DNA 计算中编码序列的优化设计方案*

崔光照[1]，张勋才[1]，王延峰[1,2]

(1. 郑州轻工业学院 电气信息工程学院，河南 郑州 450002；2. 华中科技大学 控制科学与工程系，湖北 武汉 430074)

摘 要：提出了一种优化设计方案。该方案的各项评价指标均优于根据以往文献提供的方法所能得到的最好结果。尤其是所提出的海明距离测度方法，进一步保证了特异性杂交产生的自由能远大于非特异性杂交所产生的自由能，便于进行 DNA 编码序列的设计与选择，为可控的 DNA 计算提供可靠有效的编码序列。

关键词：DNA 计算；编码序列；热力学参数；物理特性

中图分类号：TP301.5 **文献标志码**：A **文章编号**：1001-3695(2007)07-0195-04

Optimized Encoding Sequences Approach for DNA Computing

CUI Guang-zhao[1], ZHANG Xun-cai[1], WANG Yan-feng[1,2]

(1. *School of Electrical Information Engineering, Zhengzhou University of Light Industry, Zhengzhou Henan* 450002, *China*; 2. *Dept. of Control Science & Engineering, Huazhong University Science & Technology, Wuhan Hubei* 430074, *China*)

Abstract: Encoding sequences design is a crucial problem in DNA computing. Not only an optimized approach for constructing a set of encoding sequences was developed by using random generate and real-time filter algorithm, but several evaluation terms were proposed to compare with other existing sequence design systems. The comparison results show the performance of this approach outperforms the existing DNA sequence design systems, especially in preventing sequence self-complementary from forming secondary structure and keeping the uniform melting temperature among sequences. It is convenient for user to design and select proper encoding sequences in silicon, and can provide reliable and effective encoding sequences for controllable DNA computing.

Key words: DNA computing; encoding sequences; thermodynamic parameters; biophysics properties

DNA(Deoxyribo Nucleic Acid)计算中的序列编码问题可简单地定义为：系统地将一个算法问题的实例映射为特殊的 DNA 分子序列。这样的 DNA 分子序列应能够确保随后进行的生化反应不出现任何错误，而且反应产物中需包含有足够多的、稳定可靠的、能被成功提取的原始算例的解[1]。只有满足上述两个条件的 DNA 序列才能称为好的 DNA 编码序列。由此可见，编码问题几乎涵盖了 DNA 计算研究领域内所有的重点和难点。事实上，为了实现理想的生化反应以及解的检测，DNA 计算的每一成功算例均离不开设计或选择合适的 DNA 序列。所以，自 DNA 计算诞生以来编码问题就一直是该研究领域的核心问题之一。随着研究的进一步深入，其重要性愈加突显，因为它在一定程度上决定着 DNA 计算模式的未来。

为使生化反应可控，文献[2～13]分别从不同的角度提出了一些约束条件，如相似性[2～10]、汉明距离测度[2～7]、3′端汉明距离测度[2,6]、反相互补汉明距离测度[6,8,9]、二级结构[2～4,8～11]、连续性[2,3]、自由能测度[8,9,11～13]、解链温度[2,4,5,11]、GC 含量[2～8,10]、生物实验方法[12,13]、DNA 碱基约束[4]以及特殊子序列[3,5,10]等。但普遍局限于细节，缺乏整体统筹，实用性和通用性不强。本文在综合考虑上述约束条件的基础上，根据各约束条件的内在特点及其之间的相互制约关系，采用整体优化的思想，利用随机产生—实时过滤算法(RGRF)构建了一种优化设计方案。

1 总体设计策略

定义 1 $X=\{x_1, x_2, \cdots, x_n\}$。它是由满足所有约束条件的长度为 m 的 n 个 DNA 编码序列所组成的集合，用 x_i^g 表示 X 集合中任一序列 $x_i(x_i^1 x_i^2 \cdots x_i^m)$(从 5′端算起)的第 g 个碱基($1\leq i\leq n, 1\leq g\leq m$)。令 $\overline{X}=\{\overline{x_1}, \overline{x_2}, \cdots, \overline{x_n}\}$，$\overline{x_i}$是 x_i 的补序列。

总体设计思想如下：

(1)以目标对象是单链还是双链将各约束条件划分为两类，即(Ⅰ)类和(Ⅱ)类。以各约束条件的计算时间复杂度和强弱程度依次排序。

(2)随机产生第一条设定长度的单链 DNA 序列及其补序列，判断其是否满足(Ⅰ)类约束条件。若满足，则将这两个序列分别放入集合 X 和 $\overline{X}$ 中；否则，重新生成，直至产生第一条满足(Ⅰ)类约束条件的 DNA 序列及其补序列。

(3)重复(2)，产生满足(Ⅰ)类约束条件的第二条 DNA 序列及其补序列，与集合中已存在的 DNA 序列作比较(比较规则随后详述)，判断其是否满足(Ⅱ)类约束条件。若满足，将这

收稿日期：2006-03-27；修返日期：2006-07-04 基金项目：国家自然科学基金资助项目(60573190)；河南省自然科学基金资助项目(021105090, 511011600)

作者简介：崔光照(1957-)，男，河南洛宁人，教授，博士，研究方向为智能计算、数字系统可靠性(cgzh@zzuli.edu.cn)；张勋才(1981-)，男，河南郸城人，博士研究生，研究方向为 DNA 计算、算法设计与优化；王延峰(1973-)，男，河南南阳人，讲师，博士研究生，研究方向为 DNA 计算等.

