0 \rightarrow C$. 图 2 给出了一个由模板方法生成的长度为 8 的 DNA 序列的示意图.

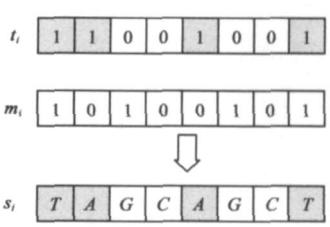

图 2　DNA 序列的生成示意图

模板编码方法的实质是将 $\sum_{DNA} = \{A, G, C, T\}$ 上的编码问题转化为 $\sum_{01} = \{0, 1\}$ 上的编码问题,即在 n 维二进制超立方体空间中搜索满足条件的模板集合 T 和映射集合 M. 模板编码方法的数学基础:当模板 t_i, t_j 或映射 m_i, m_j 具有某种性质时,其生成的二个 DNA 序列 s_i, s_j 也具有此性质.

由于 DNA 序列的方向性,我们不仅要保证编码之间在同一方向时具有足够大的距离,使得它们的互补序列能够唯一的和其对应的编码序列互补;同时还要保证任意二个编码间相互杂交的可能性很小. 我们在优化模板集合的搜索过程中要求:

(1) 任一个模板序列 t_i 满足

$$h(t_i, t_i) = \min(H_G(t_i, t_i^r), H_G(t_i, t_i)) \geq d \tag{5}$$

(2) 任意二个模板序列 t_i, t_j 满足

$$h(t_i, t_j) = \min(H_G(t_i, t_j^r), H_G(t_i, t_j)) \geq d' \tag{6}$$

其中 r 表示反序列，d，d 为正整数，且 $d \geq d'$，一般情况下 $d' \leq \lfloor n/3 \rfloor$.

在模板编码中，由于映射集合序列对编码的影响相对要小，因此，通常仅要求其中任意二个映射序列 m_i，m_j 满足汉明距离 $H(m_i, m_j) \simeq \lfloor n/2 \rfloor$. 在上述约束下，容易看出对任意二个编码 $s_i, s_j \in S$ 有

$$H_G(s_i, s_j) \geq d' \tag{7}$$

$$H_G(s_i^r, s_j^c) \geq d' \tag{8}$$

2.4 模板框的概念

对于模板集合 $T = \{t_1, t_2, t_3, \cdots, t_k\}$，称由这 k 模板序列形成的一个排列

$$p = t_1 t_2 \cdots t_k t_1 \tag{9}$$

为模板集合 T 的一个模板框(Template Frame)，在模板框 P 中要求第一个模板序列和最后一个模板序列必须相同，当模板序列的长度为 n 时，模板框 p 其长度为 $(k+1) \cdot n$. 显然，由模板集合 T 可以形成 $k!$ 不同的模板框 P. 我们定义模板序列 t 和模板框 P 的移位距离为：

$$H_M(t, p) = \begin{cases} H_M(t, p') & t = t_1 \\ \min[H_M(t, p_1'), H_M(t, p_2')] & t \neq t_1 \end{cases} \tag{10}$$

其中 p 为模板框 p 中删除第一个字符和最后一个字符得到的子串，p_1 为 p 中删除从 t_i 的最后一个字符开始的子串后得到的子串，p_2 为 p 中删除从第一个字符开始到 t_i 的第一个字符为止的子串后得到的子串(如图3).

