基于部分字的 DNA 编码设计与分析 *

李 珍, 王淑栋, 李二艳
(山东科技大学 信息科学与工程学院, 山东 青岛 266510)

摘 要: DNA 编码问题是 DNA 计算中的第一步也是最重要的一步,是 DNA 计算中的一个基本问题。引入部分字与其洞的定义,研究了部分字的洞与沃森—克里克汉明距离的内在联系,得到沃森—克里克汉明距离与 DNA 编码的关系;通过分析不完全匹配部分字中洞的出现位置,对发生错误匹配的 DNA 码进行了优化。解决了 DNA 编码中除去洞分散分布在 DNA 双链中的不完全匹配问题,有效弥补了杂交过程中出现的假阳性的缺陷,为 DNA 编码的研究注入了活力。

关键词: DNA 计算; DNA 编码; 沃森—克里克汉明距离; 部分字
中图分类号: TP18 **文献标志码**: A **文章编号**: 1001-3695(2010)01-0086-03
doi:10.3969/j.issn.1001-3695.2010.01.025

Analysis and design of DNA encoding based on partial words

LI Zhen, WANG Shu-dong, LI Er-yan
(*College of Information Science & Engineering, Shandong University of Science & Technology, Qingdao Shandong* 266510, *China*)

Abstract: This paper introduced the definitions of partial word and its holes. Researched the relationship between the holes and Watson-Crick Hamming distance. And achieved the relationship between Watson-Crick Hamming distance and DNA encoding. Optimized the DNA code with mismatches by analyzing the hole positions which presented in the partial words. The mismatch problem had been solved except that the holes distribute in the DNA strands dispersedly in DNA encoding.
Key words: DNA computation; DNA encoding; Watson-Crick Hamming distance; partial words

0 引言

1994 年,Adleman[1] 首次利用分子生物技术解决了一个具有七个顶点的有向 Hamilton 路问题。到目前为止,DNA 计算已经取得了突飞猛进的发展[2~5],DNA 编码作为 DNA 计算的一个基本问题也取得了很大的进展。1996 年,Baum[6] 提出 DNA 码字间相似度的假设,并在此基础上给出增大 DNA 码字间相似度的编码方法;1997 年,Garzon 等人[7,8] 给出 DNA 编码的定义以及分析 DNA 杂交错误的防错编码策略;2000 年,Feldkamp 等人[9] 给出 DNA 码字间相似度的定义,得到了码字间相似度的形式评价准则;1997—2001 年,Frutos[10]、Arita[11] 和 Braich[12] 等人分别提出了模板编码方法、单模板编码方法和三字母表的编码策略;2003 年,Tulpan 等人[13] 提出了随机搜索优码的方法,得到了质量较好的 DNA 码,但其计算量非常庞大。

针对 DNA 计算中出现的错误杂交,Berstel 等人[14] 于 1998 年提出了部分字的概念;2002 年,Blanchet-Sadri 等人[15] 对部分字的性质进行了详细的讨论,得到部分字的周期性以及含有不超过一个洞的部分字的相容性结论;2003 年,Blanchet-Sadri[16,17] 在此基础上研究了含有多个洞的部分字的周期性、相容性以及无关性等重要性质。事实上,部分字的相容性、无关性等性质[18] 中就体现了 DNA 编码的基本限制条件,如部分字的相容性包含了 Hamming 距离。

1 基本定义及性质

部分字的概念主要是针对 DNA 序列的错误匹配提出的。例如图 1 中的两条链在位置 3,10 发生了错误匹配,把这些既不相同也不互补的错误匹配的位置看成没有定义的位置。图 2 中的两条链虽然在位置 3,4 也发生了错误匹配,但这些位置的碱基是相同的,把它们看成是有定义的位置。

图1 发生两处错误匹配的DNA双链(对应位置碱基既不相同也不互补)

图2 发生两处错误匹配的DNA双链(对应位置碱基相同)