两个序列放入集合 X 和 $\overline{X}$ 中; 否则, 重复(3), 直至产生同时满足(Ⅰ)和(Ⅱ)类约束条件的第二条 DNA 序列及其补序列。

重复执行(3), 直至在集合中产生满足计算所需数目(假定为 n)的单链 DNA 序列及其补序列。

2 设计标准

2.1 (Ⅰ)类约束条件

(Ⅰ)类约束条件是以单链 DNA 序列的生物物理属性和热力学参数为基础建立起来的, 目的是尽量过滤掉可能形成二级结构和造成非特异性杂交的 DNA 单链序列。所考虑的因素包括 DNA 序列的热力学参数(T_m 和 ΔG)、组分(*GC* 含量)和构象(包括一级和二级结构)。

2.1.1 热力学约束

解链温度(T_m)是决定反应效率的重要参数。影响 T_m 的因素主要包括外部条件(温度和正离子的浓度)和内部结构(*GC* 含量和分子类型)。为有效地降低不完全匹配双链产生的概率, 要求参加反应的 DNA 分子的 T_m 基本一致。

目前普遍采用最邻近热力学方法来近似计算 T_m。Borer 等人最早利用最邻近理论研究 DNA 双链稳定性问题[14], 经 SantaLucia 改进后, T_m 的计算公式为

$$T_m = \Delta H/[\ \Delta S + R \times \ln(C/4)\ ] - 273.15 \tag{1}$$

其中, R 为摩尔气体常数; C 为核酸浓度; ΔS 和 ΔH 分别表示在一定温度下各碱基之间的熵变与焓变。在计算 ΔS 和 ΔH 时, 可根据寡核苷酸片段首尾碱基的不同加上不同的首尾校正值, 盐离子浓度的校正体现在 ΔS 中。37 ℃下的寡核苷酸的 ΔS、ΔH 和 ΔG 的计算参数详见文献[15]。显然, 时间复杂度为 $O(m)$。

自由能变(ΔG^{sh}) ΔG 是指进行一个反应所需要(释放)的能量, 整个反应的能量度量由 $|\Delta G|$ 表征。结构形成时放出的能量越多($|\Delta G|$ 越大), 结构越稳定。这里用 ΔG^{sh} 表示随机产生的 DNA 序列与其补序列发生特异性杂交反应时所释放的能量。为满足反应中解的稳定性, 要求依次产生的 DNA 单链序列的 $|\Delta G^{sh}|$ 尽量大。根据文献[15]中所列参数及方法, 可以很容易求出任意一条单链 DNA 序列的 $|\Delta G^{sh}|$, 保留 $|\Delta G^{sh}| \geq |\Delta G^{sh}|^*$ ($|\Delta G^{sh}|^*$ 为阈值)的序列。显然, 时间复杂度为 $O(m)$。

2.1.2 组分约束

组分约束即 *GC* 含量约束。*GC* 含量不仅对保持序列的化学性质的稳定性很重要, 还能够确保参加反应的 DNA 分子的 T_m 基本一致, 从而有效地降低不完全匹配双链产生的概率。

任意一个给定序列 $x_i(x_i^1 x_i^2 x_i^3 \cdots x_i^m)$ 的 *GC* 含量可用式(2)求得。一般要求 $f_{GC} \in [40, 60]$。

$$f_{GC} = 100/m \sum_{g,h=1}^{m} (x_i^g/G + x_i^h/C),\ (x_i^g = G, x_i^h = C) \tag{2}$$

显然, 该算法的时间复杂度为 $O(m)$。

2.1.3 构象约束

DNA 单链构象与双链一样包括一级结构和空间结构, 所以可仅依据单链来建立约束条件。

(1) 一级结构。为满足序列的可扩增性, 一个给定长度的 DNA 单链序列。一级结构应满足: ①序列的 3′端不应超过三个连续的 *G* 或 *C*, 以避免引物在 *GC* 富集序列区错误引发; ②序列中不能连续出现相同碱基, 否则 DNA 分子结构就会不稳定, 容易配对错位, 杂交反应不能得以较好控制; ③序列中不能出现重复结构。任意一个给定序列 $x_i(x_i^1 x_i^2 x_i^3 \cdots x_i^m)$ 应满足

$$\begin{cases} x_i^{m-2} x_i^{m-1} x_i^m \notin \{GGG, CCC\} \\ x_i^g x_i^{g+1} \cdots x_i^{g+p-1} \notin \overbrace{b \cdots b}^{p},\ b \in \{A, T, C, G\} \\ x_i^{g'} x_i^{g'+1} \cdots x_i^{g'+p'-1} \neq x_i^{g'+p'} x_i^{g'+p'+1} \cdots x_i^{g'+2p'-1} \end{cases} \tag{3}$$

显然, 该算法的时间复杂度为 $O(m)$。

(2) 二级结构。对于 DNA 计算中长度较短的 DNA 单链序列, 在有一定长度的反相互补序列时, 唯一的空间结构只能是发夹结构。虽然发夹结构也可应用于 DNA 计算, 并且可能使 DNA 计算实现自动控制。但为满足序列的可扩增性, 要求集合中的所有元素只包含一级结构。至此, 问题实际上已转换为随机产生的 DNA 单链序列的二级结构预测。目前单链 DNA 编码序列的二级结构预测的算法主要有两类: ①基于矩阵运算的动态规划算法[16]; ②基于茎区组合的预测算法[17]。算法①虽然预测结果较准确, 但计算烦琐[18]; 算法②准确度不高。

这两类算法虽有区别, 但其处理问题的共有前提为: 假定序列中存在着连续互补碱基个数大于 s 的两个子序列。为节约计算时间, 可赋予序列自身连续互补碱基不大于 s 的约束条件, 这样就可以最大限度地避免 DNA 序列二级结构的产生。

