doi:10.3969/j.issn.1673-4785.2011.02.015

用于DNA编码的部分字

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:寻找合理的DNA编码是DNA计算中一个基本的问题.因此要给出一种方法使得DNA序列不会产生不想要的结构,尤其是假阳性是解决此问题的关键.传统方法是要求码字间的Hamming距离足够大.因此考虑用部分字的方法来解决DNA编码问题,利用部分字的洞的定义及其性质得到了关于部分字的洞、Hamming距离和Watson-Crick Hamming距离的3个命题,通过部分字对DNA编码进行了优化,解决了DNA编码中的部分疑难问题.

关键词:DNA编码;部分字;洞;Hamming距离;Watson-Crick Hamming距离
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-4785(2011)02-0185-04

Partial words for DNA encoding

LI Zhen, WANG Shudong, LI Eryan
(College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266510, China)

Abstract: Finding a good DNA code is a very basic problem in DNA computation. A solution must be provided which ensures that the strands involved do not exhibit any undesired behavior, and especially that they should not form a false positive. The traditional solution requires the Hamming distance between the words to be big enough. The author proposed the idea of considering only partial words for the solution of the DNA encoding problem. To some degree they already include the Hamming distance in the definition of compatibility. Thus, they can be used to simultaneously guarantee a desired distance and other properties. In this paper, the definition of Hole and some properties of partial words were applied to achieve three propositions concerning Hole, Hamming distance, and Watson-Crick Hamming distance. The DNA code set was optimized by using the partial words. Thus some difficult problems were resolved in DNA encoding.

Keywords: DNA encoding; partial words; hole; Hamming distance; Watson-Crick Hamming distance

自从Adleman[1]首次利用分子生物技术解决了一个具有7个顶点的有向Hamilton路问题以来,DNA计算取得了突飞猛进的发展[2-5].DNA编码作为DNA计算的基本问题之一也取得了很大的进展[6-13].针对DNA编码中出现的错误杂交,Berstel和Boasson[14]于1998年提出了部分字的概念.2002年,Blanchet-Sadri[15]对部分字的性质做出了详细论证.2003年,Blanchet-Sadri[16-17]又对部分字的性质作了更深入的研究与探讨.部分字的一些性质[18]包含了DNA编码的一些限制条件,例如,部分字的相容性包含了Hamming距离,因此考虑用部分字的方法来解决DNA编码问题具有重要的意义.本文利用部分字的洞的定义及其性质得到了关于部分字的洞、Hamming距离和Watson-Crick Hamming距离的3个命题,通过对部分字的分析对DNA编码进行了优化.随着部分字越来越受到人们的关注,它们可能为DNA编码提供一种更为有效的工具.

收稿日期:2010-05-24.
基金项目:国家自然科学基金资助项目(60503002);中国博士后科学基金资助项目(20060400344).
通信作者:李珍.E-mail:topwayD202@163.com.

1 基本定义及性质

定义1[14] 字母表$\Sigma=\{A,T,C,G\}$上的部分字w是由部分函数$f:\{0,1,\cdots,n-1\}\to\Sigma$随机排列构成的DNA序列.$w(p)(=f(i),i\in\{0,1,\cdots,n-1\})$($p\in\{0,1,\cdots,n-1\}$)有定义的位置构成的集合称为$w$的定义域,记为$D(w)$,$H(w)=\{0,1,\cdots,n-1\}/D(w)$称为$w$的洞集合.

部分字主要是针对 DNA 序列的错误匹配提出的(见图 1).,把此位置的碱基看成是未知的,即没有定义的位置,据定义 1,这样的位置称为洞(而对于对应位置碱基相同的情况则不予考虑). 因此,可以将任意 2 条等长的 DNA 序列中相同位置碱基既不相同也不互补的位置称为洞.

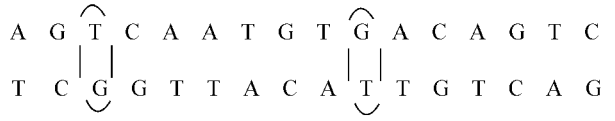

图 1 发生 2 处错误匹配的 DNA 双链

Fig. 1 DNA sequences with two mistakes

发生错误匹配的位置一般不能确定是哪个碱基发生了错误匹配.