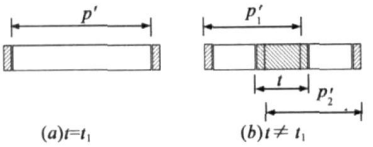

图3 模板框 p 中子串 p, p_1', p_2' 的示意图

模板集合 T 和模板框 P 的移位距离为：

$$H_M(T, p) = \min_{1 \leq i \leq k} H_M(t_i, p) \tag{11}$$

模板框 P 的平均移位的距离为：

$$\overline{H}_M(T, P) = \sum_{1 \leq i \leq k} \overline{H}_M(t_i, p) / k \tag{12}$$

从上面的定义可以看出，对模板集合 T 有

$$H_M(T, p) \leq H_G(t_i, t_j) \tag{13}$$

给定模板集合 $T = \{t_1, t_2, \cdots, t_k\}$，映射集合 $M = \{m_1, m_2, \cdots, m_q\}$，及模板框 $p = t_1 t_2 \cdots t_k t_1$，可以将编码集合 S 按照其对应的模板序列划分为 k 个不相交的子集

$$S = S_1 \cup S_2 \cup \cdots \cup S_k \tag{14}$$

其中 t_i 为子集 S_i 对应的生成模板 $(1 \leq i \leq k)$. 称 $W = (s_{1_1} s_{1_2} \cdots s_{1_m})(s_{2_1} s_{2_2} \cdots s_{2_m}) \cdots (s_{q_1} s_{q_2} \cdots s_{q_m})$ 是按照模板框 p 生成的，当切仅当 W 对应的模板序列 $(T)_W = \underbrace{t_1 t_2 \cdots t_j \cdots t_m}_{1} \times \underbrace{t_1 t_2 \cdots t_j \cdots t_m}_{2} \times \cdots \times \underbrace{t_1 t_2 \cdots t_j \cdots t_m}_{q}$

定理1：给定模板框 $p = t_1 t_2 \cdots t_k t_1$，当 $H_M(T, p) \geq d$ 时，有 $H_M(s_{i_j}, W) \geq d'' (1 \leq i \leq k, 1 \leq j \leq q)$ (证明略)

从模板框的定义可以看出，模板框其实是一个由模板序列组合起来的特定的数据结构，通过它可以将编码序列和由其表示的各种问题的解空间的 DNA 序列间的关系映射为模板序列与模板框之间的对应关系(如图4所示). 这种映射关系有二个优点：(1) 将具体模型的数据结构特征(即问题解空间)和编码方法紧密的联系起来，提高了模板编码方法应用的可靠性；(2) 将编码序列和由其表示的可能解序列的各种距离性质的计算转化为模板序列和模板框之间的各种距离性质的计算，简化了研究的难度.

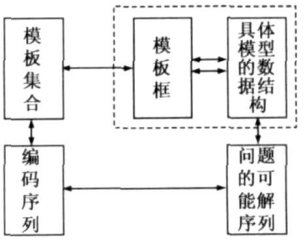

图4 映射关系示意图

显然，模板框的移位距离性质不仅取决于其中所有模板序列的移位距离性质，还与模板序列的组合方式有关. 由于模板框自身的移位距离性质稳定，如果我们能够找到一个最大移位距离的模板框，就可以有效的预防各种不希望的移位杂交的发生. 通过计算发现，如果直接将模板序列连接起来形成模板框时其移位距离都很小. 因此，我们采取一种给模板加词标的方法. 所谓词标就是一个长度较短的二进制序列(其01含量基本为50%)，可以加在模板序列的头部或尾部，起隔断各模板序列的作用，如模板框 $p = t_1 t_2 \cdots t_k t_1$ 加上词标 v 后就成为 $p = vt_1 vt_2 \cdots vt_k vt_1$. 每个模板序列所加的词标可以相同，也可以不同.

3 算法及结果讨论

当模板 T 集合的大小为 k，词标集合 V 的大小 l 时，模板框的搜索空间为 $k! \cdot l^k$. 因此，本文提出如下从一个模板序列开始的生成测试(generate and test)策略的随机搜索算法.

3.1 算法步骤

(1) 从词标集合 V 中随机选择一个词标 v_i 加到模板集合 T 中的模板 t_j 的前面，直到所有模板都加上词标为止，得到一个新的模板集合 T'.

(2) 从 T' 中随机选择一个加词标的模板 t 放入模板框集合 P_0.