任意给定一个序列 $x_i = x_i^1 x_i^2 x_i^3 \cdots x_i^m$, 如果用 x_i^g 和 x_i^h(从 5′端算起)表示 x_i 的第 g 和第 h 个碱基, 用 $x_i^g \cdot x_i^h$ 表示碱基 x_i^g 与 x_i^h 形成匹配碱基对, 则 DNA 分子二级结构定义如下[19]:

定义 2 x_i 的二级结构为顺序碱基对 $x_i^g \cdot x_i^h$ 的集合($1 \leq g < h \leq m$), 满足 $h - g > 3$; $\{x_i^g \cdot x_i^h\} \cap \{x^{g'}_i \cdot x^{h'}_i\} = \Phi$。

定义 3 定义茎为若干堆积的碱基对, 记为 $S(g, h, s)$。g、h 为 x_i 从 5′端算起的该堆积碱基的起始匹配碱基对 $x_i^g \cdot x_i^h$, s 为茎长。若干连续的非配对的碱基称为环, 记为 R; r 为环长。

根据定义 2、3, 可通过以下算法来避免单链空间结构的产生。具体可简述为: 检索给定序列中是否存在 $S(g, h, s)$, $\{1 \leq g \leq m - 2s - r + 1;\ g + s + r \leq h \leq m\}$, ($s$ 和 r 为形成发夹结构所需的最小茎长和环长), 若存在, 即认为其可能存在二级结构, 排除该序列。

对于上述算法, 若序列长度为 m, 茎长为 s, 则耗时为 $O(m^2 \cdot s)$。从分子生物学角度来讲, s 是有上界的, 所以检索长为 m 的序列茎区的时间复杂度为 $O(m^2)$。

2.1.4 (I) 类约束条件排序

RGRF 的过滤排序原则为: 先考虑计算时间复杂度, 较低的计算时间复杂度排在前。在相同的情况下, 根据约束条件的强弱程度排序。

根据(Ⅰ)类中各约束条件的计算时间复杂度分析可知, 二级结构的约束应排在最后。其余几项的计算时间复杂度均为 $O(m)$, 这时需考虑各约束条件的强弱程度。通过对随机产生的 10 组 10^6 条 DNA 序列的统计分析表明, T_m 值和 ΔG^{sh} 与 *GC* 含量和一级结构相比, 约束性更强。(Ⅰ)类中各约束条件的过滤顺序依次为热力学约束、组分约束和构象约束, 如前所述。

2.2 (Ⅱ)类约束条件

(Ⅱ)类约束条件是以序列间的生物物理属性和热力学参

数为基础建立起来的, 目的是滤除可能造成非特异性杂交的 DNA 序列。其考虑的因素有序列间的汉明距离测度和 ΔG。

2.2.1　汉明距离测度

定义 4　假定 $x_i(x_i = x_i^1 x_i^2 \cdots x_i^m)$ 为随机新产生的一条 DNA 序列, $\overline{x_i}$ 是其补序列。令 $P = \{x_i \cup \overline{x_i} \cup X \cup \overline{X}\}$, X 与 $\overline{X}$ 分别表示先前已产生的满足编码定义要求的序列及其补序列集合。用 $x_u x_v$ 表示序列 $x_u(x_u^1 x_u^2 \cdots x_u^m)$ 和 $x_v(x_v^1 x_v^2 \cdots x_v^m)$ 的顺序连接; 用 $\langle x_u x_v x_w \rangle$ 表示 x_u 和 $x_v x_w$ 或 $x_u x_v$ 和 x_w 这样一种序列组合。

当计算汉明距离时, 需考虑以下序列的连接组合:

$$\begin{cases} \langle x_u x_v \underline{x_w} \rangle & 1 \le u, v, w \le n \\ \langle x_u \underline{x_v} x_w \rangle & 1 \le u, v, w \le n, u \ne w \\ \langle x_u \underline{x_v x_w} \rangle & 1 \le u, v, w \le n, u \ne v \\ \langle x_u \underline{x_v x_w} \rangle & 1 \le u, v, w \le n(u \ne v) \wedge (u \ne w) \end{cases} \tag{4}$$

其中, x_u、x_v、$x_w \in \{x_i \cup X\}$, $\underline{x_u}$, $\underline{x_v}$, $\underline{x_w} \in \{\overline{x_i} \cup \overline{X}\}$; $\langle x_u x_v x_w \rangle$ 表示 x_u、x_v、x_w 中的任一序列与另两个序列顺序连接的所有可能组合。用 y_l(长为 $2m$) 表示满足上述条件的任一顺序连接序列。

当随机新产生一个序列时, 需要对满足上述条件的所有可能的序列组合进行移位比对, 并要求通过移位比对后所得到的汉明距离尽量大。

定义 5　令 x_i 和 y_l 为两个反向平行序列, 即如果序列 x_i 为 5′→3′ 向, 则 y_l 为 3′→5′ 向。类似地, 用 x_i^g 表示从 x_i 的 5′ 端开始的第 g 个碱基($1 \le g \le m$); 用 y_l^h 表示从 y_l 的 3′ 端开始的第 h 个碱基($1 \le h \le 2m$); 用 $x_i^g \cdot y_l^h$ 表示序列 x_i 的第 g 个碱基和序列 y_l 的第 h 个碱基形成匹配碱基对。