为了叙述方便,下面引入部分字的定义。

定义 1 设 Σ 是一个字母表,部分函数 $f:\{0,1,\cdots,n-1\}\rightarrow\Sigma$ 是定义域中某些元素没有定义的函数。

定义 2 字母表 $\Sigma=\{A,T,C,G\}$ 上的部分字是由字母表 $\Sigma\cup\{o\}$ 中元素构成的长度为 n 的序列。

定义 3 部分字 w 中有定义的位置构成的集合称为 w 的定义域,记为 $D(w)$;w 中无定义的位置构成的集合称为 w 的洞集,记为 $H(w)$;w 中洞的个数用 $|H(w)|$ 表示。

由定义 2、3 知道,部分字的洞用"o"表示。

由定义 3 可知,任意一个部分字的定义域是惟一的。不同

收稿日期: 2009-04-21; **修回日期**: 2009-06-15 **基金项目**: 国家自然科学基金资助项目(60503002,30670540)
作者简介:李珍(1982-),女,山东淄博人,硕士研究生,主要研究方向为 DNA 计算(lizhen0202@yahoo.com.cn);王淑栋(1973-),女,山西运城人,副教授,博士,主要研究方向为 DNA 计算、图与组合最优化等;李二艳(1984-),女,河北保定人,硕士研究生,主要研究方向为 DNA 计算.

的定义域必定对应不同的部分字。如图 1 和 2 中两条 DNA 双链的上链虽然相同,但因其对应部分字定义域不同,分别为{0,1,3,4,5,6,7,8,10,11,12,13,14,15}和{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15},这两个部分字不同,分别为 AG*o*CAATGT*o*ACAGTC 和 AGTCAATGTGACAGTC。同一定义域可能对应不同的部分字,如图 1 中的 DNA 双链,其上链和下链分别对应部分字 AG*o*CAATGT*o*ACAGTC 和 TC*o*GTTACA*o*TGTCAG,它们的定义域相同,都是{0,1,3,4,5,6,7,8,10,11,12,13,14,15},但它们是两个不同的部分字。

设 w 是一个部分字,长度为 n,则 $H(w)=\{0,1,\cdots,n-1\}\setminus D(w)$。

定义 4[19]　设 $x=x_1x_2\cdots x_n, y=y_1y_2\cdots y_n\in\{A,T,C,G\}^*$。$x$、$y$ 的沃森—克里克汉明距离定义为 $H'(x,y)=\sum_{i=1}^{n}f(x_i,y_i)$。其中:$f(x_i,y_i)=\begin{cases}1 & x_i\neq y_i, x_i\neq\overline{y_i}\\ 0 & \text{否则}\end{cases}$,$i=1,2,\cdots,n$;$\overline{y_i}$ 是指在沃森—克里克汉明碱基互补原则下与 y_i 配对的碱基。

事实上,字母表 $\Sigma=\{A,T,C,G\}$ 上一个无洞的部分字是一个 DNA 序列。任意两个 DNA 序列,无论它们的沃森—克里克汉明距离是多少,都分别对应一个部分字,这两个部分字可能相同也可能不同。但无论对应相同的部分字还是不同的部分字,所对应部分字中洞的个数相同。例如,x = ATGCCAGTTGCATC,y = AGACCAGTTGAAGC,则 x 与 y 都对应部分字 A*oo*CCAGTTG*o*A*o*C;x = ATGCCAGTTGCATC, y = TACGCGTTAGTAGC,则 x 与 y 分别对应部分字 ATGC*oo*TTG*o*A*o*C 和 TACGC*oo*TAG*o*A*o*C,洞的个数都为 4。

由部分字洞的定义可知,两个等长的部分字,其洞的对应位置碱基既不相同也不互补。

定义 5[14]　任意两个等长的部分字 u、v,若 $D(u)\subset D(v)$,且对 $\forall i\in D(u)$ 有 $u(i)=v(i)$,则称 u 包含于 v,记做 $u\subset v$。

定义 6[14]　任意两个等长的部分字 u、v,若存在部分字 w,使得 $u\subset w$ 且 $v\subset w$,则称 u,v 是相容的,记做 $u\uparrow v$。

定义 7　任意两个等长的部分字 u、v,它们的最小字是指包含于 u、v 的定义域规模最小的字,记做 $u\wedge v$;它们的最大字是指包含 u、v 的定义域规模最大的字,记做 $u\vee v$。亦即 $D(u\wedge v)=D(u)\cap D(v)$,$D(u\vee v)=D(u)\cup D(v)$。