(3) 对于 $p_j \in P_0 (1 \leq j \leq |P_0|)$，并计算当前模板框 p_j 中没有使用的模板集合 $T_p \in T'$.

(4) 对于 $t_i' \in T_p (1 \leq i \leq |T_p|)$，计算 T' 和 $p_j t_i$ 的移位距离，若 $H_M(T', p_j t_i) \geq d''$，则将新生成的模板框 $p_j t_i'$ 加入到模板框集合 P_1 中.

(5) 将 $P_0 = P_1$，P_1 清空.

(6) 重复步骤3、4、5，直到 P_0 中的所有模板框的长度正好为 kn.

(7) 对于 $p_j \in P_0 (1 \leq j \leq |P_0|)$,计算 T 和 p_j^t 的移位距离,若 $H_M(T', p_j^t) \geq d''$,则将新生成的模板框 p_j^t 加入到模板框集合 P_1 中.

(8) 将 $P_0 = P_1$, P_1 清空.

3.2 结果分析

由于随机搜索算法每次的结果都会不同,我们对每一种情况执行多次,然后选择其中最好的结果进行比较.由于满足最小移位距离 d 的模板框并不唯一,因此,可以用平均移位距离来衡量模板框的质量.最佳模板框是指在最小移位距离 d 的条件下模板框的平均移位距离最大的那些模板框.下面我们从单词标和多词标、词标长度等、及搜索阀值方面对模板框的性能进行讨论.

3.2.1 单词标和多词标

在 $n = 8$ 的时我们挑选一个有 10 个模板序列的模板集合 T 和一个包含 5 个词标的词标集合 V.为了提高模板框的移位距离,在搜索过程去掉了其中的二个模板 00010111 和 01110001.计算结果如表 1 所示,表中列出了部分最佳模板框.在文献[10]中我们的计算结果显示在时后不加词标的最佳模板框的最小移位距离为 1,而在加了词标后模板框的最小移位距离增加为 3.对于多词标和单词标而言,虽然多词标没有增加模板框的最小移位距离,但平均最大移位距离由 4.1 增加为 5.3.图 5 中列出了二种情况的每个模板框的平均移位距离,为了便于比较,模板框总数是一样的.这表明多词标策略可以明显改善模板框的局部移位距离性质,即对于任一模板序列,使得其和模板框中的大部分子序列的汉明距离都大于最小移位距离.

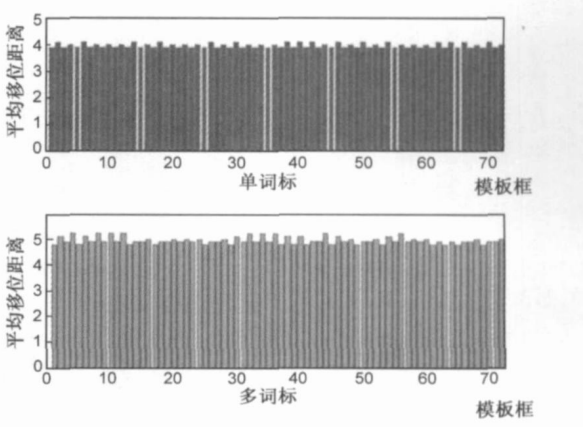

图 5　模板框平均移位距离分布 (n=8)

表 1　单词标和多词标 $n = 8$

多词标									
模板框	010101001101 100100111100 110011010100 011001011010 101011000011 010110011001 100110100101 100101100110 010101001101 010101001101 011001011010 011010011001 110011010100 101011000011 100100111100 100110100101 100101100110 010101001101								
总数	144	最佳模板框数	9	最小移位距离	3	最大移位距离	7	最大平均移位距离	5.3
单词标									
模板框	110001011010 110001100110 110011000011 110001001101 110010011001 110011010100 110010100101 110000111100 110001011010 1100								
总数	72	最佳模板框数	12	最小移位距离	3	最大移位距离	7	最大平均移位距离	4.1
词标集合	0011 1100 0101 1010 0110 1001								
模板集合	00010111 00111100 01001101 01011010 01100110 01110001 10011001 10100101 11000011 11010100								

