DNA 计算中的单模板编码方法改进研究

王向红[1,2,3],刘文斌[3],朱翔鸥[3],章林溪[2]
(1. 温州职业技术学院,浙江温州 325035;2. 浙江大学物理系,浙江杭州 310027;
3. 温州大学物理与电子工程学院,浙江温州 325035)

摘 要: 如何避免各种不期望的杂交是 DNA 计算以及微阵列技术中的一个关键问题. 为了得到稳定可靠的杂交,必须探索一种可靠的、鲁棒性的编码方法. 单模板编码方法是 Arita 提出的另一种模板编码方法,它能够保证编码间的移位距离约为 l/3. 其缺点是仅仅使用众多满足条件模板中的一个,因而编码数量有限. 本文对单模板编码方法作了进一步的研究,提出来了另外一种模板框的结构,在基本保持移位距离约为 l/3 的情况下,将单模板方法扩展为多模板方法. 这一研究大大提高了该方法的应用规模.

关键词: DNA 计算; 编码方法; 模板; 纠错码
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2009) 12-2720-05

Improving the Single Template Method in DNA Computing

WANG Xiang-hong[1,2,3], LIU Wen-bin[2], ZHU Xiang-ou[2], ZHANG Lin-xi[2]
(1. *Wenzhou Vocational and Technical Collage, Wenzhou, Zhejiang* 325035, *China*;
2. *Department of Physics, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*;
3. *College of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou, Zhejiang* 325035, *China*)

Abstract: How to avoid the various undesired hybridizations is a crucial problem in DNA based computing and other microarray applications. In order to achieve reliable hybridization, we should explore reliable and robust encoding methods. The single template method proposed by Arita can achieve a promising shift distance with l/3 between DNA strands. However, the limited codes produced by it cann't meet the requirement of any practical application because only one possible template is employed. We extend it to multiple templates case so that the final codes can be linearly increased with the number of the templates while still keeping a shift distance close to l/3. Thus, the improved method can be applied to larger applications.

Key words: DNA computing; encoding method; template, error correcting code

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 *b*、*c*、*d*、*e*),并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.

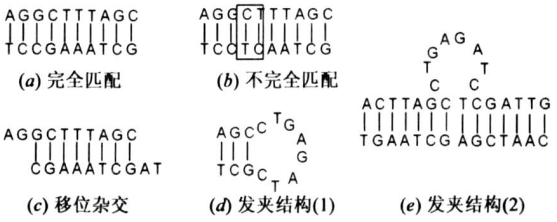

图1 DNA的几种可能的杂交形式

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的唯一识别,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

收稿日期:2008-10-09;修回日期:2009-04-20
基金项目:国家自然科学基金(No. 60403002,60970065);浙江省自然科学基金(No. Y1080227);温州市科技计划项目(No. H20080041)

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 等提出模板编码方法[9],刘文斌等对优化模板编码方法做了进一步的研究[10~12].

2002 年,Arita 提出了一个单模板编码的思想,他将编码问题约束定义为[13]:设计一个长度为 l 的编码集合 S,使得对于其中的编码 $s_1, s_2, s_3 \in S$,满足:s_1, s_1^{rc} 和 $s_2 s_3$ 中的任意长度为 l 的子序列的汉明距离大于等于 d. 显然,该条件使得编码方法要考虑到编码在实际应用中得各种线性组合情况. 应该说这正是 DNA 计算中编码问题的困难所在. Arita 提出的单模板方法具有很强得鲁棒性,其缺点是:满足条件的模板数量很多,但实际应用中只选择了其中一个,因而编码数量的有限性严重影响了其应用规模. 本文将简要介绍单模板编码方法的思想及其性质,然后将其进一步扩展为多模板的情况,从而扩大编码的数量.