例如:u = A*o*TCAT*o*C,v = AT*o*C*o*T*o*C 是两个相容的部分字。$D(u)=\{0,2,3,4,5,7\}$,$D(v)=\{0,1,3,5,7\}$,$D(u\vee v)=D(u)\cup D(v)=\{0,1,2,3,4,5,7\}$,$D(u\wedge v)=D(u)\cap D(v)=\{0,3,5,7\}$。这样,$u\vee v$ = ATTCAT*o*C,$u\wedge v$ = A*oo*C*o*T*o*C。

定理 1　任意两个相容的部分字 u、v,有 $H'(u,v)=|H(u)|+|H(v)|-|H(u)\cap H(v)|$。

证明　由相容部分字洞的定义可知,它们洞所在位置的对应碱基既不相同也不互补,而且相容部分字中非洞位置对应的碱基相同。因此,当两个相容部分字的对应位置不都是洞时,$|H(u)\cap H(v)|=0$,两部分字中洞的数目之和与它们的沃森—克里克汉明距离相等,即 $H'(u,v)=|H(u)|+|H(v)|$;当部分字中出现对应位置都是洞时,须从沃森—克里克汉明距离中去掉重复的个数,即 $|H(u)\cap H(v)|$,故 $H'(u,v)=|H(u)|+|H(v)|-|H(u)\cap H(v)|$。综上两种情况,$H'(u,v)=|H(u)|+|H(v)|-|H(u)\cap H(v)|$。

定理 2　任意两个等长的部分字 u、v,有 $H'(u,v)=|H(u)|+|H(v)|-|H(u)\cap H(v)|$。

证明　设 u、v 是两个等长的部分字,当 u、v 相容时,由定理 1,$H'(u,v)=|H(u)|+|H(v)|-|H(u)\cap H(v)|$;当 u、v 不相容时,非洞位置对应碱基相同或互补,不产生沃森—克里克汉明距离。因此,$H'(u,v)=|H(u)|+|H(v)|-|H(u)\cap H(v)|$。综上可知,对任意两个等长的部分字 u、v,有 $H'(u,v)=|H(u)|+|H(v)|-|H(u)\cap H(v)|$。

定理 3　任意两个相容的部分字 u、v,它们的最小字中洞的个数与它们的沃森—克里克汉明距离相等。

证明　设 u、v 是任意两个相容的部分字,w 是它们的最小字。由定义 7 可得 $D(w)=D(u\wedge v)=D(u)\cap D(v)$。因此,$H(w)=\overline{D(w)}=\overline{D(u)\cap D(v)}=H(u)\cup H(v)$,故 $|H(w)|=|H(u)|+|H(v)|-|H(u)\cap H(v)|$。再由定理 1,$H'(u,v)=|H(u)|+|H(v)|-|H(u)\cap H(v)|$,于是 $|H(w)|=H'(u,v)$,即 u、v 的最小字中洞的个数与它们的沃森—克里克汉明距离相等。由 u、v 的任意性可知,结论成立。

2　DNA 杂交反应与 DNA 编码优化

定义 8[14]　部分字 w 的穿洞率定义为 $r(w)=\frac{|H(w)|}{|w|}$。其中:$|w|$ 表示 w 的长度。

穿洞率的大小与任意两个部分字的杂交情况有着密切的联系。因前面已经提到 DNA 序列是特殊的部分字,任意两个 DNA 序列都分别对应一个部分字,所以这两个 DNA 序列的穿洞率可通过计算它们对应的部分字的穿洞率得到。设 x、y 是任意两个 DNA 序列,且 $|x|=|y|=n$。当 $r(x)\geqslant 1/2$ 时有 $r(y)\geqslant 1/2$,此时 $|H(x)|\geqslant n/2$,$|H(y)|\geqslant n/2$。当 $|H(x)\cap H(y)|\leqslant n/2$ 时,由定理 2,$H'(x,y)=|H(x)|+|H(y)|-|H(x)\cap H(y)|\geqslant n/2+n/2-n/2=n/2$;当 $|H(x)\cap H(y)|>n/2$ 时,$H'(x,y)\geqslant|H(x)|\geqslant n/2$ 或 $H'(x,y)\geqslant|H(y)|\geqslant n/2$,即当 $r(x)\geqslant 1/2$ 且 $r(y)\geqslant 1/2$ 时,$H'(x,y)\geqslant n/2$。