汉明距离测度准则为

$$|y_l, x_i|: = \min_{0 < d < m-1} H(y_l, \sigma^d x_i) \tag{5}$$

这里, $H(*, *)$ 定义为汉明距离; d 定义为移动碱基个数。

具体算法可简要描述为: ①分别任意地定义两个待比对序列 y_l 为目标序列, x_i 为比对序列; ②比对序列与目标序列移位比对, 每移位一次得到一个汉明距离, 记为 $H_{l,i}^{\sigma^d}$; ③依次递推, 最后取其汉明距离最小值 $H(y_l, x_i) = \min\{H_{l,i}^{\sigma^1}, H_{l,i}^{\sigma^2}, \cdots, H_{l,i}^{\sigma^{m-1}}\}$; ④若 $H(x_i, x_j) \ge H^*$ (H^* 为阈值), 则保留; ⑤变换目标序列, 重复② ~④, 直至遍历式(4) 中所有的 y_l 序列组合。该算法的计算时间复杂度为 $O(n^2 m^2)$。

2.2.2　热力学约束

$|\Delta G_{\min}^{\text{non-sh}}|$值。为满足 DNA 编码序列定义中的约束条件①, 要求任一随机产生的 DNA 序列 x_i 与 y_l 发生非特异性杂交反应时所产生的最小自由能 $|\Delta G_{\min}^{\text{non-sh}}|$ 尽量小。

x_i 与 y_l 发生非特异性杂交反应时所产生的最小自由能 $|\Delta G_{\min}^{\text{non-sh}}|$ 的计算公式如下:

$$\Delta G_{\min}^{\text{non-sh}} = \min_{\substack{1 \le i \le m \\ 1 \le l \le 2m}} \{D(x_i^1, x_i^g, y_l^l, y_l^h) + V(x_i^g, y_l^h)\} \tag{6}$$

在式(6) 中, $D(x_i^1, x_i^g, y_l^1, y_l^h)$ 表示 x_i 的子序列 $x_i^1 x_i^2 \cdots x_i^g$ 与邻近 $x_i^g \cdot y_l^h$ 的 y_l 子序列 $y_l^1 y_l^2 \cdots y_l^h$ 之间的悬挂末端或自由末端的自由能; $V(x_i^g, y_l^h)$ 表示 x_i 的子序列 $x_i^g x_i^{g+1} \cdots x_i^m$ 与邻近 $x_i^g \cdot y_l^h$ 的 y_l 的子序列 $y_l^h y_l^{h+1} \cdots y_l^{2m}$ 之间的自由能的最小值:

$$V(x_i^g, y_l^h) = \min\left\{\begin{array}{l} D(y_l^{2m}, y_l^h, x_i^m, x_i^g), VBI(x_i^g, y_l^h), \\ ex_i(x_i^g, x_i^{g+1}, y_l^h, y_l^{h+1}) + V(v_i^{g+1}, y_l^{h+1}) \end{array}\right\} \tag{7}$$

其中, $D(y_l^{2m}, y_l^h, x_i^m, x_i^g)$ 表示邻近的 $x_i^g \cdot y_l^h$ 包含 y_l 的子序列 $y_l^{2m} y_l^{2m-1} \cdots y_l^h$ 和 x_i 的子序列 $x_i^m x_i^{m-1} \cdots x_i^g$ 的悬挂或自由末端的自由能; $VBI(x_i^g, y_l^h)$ 表示邻近的 $x_i^g \cdot y_l^h$ 形成一个凸环或内环所需的最小自由能; $ex_i(x_i^g, x_i^{g+1}, y_l^h, y_l^{h+1})$ 表示 x_i 的子序列 $x_i^g x_i^{g+1}$ 和 y_l 的子序列 $y_l^h y_l^{h+1}$ 之间形成堆积碱基对的自由能; $ex_i(x_i^g, x_i^{g+1}, y_l^h, y_l^{h+1}) + V(x_i^{g+1}, y_l^{h+1})$ 之和表示趋向于形成堆积碱基对的情况。其中, $V(x_i^{g+1}, y_l^{h+1})$ 可递归求出, 由下式求得

$$VBI(x_i^g, y_l^h) = \min_{\substack{g < g', h < h' \\ g' - g + h' - h > 2}} eL(x_i^g, x_i^{g'}, y_l^h, y_l^{h'}) + V(x_i^{g'}, y_l^{h'}) \tag{8}$$

在式(8) 中, $eL(x_i^g, x_i^{g'}, x_l^h, y_l^{h'})$ 表示的是邻近两个匹配碱基对 $x_i^g \cdot y_l^h$ 和 $x_i^{g'} \cdot y_l^{h'}$ 的 x_i 的子序列 $x_i^g x_i^{g+1} \cdots x_i^{g'}$ 和 y_l 的子序列 $y_l^h y_l^{h+1} y_l^{h'}$ 之间的环的自由能。

$\Delta G_{\min}^{\text{non-sh}}$可根据式(6) ~(8) 采用动态规划方法递归求出。在计算最小自由能的过程中, 所需参数的定义及来源如下: 堆积碱基对定义为($x_i^g \cdot y_l^h$, $x_i^{g+1} \cdot y_l^{h+1}$), 其自由能计算参数来源于文献[20]; 悬挂末端定义为(x_i^1, $x_i^2 \cdot y_l^1$), (y_l^1, $x_i^1 \cdot y_l^2$), ($x_i^{m-1} \cdot y_l^{2m}$, x_i^m) 或($x_i^m \cdot y_l^{2m-1}$, y_l^{2m}), 其自由能计算参数来源于文献[21]。