2　单模板编码方法及其性质[13]

下面我们介绍几个概念,对于一个二进制串 $x = x_1 x_2 \cdots x_l$,$|x|$ 表示 x 的长度,即其中包含 0、1 的个数 l;$x^r = x_l x_{l-1} \cdots x_1$ 表示 x 的反串;x^c 表示 x 的补串,即将其中的 01 相互反过来;$x^{\rightarrow i} = x_{i+1} x_{i+2} \cdots x_l x_1 \cdots x_i$ 表示将 x 的前 i 个字符依次移到其尾部;$w(x)$ 表示序列 x 的权重,即其中 1 的个数;$\langle x \rangle = x_2 \cdots x_{l-1}$ 表示将序列 x 中第一个和最后一个字母去掉后所得的序列. 显然,$\langle xx \rangle$ 中包含 $l-1$ 个子串 $x^{\rightarrow i}$,$[x^r x^r]$ 中包含 $l-1$ 个子串 $(x^r)^{\rightarrow i}$ $(1 \leq i \leq l-1)$. 此外,为了衡量一个二进制串 x 自身的移位距离,Arita 引入了一个串 x 的质(Mass) $\| x \|$ 的概念.

定义 1

$$\| x \| = \min(H(x, x^r), H_M(x, \langle xx \rangle), H_M(x, \langle x^r x^r \rangle))$$

单模板编码的思想是给定:

(1) 一个二进制模板序列 $\| x \| = d$;

(2) 一个汉明距离大于等于 d 的二进制纠错码集合 E;

可以生成一个 DNA 的编码集合 $S = \{ x \cdot e \mid e \in E \}$,其生成规则和 Frutos 的模板编码方法一样[9]. 显然,对于任意 DNA 编码 $s, s', s'' \in S$,一定满足:

(1) $H(s, s^{rc}) \geq d$

(2) $H_M(s, \langle ss' \rangle) \geq d$, $H_M(s, s's'') \geq d$, $H_M(s, s's'') \geq d$

(3) $H_M(s^{rc}, \langle ss' \rangle) \geq d$, $H_M(s^{rc}, s's'') \geq d$, $H_M(s^{rc}, s's'') \geq d$

上面的(2)、(3)解决 DNA 计算中出现的各种移位杂交现象,因此具有很强的鲁棒性. 由于在传统的电子计算机中对于纠错码的研究已经很成熟,并有很多现成的算法,在此我们将不再作过多讨论.

定义 2　长度为 l 的二进制串 x 和 x^r,通过移位操作 $\rightarrow i$ 可以生成一个循环码集合

$$C_x = \{ x^{\rightarrow i} \mid 0 \leq i \leq l-1 \} \cup \{ (x^r)^{\rightarrow i} \mid 0 \leq i \leq l-1 \}$$

由质(mass)的定义,$\| x \|$ 是 x 与循环码集合 C_x 中的其它循环码 $x' \in C_x (x' \neq x)$ 的最小汉明距离. 下面列出单模板的一些性质:

性质 1　对于二进制串 x,$(x^{\rightarrow i})^{\rightarrow j} = (x)^{\rightarrow i+j}$ $(1 \leq i, j \leq l$ 为正整数)

性质 2　对长度为 l 的二进制串 x,$x^{\rightarrow i+l} = x^{\rightarrow i}$ $(1 \leq i \leq l$ 为正整数)

性质 3　对二进制串 x,$(x^r)^{\rightarrow i} = (x^{\rightarrow -i})^r$ $(1 \leq i \leq l$ 为正整数)

性质 4　对二进制串 x,$H(x^{\rightarrow i}, x^{\rightarrow j}) = H(x, x^{\rightarrow j-i})$ $(1 \leq i, j \leq l$ 为正整数)

性质 5　对二进制串 x,$H(x^{\rightarrow i}, x^{\rightarrow j}) = H(x^{\rightarrow i+k}, x^{\rightarrow j+k})$ $(1 \leq i, j, k \leq l$ 为正整数)

性质 6　对二进制串 x,有 $\| x \| = \| x^r \| = \| x^{\rightarrow i} \| = \| (x^r)^{\rightarrow j} \|$ $(i, j$ 为正整数).

由此性质,$\| x \|$ 为循环码集合 C_x 中的最小汉明距离.

性质 7　对长度为 l 的任意二进制串 x,当 $\| x \| = 0$ 时,$|C_x| < 2l$;否则 $|C_x| = 2l$

性质 8　对于任一 $x' \in C_x$,有 $C_x = C_{x'}$

性质 9　对二进制串 x,$\| x \| = \| x^c \|$.