研究文献[14]表明,对任意两个部分字 x、y,当 $r(x)$ 与 $r(y)$ 都大于等于 1/2 时,除洞外,即使其余对应位置的碱基互补,x 和 y 也不能杂交。因此,当 $H'(x,y)\geqslant n/2$ 时,x 和 y 不能杂交。当 $r(x)$ 与 $r(y)$ 之一小于 1/2 时,任意两个等长的 DNA 序列错误匹配情况如图 3 所示。

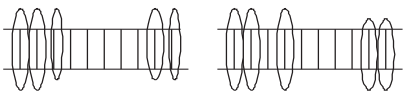


图3　洞集中分布的DNA双链两端

a)当洞分散分布在 DNA 双链中,其形式如图 4 所示。

b)当洞集中分布在 DNA 双链两端时,其形式如图 3 所示。

在部分字的左右两端分别选取下链和上链(因为上链从左到右方向为 5′→3′,而聚合酶链式反应的方向为 5′→3′),将包含洞的部分用外切酶切去,出现如图 5 所示的形式。对于这种形式的 DNA 分子,在试管中加入游离的核苷酸和聚合酶进行聚合酶链式反应,从而实现 DNA 分子的完全匹配;反复进行上述操作,直到所有这种形式的 DNA 分子全部实现完全匹配。将这些完整 DNA 双链固定在充满聚丙烯酰胺凝胶体的玻璃板上;再将玻璃板加热至 94℃并保持恒温,用 94℃缓冲液与玻璃

板充分均匀混合发生变性反应;然后用 94℃缓冲液冲洗玻璃板,冲洗后留在玻璃板上[20]的产物即为改良后的 DNA 单链分子,从而实现对编码的优化。

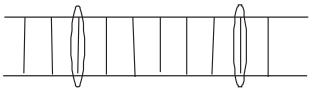


图4　洞分散分布在DNA双链中

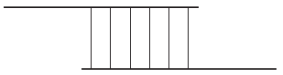


图5　图3切割之后的形式

c)当洞集中分布在 DNA 双链一端时,其形式如图 6 所示。

在出现洞的一端选取下链,将包含洞的部分用外切酶切去。同上述情况,加入游离的核苷酸和聚合酶进行聚合酶链式反应,从而实现 DNA 分子的完全匹配,其余操作与 b)类似。

d)当洞集中分布在 DNA 双链中间时,其形式如图 7 所示。

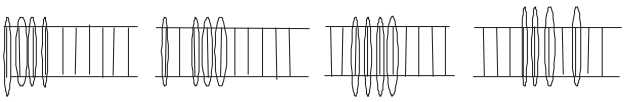
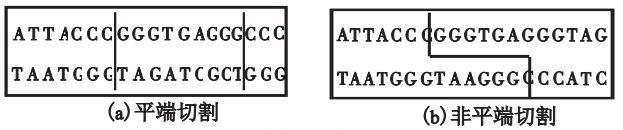


图6　洞集中分布在DNA双链一端　图7　洞集中分布在DNA双链中部

用限制性内切核酸酶对包含洞的部分进行切割,其切割可分为平端切割和非平端切割(这两种切割形式可用图 8 所示例子进行说明)。



(a)平端切割　(b)非平端切割

图8　示例

通过凝胶电泳实验将切割所得的 DNA 片段去除(在凝胶电泳中,由于较大的 DNA 片段会被构成凝胶的琼脂糖纤维网的障碍所阻滞,因而较小的线性片段比较大片段移动得快,并且完整双链比不完整双链迁移快,故完整双链最先到达阳极,凝胶中剩余的 DNA 片段可被去除。通过测定到达阳极的完整 DNA 双链长度,将长度小于给定长度的 DNA 双链去除),剩余的即为理想的 DNA 双链分子,然后执行 b)中所述操作,得到理想的 DNA 单链分子。

DNA 序列的不完全匹配包含以上四种情况,通过对其讨论可知,b)~d)这三种情况可以改良或去除,从而实现对 DNA 编码的优化。也就是说,对于 DNA 编码中的不完全匹配问题,除去情况 a)外,都可通过生物操作的方法加以解决。