自由末端定义为 x_i 的子序列 $x_i^1 x_i^2 \cdots x_i^g$ 与邻近 $x_i^g \cdot y_l^h$ 的 y_l 的子序列 $y_l^1 y_l^2 \cdots y_l^h$ 之间或 x_i 的子序列 $x_i^g x_i^{g+1} \cdots x_i^m$ 与邻近 $x_i^g \cdot y_l^h$ 的 y_l 的子序列 $y_l^h y_l^{h+1} \cdots y_l^{2m}$ 之间没有形成匹配碱基对, 其自由能计算参数来源于文献[22]; 单凸环定义为($x_i^g \cdot y_l^h$, $x_i^{g+2} \cdot y_l^{h+1}$) 或者($x_i^g \cdot y_l^h$, $x_i^{g+1} \cdot y_l^{h+2}$), 其自由能计算参数来源于文献[20]; 其他类型的环(包括内环和多凸环) 的自由能的计算参数来源于文献[22, 23]。其中, 内环可定义为($x_i^g \cdot y_l^h$, $x_i^{g+u} \cdot y_l^{h+v}$) $\wedge$ ($u, v \ge 2$); 多凸环可定义为$\{(x_i^{g'} \cdot y_l^h, x_i^{g+u} \cdot y_l^{h+1}) \wedge (u \ge 3)\}$ 或$\{(x_i^g \cdot y_l^h, x_i^{g+1} \cdot y_l^{h+u}) \wedge (u \ge 3)\}$。

在 RGRF 中, 随机产生的首条 DNA 编码序列及其补序列无须进行 $\Delta G_{\min}^{\text{non-sh}}$过滤, 而随后随机产生的每一条 DNA 序列及其补序列, 都需要进行 $\Delta G_{\min}^{\text{non-sh}}$过滤, 依次计算出 $\Delta G_{\min}^{\text{non-sh}}$。如果 $|\Delta G_{\min}^{\text{non-sh}}| \le |\Delta G_{\min}^{\text{non-sh}}|^*$ ($|\Delta G_{\min}^{\text{non-sh}}|^*$ 为阈值), 则保留该序列及其补序列; 否则舍弃。其计算时间复杂度为 $O(n^2 m^3)$[24]。

3　结果与分析

这里针对上述约束条件建立了一套评价体系, 通过将 RGRF 随机产生的一组(10 个) DNA 编码序列与文献[24 ~26] 所提供的或随机产生的 DNA 编码序列作比较。比较时分两种情况考虑: ①DNA 编码序列的物理属性(包括连续性、二级结构和汉明距离测度); ②DNA 编码序列的热力学特性(包括 T_m 和 $|\Delta G_{\min}^{\text{non-sh}}|$)。其中, 连续性评价指标沿用文献[27], 汉明距离测度评价指标包括文献[27] 中的相似性和 H-measure。其余各评价指标如下:

(1) 二级结构的评价指标

$$f(x) = \sum_{s=S_{\min}}^{(m-R_{\min})/2} \sum_{r=R_{\min}}^{(m-2s)} \sum_{g}^{(m-2s-r)} [\sum_{g'=1}^{S} bp(x_i^{s+g-g'}, x_i^{s+g+r+g'}), s-1] \tag{9}$$

其中, m 为序列长度; r、s 为茎长和环长; $S_{\min}$、$R_{\min}$ 分别为设定的最小茎长和最小环长; x_i^g 表示 DNA 序列集合 X 中任一序列 x_i 从 5′ 端开始的第 g 个碱基; g 表示包含的碱基个数。且有

$$bp(a, b) = \begin{cases} 1 & a = \overline{b} \\ 0 & a \ne \overline{b} \end{cases} \tag{10}$$

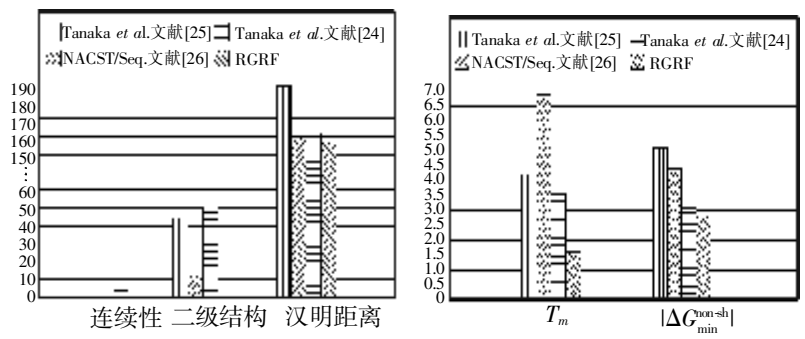
其中, $a, b \in \{A, T, C, G\}$; $\overline{a}$ 是 a 的互补碱基。