该性质说明二进制串 x 的质 $\| x \|$ 以权值 $w(x) = \lfloor x/2 \rfloor$ (l 为偶数)或 $w(x) = \lfloor l/2 \rfloor$ 和 $w(x) = \lfloor l/2 \rfloor + 1$ (l 为奇数)的列对称.

性质 10　对二进制串 x,$\| x \|$ 总为偶数

性质 11　在 l 维超立方体中,权重为 w 的列中至少由 $C_l^w / 2l$ 不同的循环码集合.

性质 12　对任一长度为 l 的二进制串 x,$\| x \| \leq \lfloor l/2 \rfloor$

该性质当且仅当 $w = l/2$ 时,等式成立. 由于 w 从 0 到 $\lfloor l/2 \rfloor$ 变化时 $\| x \|$ 逐渐增大,权值 $w(x) = \lfloor l/2 \rfloor$ 附近的二进制串的 $\| x \|$ 最大.

性质 13　长度为 l 的二进制串 x,$\| xx \| = 2 \| x \|$

该性质说明可以通过二进制串 x 构造长度和 都加倍的模板.

3　单模板编码方法的改进

定义 3　二个长度为 l 二进制串 x, y 的质为

$$\| x, y \| = \min\{ (H(x, y), H(x, y^r), H_M(x, \langle yy \rangle), H_M(x, \langle y^r y^r \rangle) \}$$

性质 14

$\| x, y \| = \| y, x \|$

证明

(1) 显然 $H(x, y) = H(y, x)\ H(x, y^r) = H(y, x^r)$

(2) $H_M(x, <yy>)$

$= \min\{ H(x, y^{\to 1}), \cdots, H(x, y^{\to i}), \cdots, H(x, y^{\to l-1}) \}$

$= \min\{ H(x^{\to l-1}, y), \cdots, H(x^{\to l-i}, y), \cdots, H(x^{\to (l-1)}, y) \}$

$= \min\{ H(x^{\to l-1}, y), \cdots, H(x^{\to l-i}, y), \cdots, H(x^{\to 1}, y) \}$

$= H_M(y, <xx>)$

(3) 同理 $H_M(x, <y^r y^r>) = H_M(y, <x^r x^r>)$

最后,由式(1)～(3)及 $\| x, y \| = \| y, x \|$ 得证.

定义 4 长度为 l 的二进制串集合 X 的质 $\| X \| = \min\limits_{1 \le i < j \le |X|} (\| x_i, x_j \|)$

显然,当 $x_i, x_j \in C_x$,有 $\| X \| = 0$.

性质 15 给定模板集合 $X = \{ x_1, x_2, \cdots x_k \}$ 和汉明距离为 $\| X \|$ 的纠错码集合 E,可以生成编码集合 $S = S_1 \cup \cdots \cup S_k$,其中 $S_i = \{ x_i \cdot e \mid e \in E, 1 \le i \le k \}$. 对编码 $s \in S_i$ 和编码 $s', s'', s''' \in S_j (1 \le i, j \le k)$ 有:

(1) 同一模板产生的编码之间满足 $H_M(s''', s's'') \ge \| X \|$ $H_M((s^{rc}), s's'') \ge \| X \|$;

(2) 任意二个模板产生的编码之间满足 $H_M(s, s's'') \ge \| X \|$ $H_M(s^{rc}, s's'') \ge \| X \|$.

定义 5 二进制模板集合 $X = \{ x_1, x_2, \cdots x_k \}$ 的模板框

$$F = x_{i_1} x_{i_1} g_1 x_{i_2} x_{i_2} g_2 \cdots g_{k-1} x_{i_k} x_{i_k}$$

其中 $i_1, \cdots, i_k$ 为 $1, \cdots, k$ 的一个排列,$G = \{ g_1, \cdots g_{k-1} \}$ 为模板之间起间隔作用的二进制串集合.