3　结束语

DNA 计算最主要和核心的反应为 DNA 分子间的杂交反应,其效率和准确性直接影响到 DNA 计算的结果。DNA 计算过程中的错误杂交分为假阳性和假阴性。假阴性的产生主要是由反应条件及生化操作本身引起的,可通过控制生化反应条件来避免;假阳性主要包括不完全匹配、移位杂交、双链形成的发卡结构等。研究表明,合理的编码可以最大限度地避免假阳性的出现。本文即针对这一问题进行了较为深入的研究,解决了 DNA 编码中除去洞分散分布在 DNA 双链中的不完全匹配问题,有效弥补了杂交过程中出现的假阳性的缺陷。随着研究的继续深入,移位杂交与双链形成的发卡结构问题将成为笔者下一步的研究课题。

参考文献:

[1] ADLEMAN A M. Molecular computation of solution to combinatorial problems[J]. Science, 1994,266(5187):1021-1024.

[2] LIPTON R J. DNA solution of hard computational problems[J]. Science, 1995,268(5210): 542-545.

[3] BONEH D, DUNWORTH C, LIPTON R. Breaking DES using a molecular computer[C]//Proc of the 1st DIMACS Workshop on DNA Based Computers. Providence: American Mathematics Society, 1995: 37-65.

[4] GARZON M, DEATON R, NEATHERY P. On the encoding problem for DNA computing[C]//Proc of the 3rd DIMACS Workshop on DNA Based Computers. 1997:230-237.

[5] OUYANG Qi, KAPLAN P D, LIU Shu-mao. DNA solution of the maximal clique problem[J]. Science, 1997, 278(5337): 446-449.

[6] BAUM E B. DNA sequences useful for computation[C]//Proc of the 2nd Annual Meeting DNA-based Computers. Providence: American Mathematical Society, 1996.

[7] GARZON M, NEATHERY P, DEATON R, *et al*. A new metric for DNA computing[C]//Proc of the 2nd Annual Genetic Programming Conference. San Francisco: Morgan Kaufmann Publisher, 1997:472-487.

[8] GARZON M, DEATON R, NINO L F, *et al*. Genome encoding for DNA computing[C]//Proc of the 3rd DIMACS Workshop on DNA based Computing. Providence: American Mathematics Society, 1997: 230-273.

[9] FELDKAMP U, BANZHAF W, RAUHE H, *et al*. A DNA sequence complier[C]//Proc of the 6th DIMACS Workshop on DNA-based Computing. Providence: American Mathematics Society, 2000:253.

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

[11] ARITA M, KDBAYASHI S. The power of sequence design in DNA computing[C]//Proc of the 4th International Conference on Computational Intelligence and Multimedia Applications. 2001:163-167.

[12] BRAICH R S, JOHNSON C, ROTHEMUND P W, *et al*. Solution of a satisfy problem on a gel-based DNA computer[C]//Proc of the 6th International Workshop on DNA-based Computing. London: Springer-Verlag, 2001:27-42.

[13] TULPAN D C, HOLGER H, CONDON A. Stochastic local search algorithms for DNA word design[C]//Proc of the 8th International on DNA-based Computers. Berlin: Springer, 2003:229-241.

[14] BERSTEL J, BOASSON L. Partial words and a theorem of Fine and Wilf[J]. Theoretical Computer Science, 1999, 218(1): 135-141.

[15] BLANCHET-SADRI F, HEGSTROM A. Partial words and a theorem of Fine and Wilf revisited[J]. Theoretical Computer Science, 2002,270(1-2): 401-419.

[16] BLANCHET-SADRI F. Primitive partial words[J]. Discrete Applied Mathematics, 2003,148(3):195-213.

[17] BLANCHET-SADRI F. Codes, orderings and partial words[J]. Theoretical Computer Science, 2003,329(1-3):177-202.

[18] LEUPOLD P. Partial words for DNA coding[C]//Proc of the 10th International Workshop on DNA Computing. 2005:224-234.

[19] 王淑栋, 宋　. DNA Golay 码的设计与分析[J]. 电子学报, 2009,37(7):1542-1545.

[20] 周康, 同小军, 许进. 最优指派问题 DNA 算法[J]. 系统工程与电子技术, 2007,29(7):1183-1187.