关于 DNA 编码序列的物理属性的比较结果如表 1 所示。

表中各评价指标经加和平均后如图 1(a) 所示。

表 1 文献[24 ~28] 与 RGRF 的 DNA 编码序列比较结果

序列	连续性	二级结构	汉明距离	T_m	$\Delta G0$	$\Delta G_{\min}^{\text{non-sh}}$
Tanaka *et al.* 文献[25]						
CGAGACATCGTGCATATCGT	0	6	170	64.595 8	- 27.12	- 12.33
TATAGCACGAGTGCGCGTAT	0	13	197	66.132 9	- 27.81	- 14.25
GATCTACGATCATGAGAGCG	0	13	191	61.986 3	- 25.78	- 14.12
TCTGTACTGCTGACTCGAGT	0	0	199	64.678 6	- 26.70	- 14.78
CGAGTAGTCACACGATGAGA	0	3	198	63.219 7	- 26.25	- 11.90
AGATGATCAGCAGCGACACT	0	0	196	66.058 5	- 27.41	- 14.52
TGTGCTCGTCTCTGCATACT	0	3	188	65.680 1	- 27.25	- 15.23
AGACGAGTCGTACAGTACAG	0	6	178	62.864 2	- 26.06	- 14.07
ATGTACGTGAGATGCAGCAG	0	0	177	64.809 8	- 27.00	- 12.67
ATCACTACTCGCTCGTCACT	0	0	180	65.075 7	- 27.02	- 13.41
NACST/ Seq. 文献[26]						
GTGACTTGAGGTAGGTAGGA	0	0	158	62.223 9	- 25.36	- 12.33
ATCATACTCCGGAGACTACC	0	0	156	62.141 3	- 25.42	- 13.26
CACGTCCTACTACCTTCAAC	0	0	178	61.939 7	- 25.55	- 15.89
ACACGCGTGCATATAGGCAA	0	3	154	67.266 2	- 28.22	- 15.36
AAGTCTGCACGGATTCCTGA	0	0	163	65.966 9	- 27.24	- 10.25
AGGCCGAAGTTGACGTAAGA	0	0	164	65.902 4	- 27.35	- 9.38
CGACACTTGAAGCACACCTT	0	0	166	65.459 4	- 27.25	- 10.06
TGGCGCTCTACCGTTGAATT	0	0	151	66.895 3	- 27.90	- 12.89
CTAGAAGGATAGGCGATACG	0	3	151	61.077 7	- 25.22	- 9.47
CTTGGTGCGTTCTGTGTACA	0	0	149	65.161 2	- 27.14	- 9.78
Tanaka *et al.* 文献[24]						
TTAGAAGTCCCACTGCACTG	9	6	159	64.030 2	- 26.34	- 7.16
TCCAGCCATTAGACTGACGA	0	3	189	65.091 9	- 26.82	- 7.84
GGCAATACGGTATGGAGGAT	0	3	166	63.764 3	- 26.23	- 8.01
GCTGCTTTTACCAGCACGTA	16	20	142	65.580 9	- 27.40	- 9.13
GCCTCCCAAATCATACGACA	18	0	158	64.524 6	- 26.68	- 8.67
ATCCGCGTTGTTCAAGCATC	0	0	158	66.254 8	- 27.83	- 8.18
AGAACTCCGTCACAAGACGA	0	10	170	65.518 0	- 27.18	- 9.03
GTATACCCCCTTCTTGTTGC	25	0	170	62.630 9	- 25.72	- 9.19
TCTCCCTTTCTCACATAGGC	18	3	155	63.260 4	- 25.88	- 6.70
GCTGTCCTTTAGGTGTTTCG	18	0	155	63.430 2	- 26.33	- 7.14
RGRF						
GGTTGTCCTGAGTGTTGTGT	0	0	145	64.814 8	- 26.73	- 6.36
TACACGTAGCCTGAACCATC	0	0	152	63.848	- 26.40	- 8.21
AGAGTCAGCAGCGTAGAGAT	0	0	173	64.842 3	- 26.77	- 7.51
TACAGTCGGTTCGGTTATGG	0	0	152	63.838 4	- 26.44	- 8.17
GCGGAAGTAATCGGAAGTGA	0	0	148	64.460 0	- 26.81	- 6.83
CGTCTAGGCCGGATCAATAT	0	0	161	63.502 8	- 26.25	- 8.05
AGTCATTCTCCTGGCATTGC	0	0	168	65.127 4	- 26.90	- 7.59
GCACTGGTCAAGACATGGAT	0	0	150	64.761 8	- 26.71	- 7.27
ACGCCACAGTATATCATCGC	0	0	151	64.652 7	- 27.01	- 7.84
GCATCCTGTACTAACGTGGT	0	0	176	64.136 9	- 26.54	- 8.39



(a) 物理属性比较 (b) 热力学特性比较

图 1 RGRF 与文献[24~26]关于 DNA 编码序列的比较结果图

四种方法所产生的 DNA 编码序列的 *GC* 含量均为 50%, 所以图表中均未列出。从图中可以看出, 根据 RGRF 所产生的序列在防止编码序列自身形成二级结构方面要远远优于其他三种方法, 这意味着较低的漏解产生。在其他两项评价指标中, 虽然优势不十分明显, 但在 DNA 编码序列的物理属性方面, 总体评价效果要优于文献[24 ~26] 。

(2) T_m。一般要求参加反应的 DNA 分子的 T_m 基本一致, 这样可以有效地降低不完全匹配双链产生的概率。为此, 可建立评价指标为 $f(T_m) = \sqrt{\sum_{i=1}^{n}(T_m(x_i) - \overline{T_m})^2}$ (11)

其中, $T_m(x_i)$ 表示 DNA 序列集合 X(n 个序列) 中任一序列 x_i 的解链温度; $\overline{T_m}$ 表示这 m 个序列的解链温度的平均值。显然 $f(T_m)$ 越小代表序列彼此间的解链温度越接近。

(3) $|\Delta G_{\min}^{\text{non-sh}}|$ 如前所述, 要求集合中 DNA 编码序列间的 $|\Delta G_{\min}^{\text{non-sh}}|$ 在适当范围内尽量小。为此, 建立评价指标为

$$f(|\Delta G_{\min}^{\text{non-sh}}) = \sum_{i=1}^{n} \max(|\Delta G_{\min}^{\text{non-sh}}|_{x_i, y_l}) / |\Delta G^{\text{non-sh}}|_{x_i} \quad (12)$$

在计算 $f(|\Delta G_{\min}^{\text{non-sh}}|)$ 时, 只需考虑以下组合:

$$\begin{cases} \langle x_i x_v x_w \rangle & 0 \le i, v, w \le n \\ \langle x_i \underline{x_v} x_w \rangle & 0 \le i, v, w \le n, i \ne w \\ \langle x_i \underline{x_v x_w} \rangle & 0 \le i, v, w \le n, i \ne v \\ \langle x_i \underline{x_v x_w} \rangle & 0 \le i, v, w \le n(i \ne v) \wedge (i \ne w) \end{cases} \quad (13)$$