定义 6

$\| F \| = \min\limits_{1 \le j \le k} \{ H_M(x_{i_j}, (x_{i_1} x_{i_1} g_1 x_{i_2} x_{i_2} \cdots g_{j-1} x_{i_j} \rangle), H_M(x_{i_j}, \langle x_{i_j} g_j \cdots g_{k-1} x_{i_k} x_{i_k})), H_M(x_{i_j}^r, (x_{i_1} x_{i_1} g_1 x_{i_2} x_{i_2} \cdots g_{j-1} x_{i_j} \rangle, H_M(x_{j_i}^r, \langle x_{i_j} g_j \cdots g_{k-1} x_{i_k} x_{i_k})) \}$

定义 7 给定一 DNA 序列 $W = W_1 G_1 W_2 G_2 \cdots G_{k-1} W_k$,当且仅当

(1) $W_j (1 \le j \le k)$ 是由模板 x_{i_j} 生成的所有编码的一个任意的线性组合;

(2) $G_j = g_j \cdot e_0 (1 \le j \le k-1)$,其中 e_0 为任一和间隔串 $G = \{ g_1, \cdots g_{k-1} \}$ 长度相同的二进制串;

则称 W 由模板框 F 编码.

性质 16 给定模板集合 X 及其任一模板框 F, $\| F \| \le \| X \|$

性质 17 设 $S = \{ x \cdot e \mid x \in X, e \in E \}$, W 按 X 的模板框 F 编码,则对任一编码 $s \in S$,和 W 中任一长度为 l 的子串 $s' \ne s$,条件 $H(s, s') \ge F$ 和 $H(s^{rc}, s') \ge \| F \|$ 同时成立. (证明略)

4 算法及计算结果

文献[13]的计算结果表明,单模板方法的最大移位距离 $d = \| \cdot \|_{\max}$ 大于等于 $l/3$,这是目前所有编码方法得到的最好结果. 从杂交的热力学过程来说,在适当的物理化学条件下,此约束能够保证错误杂交的可能性很小. 由于具有最大移位距离 $d = \| \cdot \|_{\max}$ 的模板并不唯一,在长度 $l = 6$～30 时,其数量范围为 $|T| = 1$～8072. 因此,如果仅使用一个模板显然是一种编码的浪费. 但是,在多模板情况(即模板框中)不一定能够仍然保持最大移位距离. 因此,我们先搜索 $V = \| X \| \approx \| \cdot \|_{\max}$ 的最大模板集合 $X \subseteq T$,然后再从模板集合 X 中搜索具有 $d'' = \| F \|_{\max}$ 的模板框.

因此,计算过程为:

(1) 搜索集合 T;

(2) 搜索最大模板集合 $X \subseteq T$;

(3) 求 X 的最佳模板框 F;

显然,集合 T 可以采用穷举搜索方法直接得到. 而求最大模板集合 $X \subseteq T$ 则对应于超立方体中的一个最大独立集问题,这是一个 NP - 完全问题,因此,我们采用蚁群优化搜索算法,经过多次搜索求得最大模板集合 X. 由定义 7 和性质 17 可以看出,模板框 F 提供了一个将由各模板 $x_i (1 \le i \le k)$ 生成的 DNA 编码序列分组使用的方法,并且能够保证移位距离测度不小于 $\| F \|$. 因此,如何得到具有最大值 $d'' = \| F \|_{\max}$ 的模板框就是问题的关键.

4.1 模板框的算法

在模板框 F 的定义中加入间隔串 $G = \{ g_1, \cdots g_{k-1} \}$ 主要目的是隔开各模板串之间的直接影响,提高模板框的 $\| F \|$. 显然模板框的 $\| F \|$ 与模板序列 $X = \{ x_1, x_2, \cdots x_k \}$、间隔串 $G = \{ g_1, \cdots g_{k-1} \}$ 及它们的相互位置有关. 当间隔串的长度 l_g 小于模板串的长度 l 时,显然不能完全消除之间的相互影响. 例如,对于 $H_M(x_i, x_p g_j x_q)$,其中包含一些长度为 l 的子串 $* g_j *$,其前后有部分子串分别来自 x_p, x_q. 如果直接采用穷举搜索算法,计算复杂度将很大. 当 $l_g = l$ 时,只要分别计算 $H_M(x_i, x_p g_j)$ 和 $H_M(x_i, g_j x_q)$ 即可. 此时,模板串在模板框中的相对位置对 $\| F \|$ 影响很小. 下面我们给出二种搜索最佳模板框 F 的方法:

(1) $g_1 = \cdots = g_{k-1} = g$, $\| F \|$ 仅由下面 $2k$ 个长度为 $2l - 1$ 的子串

$$< x_{i_1} g), (g x_{i_2} >$$
$$\cdots, \cdots$$
$$< x_{i_{k-1}} g), (g x_{i_k} >$$

和各模板 x_{i_j} 及其反 $x_{i_j}^r (1 \le j \le k)$ 的移位距离 $H_M(\cdot, \cdot)$ 决

定.

(2) $g_1, \cdots, g_{k-1}$不固定,$\|F\|$将由下面 $2k$ 个长度为 $2l-1$ 的子串

$$< x_{i_1} g_1) , (g_1 x_{i_2} >$$
$$\cdots , \cdots$$
$$< x_{i_{k-1}} g_{k-1}) , (g_{k-1} x_{i_k} >$$

和各模板 x_j 及其反 $x_j^r (1 \leq j \leq k)$ 的移位距离 $H_M(\cdot, \cdot)$ 决定.

由于使用相同的间隔串,因而方法(1)计算简单,其缺点是灵活性差. 方法(2)间隔串可以不同,因而对前后的模板串的适应性强,有利于提高模板框的$\|F\|$. 此外,由于与模板的顺序关系不大,因而,可以分别搜索各个间隔串使得$\|F\|$达到最大值. 此外,由$\|F\|$的定义可以看出,要保证$\|F\|$具有最大值,间隔串 g_i 与各模板 x_j 及其反 x_j^r之间的汉明距离必须足够大. 考虑热力学方面的因素及单个模板的$\|\cdot\|$,模板集合的 $w(x) \approx \lfloor l/2 \rfloor$,所以我们可以从权值小的二进制串$(0)_l$开始搜索.

4.2 计算结果

根据上节的算法,表 1 列出了我们得到的计算结果,可以看出:

(1) 对于单个模板的最大$\|x\| \geq \lfloor l/3 \rfloor$(仅 $l=24$ 时 $\|x\|=8$);

(2) $d' = \|X\|$除 $l=26,30$ 时,$d' = \|X\| = d-2$ 外,其余均 $d' = d$,且约等于 $l/3$;

(3) $d'' = \|F\|$基本都在$(1/4 \sim 1/3)\, l$ 之间;

(4) 具有最大$\|X\|$的模板集合 X 的大小都小于 10,且$|X| \ll |T|$. 只有当 $l=24,30$ 时,最大的$|X|=9$,从这个角度,编码长度取 $l=24,30$ 可能最合适,能够得到的编码的集合将是原单模板编码的 9 倍.

表 1　编码长度 $l=16 \sim 30$ 的计算结果

l	w	d	d'	d''	$\|T\|$	$\|X\|$	l	w	d	d'	d''	$\|T\|$	$\|X\|$
16	7	6	6	4	8	4	24	12	8	8	6	8072	9
17	8	6	6	5	16	2	25	11	10	10	8	20	2
18	9	6	6	5	195	5	26	13	10	8	7	1176	5
19	9	8	8	8	1	1	27	13	10	10	8	2151	3
20	10	8	8	6	24	7	28	14	12	10	9	26	2
21	10	8	8	6	15	4	29	14	12	12	9	2	2
22	11	8	8	7	40	3	30	14	12	10	7	1092	9
23	11	10	10	10	1	1							

最后,尽管我们得到的最大模板框的 $d'' = \|F\| = (1/4 \sim 1/3)\, l$,但是当一个 DNA 序列 W 按照模板框 F 的结构进行编码时,任一编码 $s \in S$ 和 W 中任一长度为 l 的子串 $s' \neq s$ 之间的汉明距离 $H(s, s')$和 $H(s^{rc}, s')$的分布具有如下二个特点(如图 2 所示):