定义 2[1] 设 $x=x_1x_2\cdots x_n, y=y_1y_2\cdots y_n\in\{A,T,C<G\}^*$. x、y 的 Watson-Crick Hamming 距离定义为

$$H'(x,y)=\sum_{i=1}^{n}f(x_i,y_i).$$

式中:$f(x_i,y_i)=\begin{cases}1, x_i\neq y_i, x_i\neq\bar{y}_i\\ 0, \text{其他}\end{cases}$, $i=1,2,\cdots,n$, $\bar{y}_i$ 是指在 Watson-Crick 碱基互补原则下与 y_i 配对的碱基.

定义 3[19] 设 $\Sigma=\{A,T,C,G\}$, $\Sigma_0=\{A,T\}$, $\Sigma_1=\{C,G\}$ 则显然有 $\Sigma_0\cap\Sigma_1=\varnothing$. $\Sigma_0\cup\Sigma_1=\Sigma$. Σ_0、Σ_1 定义为 2 个不同的类,则 x 和 y 中对应位置属于不同类的分量个数称为 Watson-Crick Hamming 距离,记为 $H'(x,y)$.

定义 4 字母表 $\Sigma\cup\{o\}$ 上的完全字 w_o 称为部分字 w 的伴随,如果

$$w_o[i]=\begin{cases}w[i], & i\in D(w);\\ o, & i\notin D(w)\text{ 且 }0\leqslant i\leqslant|w|.\end{cases}$$

为简便起见,用部分字的伴随来代替部分字. 例如,用"部分 *oAoT* 字"来代替"部分字的伴随 *oAoT*".

定义 5[14] 任意 2 个等长的部分字 u、v, $D(u)\subset D(v)$ 且由 $i\in D(u)$ 可得 $u(i)=v(i)$,则称 u 包含于 v,记作 $u\subset v$.

定义 6[14] 任意 2 个等长的部分字 u、v,若存在部分字 w 使 $u\subset w$ 且 $v\subset w$,则称 u、v 是相容的,记作 $u\uparrow v$.

由定义 3 ~5 得到如下定义.

定义 7 任意2个相容部分字 u、v、$u\vee v$ 表示包含 u、v 的最小的字,即 $D(u\vee v)=D(u)\cup D(v)$;$u\wedge v$ 表示包含于 u、v 的最大的字,即 $D(u\wedge v)=D(u)\cap D(v)$.

例 1 设 $u=AoTCAToC$, $v=AToCoToC$,则 $D(u)=\{0,2,3,4,5,7\}$, $D(v)=\{01,3,5,7\}$ 于是 $D(u\vee v)=D(u)\cup D(v)=\{0,1,2,3,4,5,7\}$, $D(u\wedge v)=D(u)\cap D(v)=\{0,3,5,7\}$, 从而 $u\vee v=ATTCAToC$, $u\wedge v=AooCoToC$.

由以上定义可以得到如下命题.

命题 1 任意两相容的部分字,其最大字的洞与 Watson-Crick Hamming 距离是一一对应的,即它们最大字的洞的个数与 Watson-Crick Hamming 距离相等.

证明 设任意两相容的部分字为 u、v,包含于 u、v 的最大字为 w,由定义 6 可得 $D(w)=D(u)\cap D(v)$,从而 $H(w)=\overline{D(w)}=\overline{D(u)\cap D(v)}=H(u)\cup H(v)$,故 $|H(w)|=|H(u)|+|H(v)|-|H(u)\cap H(v)|$,又 $H'(u,v)=|H(u)|+|H(v)|-|H(u)\cap H(v)|$,故二者相等,结论成立.

2 DNA 杂交反应与 DNA 编码优化

实验结果研究表明[14],2 条 DNA 序列是否杂交,不是取决于错误匹配的绝对数目,而是取决于错误匹配的频率(错误匹配所占的比率). 为此,引入穿洞率的定义.

定义 8 部分字 w 的穿洞率定义为 $r(w)=\dfrac{\text{Hole}(w)}{|w|}$,其中 Hole($w$)表示 w 中洞的个数.