3.2.2 词标长度

由于词标是在二个模板序列间起间隔的作用,因此,适当增加词标的长度应该可以改善模板框的移位距离性质.表 2 以 $n = 16$ 为例列出了词标长度 $n' = 4$ 和 $n' = 5$ 的计算结果,可以看出,最小移位距离和最大平局移位距离分别由 4 增加到 6 和 8.3 增加为 9.2.图 6 给出了二种情况模板框中每个模板序列和模板框的平局移位距离的分布图.

表 2　词标长度 $n = 16$

	模板框数	最佳模板框数	最小移位距离	最大移位距离	最大平均移位距离
$n' = 4$	17	17	4	12	8.3
$n' = 5$	58	11	6	12	9.2
词标集合 1	0011 1100 0101 1010 0110 1001				
词标集合 2	01011 11010 01100 10101				
模板集合	0001110100101110 0010100111100101 0010111000111010 1001001110110001 1011100100011010 1100010110100011 1100110100010110 1101011000011010 1111111000100000				

表 3　搜索阀值

最小移位距离	模板框数	最佳模板框数	最大移位距离	最大平均移位距离
7	7	1	15	12.8
6	138	16	16	11.6
词标集合	0011 1100 0101 1010 0110 1001			
模板集合	000000110110111010100111 0 111111111001010100000000 0 000110111000101001111010 0 000011101011001100101011 000111010010110110001011 0 101001110000011010011011 0 000101001001111010110011			

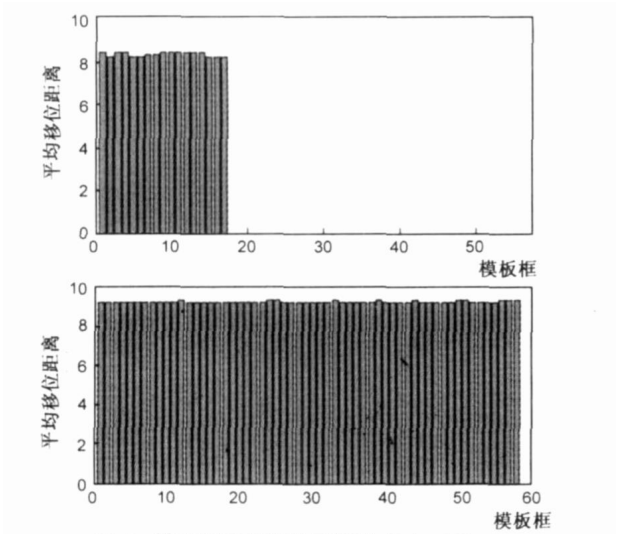

图 6 模板框平均移位距离分布 (n=16)

3.2.3 搜索阈值

由于在多词标方法下模板框的搜索空间为 $k!\ l^k$,为了提高搜速算法的效率,我们可以设定合理的搜索阈值(即最小移位距离). 表 3 以模板长度 $n=24$ 为例列出了二种不同阈值 $d''=6$ 和 $d''=7$ 的计算结果. 显然,随着搜索阈值的增加,最终得到的模板框的数量急剧减少. 这说明当算法进行到一定程度后,很多候选模板框将随着新模板序列的加入而被淘汰掉. 从我们的计算结果看,初始的搜索阈值设为 $d''\approx \lfloor n/3 \rfloor$ 较为适宜,然后在根据搜索结果作适当调整.