关于 DNA 编码序列的热力学属性的比较结果见表 1。将表中数据按照式(10) (11) 处理后, 结果如图 1(b) 所示。可以看出, RGRF 产生的序列在保持解链温度一致性方面要远远优于其他三种方法, 这意味着较低的不完全匹配双链产生的概率。在 $|\Delta G_{\min}^{\text{non-sh}}|$ 评价指标中, 与文献[24] 相当, 与文献[25, 26] 相比, 优势明显, 这意味着较低的非特异性杂交概率。综合上述分析, 总体的评价效果要优于文献[24 ~26] 。

4 结束语

在 DNA 计算中, 根据相关约束条件建立一套通用的 DNA 编码序列设计方案是一件非常困难的事情。一些约束条件之间存在着相互制约, 并且不同的约束条件所要求的生物实验条件和方法不同。由于生化反应受多种条件的影响, 在分子生物学中, 微小的条件变化可能不会影响整体实验效果及定性分析, 但对于 DNA 计算来说, 这种变化却可能是致命的。选择标准时应特别小心。可能的解决方案之一是建立一套多目标评价体系, 根据所要解决实际问题的需求, 通过赋予不同目标函数不同的权值来实现针对具体问题的 DNA 计算的编码序列的优化设计与选择。另外, 从本质上来讲, DNA 计算是以牺牲空间来换取时间的一种计算模式。当计算规模比较大时, 所需 DNA 编码序列的长度和数目随之增长。在进行编码序列设计与选择时, 计算规模就会随之增长, 为节省计算机处理时间, 还需针对具体约束条件(如 ΔG) 进行算法优化。

参考文献:

[1] ARZON M, DEATON R, NEATHERY P, *et al*. On the encoding problem for DNA computing: proc. of the 3rd DIMACS Workshop on DNA-based Computer[C] . [S. l.] : [s. n.] , 1997: 230-237.

[2] TANAKA F, NAKATSUGAWA M, *et al*. Developing support system for sequence design in DNA computing: proc. of the 7th Int. Workshop DNA-based Computer[C] . [S. l.] : [s. n.] , 2001: 340-349.

[3] FRUTOS A G, LIU Q, *et al*. Demonstration of a word design strategy for DNA computing on surfaces[J] . Nucleic Acids Research, 1997, 25(23) : 4748-4757.

[4] FAULHAMMER D, CUKRAS A R, *et al*. Molecular computation: RNA solutions to chess problems: proc. of the National Academy of Sciences[C] . [S. l.] : [s. n.] , 2000: 1385-1389.

[5] ARITA M, KOBAYASHI S. DNA sequence design using templates [J] . New Generation Computer, 2002, 20: 263-277.

[6] ARITA M, NISHIKAWA A, *et al*. Improving sequence design for DNA computing: proc. of Genetic Evol. Comput. Conf. (GECCO) [C] . [S. l.] : [s. n.] , 2000: 875-882.

[7] TUPLAN D C, HOOSE H, *et al*. Stochastic local search algorithms for DNA word design: proc. of the 8th Int. Workshop DNA Based Computer[C] . London: Springer-Verlag, 2002: 229-241.

[8] ANDRONESCU M, DEES D L, *et al*. Algorithms for testing that DNA word designs avoid unwanted secondary structure: proc. of the 8th Int. Workshop DNA-based Computer[C] . [S. l.] : [s. n.] , 2002: 182-195.

(下转第 201 页)

保留突变点信息。

3.3　信噪比分析

下面从信噪比方面对以上算法进行比较,信噪比为

$$SNR = 10\lg[\sum_{i=1}^{M} x^2(i)] / \{\sum_{i=1}^{M}[\hat{x}(i) - x(i)]^2\} \quad (16)$$

其中, M 为信号长度; i 为信号点的位置; $x(i)$ 为第 i 点原始信号值; $\hat{x}(i)$ 表示信号处理后的值。

表 2 给出了电压暂降信号和电压崩溃信号经各算法去噪后的信噪比。分析表中数据可知,本文算法的信噪比大于五点均值滤波和空间自适应滤波的信噪比,说明本文算法在去噪方面也远远优于均值滤波和空间自适应滤波。

表 2　信噪比比较表

信噪比	电压暂降信号	电压崩溃信号
叠加噪声之后信噪比(SNR)	27.117 957	25.919 068
均值滤波去噪后的信噪比(SNR′)	31.512 245	31.798 306
空间自适应滤波去噪后的信噪比(SNR′)	23.082 567	28.727 947
本文算法去噪后的信噪比(SNR′)	32.630 892	32.207 760

通过上述实验结果、波形和实验仿真数据表、信噪比比较表的分析以及其他电能质量信号的实验仿真验证,本文所采用的算法,克服了空间自适应滤波中大量噪声对参数的影响和均值滤波无法准确保留突变点信息的缺点,可以良好地应用于电能质量检测信号去噪和突变点信息保留。

4　结束语

本文提出的基于维纳滤波的电能质量检测去噪算法,继承了均值滤波能很好地去除大量高斯噪声的优点,结合维纳滤波的最小均方误差准则及阈值处理技术,同时实现了大面积噪声的去噪和原始信号突变点保留。因而对改善电能质量、降低损耗具有重要作用,具有很好的推广应用价值。

参考文献:

[1] 肖湘宁. 电能质量分析与控制[M]. 北京: 中国电力出版社, 2004.