(1) 在间隔序列 G_i 及其二边 $l-1$ 的区域(总长为 $3l-2$),$H(*, *) \geq d''$;

(2) 在编码序列区域 W_{i+1}中间总长为 $(m-2)\, l+2$ 的区域,$H(*, *) \geq d'$(其中纠错码$|E| = m$);

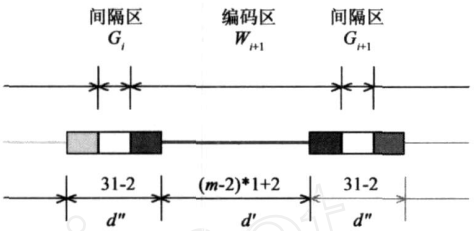

图2　$H(s,s')$和$H(s^{rc},s')$的分布示意图

由于编码远大于间隔区域,模板框的结构能够保证在编码区的绝大部分区域保证 $H(*, *) \geq d'$,即移位距离大于等于 $l/3$. 当$|E| = m = 100$ 时,在编码区中大约有 98 %的区域的 $H(*, *) \geq d'$. 由于只有在间隔区中 $H(*, *) \geq d''$,因此,我们能够知道可能发生错误杂交的区域,从而可以采取一些防范的措施. 如可以将间隔区二边的二个编码序列不用作编码任何信息,这样就可以保证编码区的移位距离足够大,提高 DNA 计算的可靠性.

5 应用

基于 DNA 的大规模数据库具有高密度和易于实现相关搜索(Association Search)的优点. Duke 大学的 Reif 等正在从事这方面的研究[14]. 他们提出了一种基于 DNA 分子的多进制数据表示方法. 表示数据库数据的所有的 DNA 都固定在一个由氨基团覆盖的磁珠上,如图 3 所示,每个数据是由 4 位信息位组成,每个信息位有 k 个不同的元素.

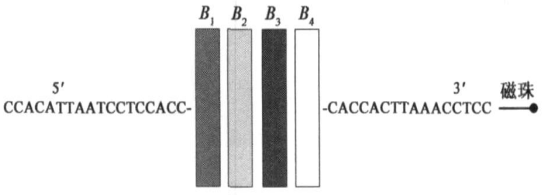

图3　DNA数据库的编码示意图

显然,为了构建大规模的数据库,信息位的数目及每个信息位中的元素也必须随之增加. 他们目前已经建立了一个规模为 127 的数据库,使用的编码列在表 2,分为 7 个编码块,每个编码块包含 12 个长度为 5 的编码序列. 从每个块中取一个编码线性连接后就现成一个库序列. 显然,在这种编码方法中第一个序列 AAACC 可能在库序列中多次出现,而使用本章提出的改进的单模板方法就可以明显的改善这种情况.

由于在改进的单模板方法中,编码的使用实际上是以其对应的模板来分子使用的,而每个模板能够生成的编码数量由纠错码的大小$|E|$ 决定的. 假定模板集合的大小 $k = |X|$,$|E|$ 编码能编码的数据库的容量为 $f(|E|$

|) ,那么我们可以采用二种方法来编码的数据库:(1)由一个模板生成的编码单独编码一个子数据库,那么,数据库的容量 $N = k * f(|E|)$;(2)按照模板框的结构直接生成数据库,则数据库的容量 $N = \left(f(|E|)\right)^k$. 显然,由性质 3.2,前者的可靠性高于后者,而数据库的容量远远小于后者.

表 2 Reif 使用的编码表

Block1	Block2	Block3	Block4	Block5	Block6	Block7
AAACC	AATCC	AACCA	AACCT	AATCC	ACACA	AACCA
ACCAA	ACACT	ACATC	ACCTA	ACAAC	ACCAT	ACACT
ACTCT	ATCAC	ACCAT	ACTAC	ACCTT	ATCTC	ACTTC
ATCTC	CAAAC	ATTCC	ATACC	ATCCA	CACAA	ATCAC
CATAC	CCATA	CACTT	CAAAC	CAACT	CATTC	CATAC
CCTTA	CCTAT	CATAC	CCATT	CCATA	CCATT	CCAAA
CTACA	CTCTT	CCAAA	CTCAA	CTCAT	CTAAC	CTCTA
CTCAT	CTTCA	CTACT	CTTCT	CTTTC	CTTCA	CTTCT
TACCA	TACCT	TAACC	TATCC	TACTC	TAACC	TACTC
TCAAC	TCCAA	TCCTA	TCACA	TCCAA	TCCTA	TCCAT
TCCTT	TCTTC	TCTCT	TCCAT	TCTCT	TCTAC	TCTCA
TTTCC	TTACC	TTCAC	TTCTC	TTACC	TTCCT	TTACC