4 结束语

编码问题是 DNA 计算中的一个基本问题,它直接影响到 DNA 计算的可靠性和计算效率. 由于 DNA 计算的特殊性,在得到编码后如何有效的使用这些编码则是 DNA 计算编码问题研究中面临的又一个难题. 本文在模板编码方法的基础上,介绍了模板框的概念,并对其移位距离性质进行了研究. 为了提高模板框的移位距离性质,我们提出多词标的策略,由于其具有更大的灵活性,计算结果表明:和单词标相比,多词标可以进一步提高模板框的移位距离性质.

模板框使得我们可以有效地使用编码,避免各种不希望的移位杂交现象的发生. 因此,如何进一步优化模板框将是一个值得进一步深入研究的问题,如能否将模板集合分组? 此外,随着编码长度的增加,词标长度及其数量也随之有所增加,如何选取更为合理的词标集合? 最后,应该说明的是,虽然模板框有良好的性质,但其应用范围也是有限的,即对于那些解空间具有与编码顺序无关的问题,如可满足问题、各种 Sticker 模型及构建基于 DNA 的大规模数据库等. 因此,对于其它问题有没有类似模板框的结构可以应用也是值得研究的问题.

参考文献:

[1] L Adleman. Molecular computation of solution to combinatorial problems[J]. Science, 1994, 266(11): 1021 - 1024.

[2] M Garzon et al. A new metric for DNA computing[A]. Proceedings of the 2nd Annual Genetic Programming Conference GP-97[C]. Morgan Kaufmann, Stanford University, 1997. 472 - 487.

[3] Garzon M, Deaton R, Nino L F, Stevens S E, Wittner M. Encoding genome for DNA computing[A]. The Third DIMACS Workshop on DNA-based Computing. American Mathematical Society[C]. Philadelphia, 1997. 230 - 237.

[4] E B Baum. DNA sequences useful for computation[A]. Proc. Second Annual Meeting on DNA Based Computers, American Mathematical Society[C]. University of Leiden, The Netherlands, 1996. 122 - 127.

[5] Feldkamp, et al. A DNA sequence compile[A]. Proceedings of 6th DIMACS Workshop on DNA Based Computers[C]. University of Leiden, The Netherlands, 2000. 253 - 263.

[6] A Suyama, et al. DNA chips-integrated chemical circuits for DNA diagnosis and DNA computers[A]. Proc 3rd International Micromachine Symp [C]. Tokyo: Japan Science museum, 1997. 7 - 12.

[7] Encoding Choices for Error Resistant DNA Computers[DB/OL]. www. csd. uwo. ca/～morey/ dnatalk/ kevin/ dna/ dnaerror. html

[8] Ravinderjit S Braich, Cli. Johnson, Paul W K Rothemund, Leonard M Adleman. Solution of a satisfiability Problem on a Gel-Based DNA Computer[A]. The 6th International Workshop on DNA-Based Computers[C]. Leiden, The Netherlands, LCNS 2054, 2001. 27 - 42.

[9] A G Frutos, et al. Demonstration of a word design strategy for DNA computing on surface[J]. Nucleic Acids Research, 1997, 25(23): 4748 - 4757.

[10] Wenbin Liu, Shudong Wang, Lin Gao, Jin Xu. DNA sequence design based on template strategy[J]. Chem Inf Comput Sci, 2003, 43(6), 2014 - 2018.

[11] 刘文斌. DNA 计算中的编码问题及模型研究[D]. 华中科技大学博士学位论文. 武汉. 2004. 1.

[12] M garzon, P. Neathery, P Deaton. A new metric for DNA computing[A]. In Proc of 2nd Annual Genetic Programming Conference[C]. Morgan Kaufmann, 1997. 472 - 478.

[13] Arita M, Kobayashi S. DNA sequence design using template [J]. New Generation Comput, 2002, 20(3): 263 - 277.

作者简介:

刘文斌 男,博士后,副教授,2004 年获华中科技大学博士学位,目前在温州大学计算机科学与工程学院工作. 研究领域为 DNA 计算、神经网络、遗传算法及生物信息学. E-mail: wbliu @mail. hust. edu. cn