[2] MOHARAM M G, GRANN E B, POMMENT D A, *et al*. Stable implementation of the rigorous coupled-wave analysis of for surface-relief grating: enhanced transmittance matrix approach [J]. J. Opt. Soc. Am. A., 1995, 12(5): 1077-1086.

[3] VASEGHI S V. Advanced digital signal processing and noise reduction[M]. Chichester, England: Wiley, 2000: 178-202.

[4] PORTILLA J, STRELA V, WAINWRIGHT M J, *et al*. Adaptive Wiener de-noising using a Gaussian scale mixture model in the wavelet domain: proc. of the 8th Int. Conf. Image Processing[C]. [S. l.]: [s. n.], 2001: 327-331.

[5] CHANG S G, YU B, VETTERLI M. Spatially adaptive wavelet threshold with context modeling for image de-noising [J]. IEEE Trans. on Image Processing, 2000, 9(9): 1522-1531.

[6] CHANG S G, VETTERLI M. Spatial adaptive wavelet threshold for image de-noising: proc. of ICIP '97[C]. Washington DC: [s. n.], 1997: 374-377.

[7] 赵艳明, 全子一. 一种有效的小波: Wiener 滤波去噪算法[J]. 北京邮电大学学报, 2004, 27(4): 41-45.

[8] 熊四昌, 陈茂军, 胡金华. 一种基于小波变换图像去噪方法[J]. 计算机与数字工程, 2005, 33(4): 24-27.

[9] HILTON L, OGDEN T. Data analytic wavelet threshold selection in 22D signal de-noising[J]. IEEE Trans. on Signal Processing, 1997, 45(2): 496-500.

[10] 欧阳森, 宋政湘, 陈德桂, 等. 小波软阈值去噪技术在电能质量检测中的应用[J]. 电力系统自动化, 2002, 26(19): 56-60.

(上接第 198 页)

[9] HANG B T, SHIN S Y. Molecular algorithms for efficient and reliable DNA computing: proc. of Genetic Program[C]. [S. l.]: [s. n.], 1998: 735-742.

[10] FELDKAMP U, SAGHAFI S, *et al*. DNA sequence generator: a program for the construction of DNA sequences: proc. of the 7th Int. Workshop DNA-based Computer[C]. [S. l.]: [s. n.], 2001: 179-188.

[11] HARTEMINK A J, GIFFORD D K, *et al*. Automated constraint based nucleotide sequence selection for DNA computation: proc. of the 4th DIMACS Workshop DNA-based Computer[C]. [S. l.]: [s. n.], 1998: 227-235.

[12] DEATON R, CHEN J, *et al*. A software tool for generating noncrosshybridization libraries of DNA oligonucleotides: proc. of the 8th Int. Workshop DNA-based Computer[C]. [S. l.]: [s. n.], 2002: 252-261.

[13] DEATON R, CHEN J, *et al*. A PCR-based protocol for in vitro selection of noncrosshybridizing olgionucleotides: proc. of the 8th Int. Workshop DNA-based Computer[C]. [S. l.]: [s. n.], 2002: 196-204.

[14] BORER P N, DENGLER B, *et al*. Stability of ribonucleic acid double-stranded helices[J]. Journal of Molecular Biology, 1974, 86(4): 843-853.

[15] SANTALUCIA J Jr. An unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics [J]. Proc. of the National Academy Sciences, 1998, 95(4): 1460-1465.

[16] NUSSINOV R, JACOBSON A B. Fast algorithm for predicting the secondary structure of single strand RNA[J]. Proc. of the National Academy Sciences, 1980, 77(11): 6309-6313.

[17] BENEDETTI G. SANTIS P D, *et al*. A new method to find a set of energetically optimal RNA secondary structures[J]. Nucleic Acids Research, 1989, 17(13): 5149-5161.

[18] WATERMAN M S. Introduction to computational biology[M]. London: Chapman & Hall, 1995: 334-337.

[19] SANKOFF D. Simultaneous solution of the RNA folding. alignment and protosequence problems[J]. SIAM Journal on Appled Mathematics, 1985, 45(5): 810-825.

[20] TANAKA F, KAMEDA A, *et al*. Thermodynamic parameters based on a nearest-neighbor model for DNA sequences with a single-bulge loop [J]. Biochemistry, 2004, 43: 7143-7150.

[21] BOMMARITO S, PEYRET N, *et al*. Thermodynamic parameters for DNA sequences with dangling ends[J]. Nucleic Acids Research, 2000, 28: 1929-1934.

[22] ZUKER M. Mfold Web server for nucleic acid folding and hybridization prediction[J]. Nucleic Acids Research, 2003, 31(13): 3406-3415.

[23] PERITZ A E, KIERZEK R, *et al*. Thermodynamic study of internal loops in oligoribonucleotides: symmetric loops are more stable than asymmetric loops[J]. Biochemistry, 1991, 30(26): 6328-6436.

[24] TANAKA F, KAMEDA A, *et al*. Design of nucleic acid sequences for DNA computing based on a thermodynamic approach[J]. Nucleic Acids Resarch, 2005, 28(3): 903-911.

[25] TANAKA F, NAKATSUGAWA M, *et al*. Toward a general-purpose sequence design system in DNA computing: proc. of Congr. Evol. Comput. (CEC) [C]. USA: [s. n.], 2002: 73-78.

[26] SHIN S Y, KIM D M, *et al*. Evolutionary sequence generation for reliable DNA computing: proc. of Congr. Evol. Comput. (CEC) [C]. USA: [s. n.], 2002: 79-84.

[27] SHIN S Y, LEE I H, *et al*. Multi-objective evolutionary optimization of DNA sequences for reliable DNA computing[J]. IEEE Transactions on Evolutionary Computation, 2005, 9(2): 143-158.