6 小结

单模板编码方法是 Arita 于 2002 年提出来的另一种具有良好鲁棒性的模板编码方法,由于仅仅使用一个模板导致其编码数量有限. 本文对单模板编码方法的数学理论作了进一步的研究,并将我们提出的模板框结构和这种单模板方法结合起来. 结果表明:在基本保持原单模板方法的优良移位距离性质 $\|\cdot\|$ 的前提下,编码数量明显扩大(主要依赖于模板框的大小 k = |X|).

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]. San Fransisco: Morgan Kaufmann, 1997. 472 - 487.

[3] M GARZON et al. Encoding genome for DNA computing[A]. The Third DIMACS Workshop on DNA-based Computing, American Mathematical Society[C]. San Fransisco: Morgan Kaufmann, 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]. Washington: American Mathematical Society, 1996. 122 - 127.

[5] U FELDKAMP, W BANZHAF, H RAUHE. A DNA sequence compile[A]. Proceedings of 6th DIMACS Workshop on DNA Based Computers[C]. Heidelberg: Springer, 2001. 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]. Washington: American Mathematical Society, 1997. 7 - 12.

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

[8] N CHELYAPOV, L M ADLEMAN. Solution of a satisfiability problem on a Gel-based DNA computer[A]. The 6th International Workshop on DNA-Based Computers[C]. Heidelberg: Springer, 2001. 27 - 42.

[9] A FRUTOS, Q LIU, A THIEL, A SANNER, A CONDON, L SMITH, R CORN. Demonstration of a word design strategy for DNA computing on surface[J]. Nucleic Acids Research, 1997, 25(23): 4748 - 4757.

[10] W LIU, S WANG, L GAO, J XU. DNA sequence design based on template strategy[J]. Journal of Chemical Information and Computer Sciences, 2003, 43(6): 2014 - 2018.

[11] 刘文斌,朱翔鸥,王向红,张强,马润年. 一种优化 DNA 计算模板性能的新方法[J]. 电子与信息学报, 2008, 30(5): 1131 - 1135.
LIU Wen-bin, ZHU Xiang-ou, WANG Wiang-hong, ZHANG Qiang, MA Run-nian. A new method to optimize the template set in DNA computing[J]. Journal of Electronics & Information Technology, 2008, 30(5): 1131 - 1135. (in Chinese)

[12] 刘文斌,陈丽春,白宝钢,朱翔鸥,张强,马润年. DNA 计算中的模板框优化方法研究[J]. 电子学报, 2007, 35(8): 1490 - 1494.
LIU Wen-bin, CHEN Li-chun, BAI Bao-gang, ZHU Xiang-ou, ZHANG Qiang, MA Run-nian. Research on optimizing the template frame in DNA computing[J]. Acta Electronica Sinica, 2007, 35(8): 1490 - 1494. (in Chinese)

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

[14] J H REIF, et al. Experimental construction of very large scale DNA databases with associative search capability[A]. The 7th International Workshop on DNA-Based Computers: DNA Computing[C]. Heidelberg: Springer, 2002. 231 - 247.

作者简介:

王向红 女,1964 年 6 月生,硕士,教授,硕士生导师. 研究领域为高分子物理、蛋白质折叠、生物信息、DNA 计算等,已发表 40 多篇学术论文. E-mail: wangxh@wzu.edu.cn

刘文斌 男,1969 年 6 月生,博士,副教授,硕士生导师. 2004 年获华中科技大学博士学位,研究领域为生物信息、DNA 计算、神经网络、遗传算法. 已发表这方面的学术论文 40 余篇.

E-mail: wbliu6910@126.com