穿洞率的大小与任意 2 个部分字的杂交情况有着密切的联系. 研究表明,当 $r(w)\geqslant\dfrac{1}{2}$时,任意 2 个部分字即使相应位置剩余所有碱基互补,则它们也不能杂交. 于是得到如下命题.

命题 2 任意两 DNA 序列 x、y, $|x|=|y|=n$, 当 $H'(x,y)\geqslant\dfrac{n}{2}$时,$x$ 与 y 不会发生杂交反应.

证明 当 $r(x)\geqslant\dfrac{1}{2}$ 且 $r(y)\geqslant\dfrac{1}{2}$ 时,有 $|H(x)|\geqslant\dfrac{n}{2}$, $|H(y)|\geqslant\dfrac{n}{2}$, $|H(x)\cap H(y)|\leqslant\dfrac{n}{2}$. 由命题 1, $H'(x,y)=|H(x\wedge y)|=|H(x)|+|H(y)|-|H(x)\cap H(y)|\geqslant\dfrac{n}{2}+\dfrac{n}{2}-\dfrac{n}{2}=\dfrac{n}{2}$,有以上论述知命题成立.

当 $r(w)<\dfrac{1}{2}$时,任意两等长 DNA 序列错误匹配情况如图 2 所示.

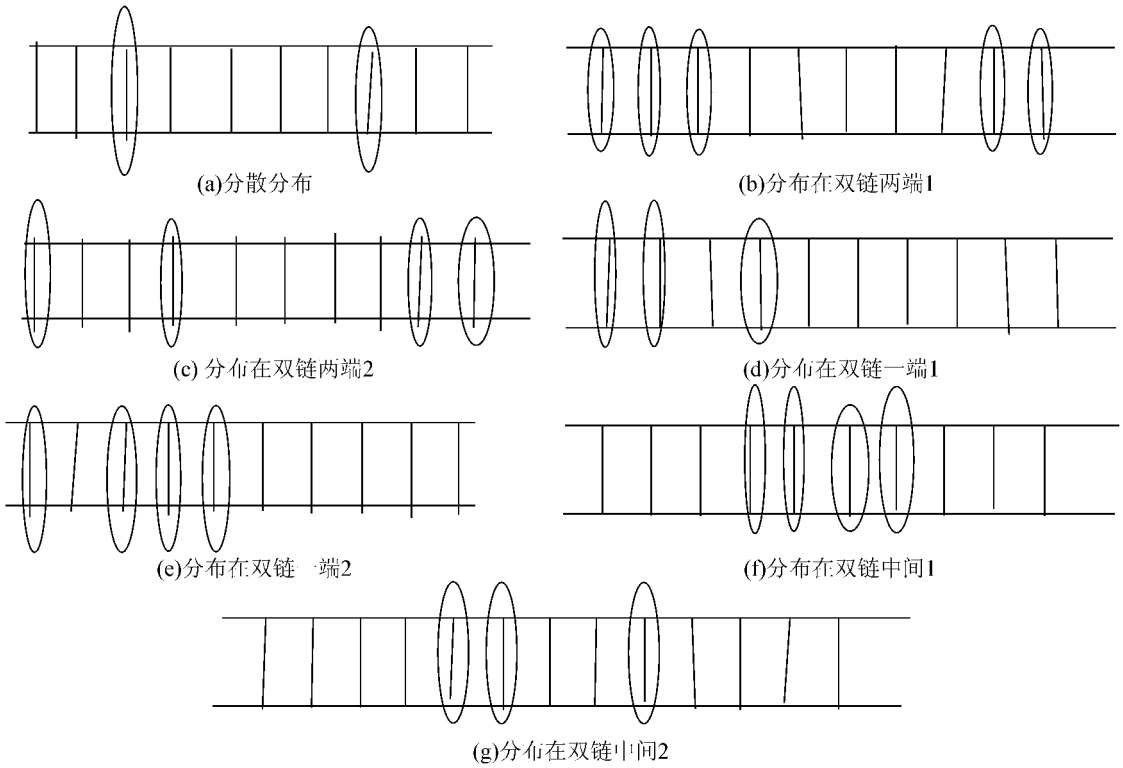

图 2　任意两等长 DNA 序列错误匹配情况

Fig. 2　All mistakes in DNA sequences

1)当洞分散分布在 DNA 双链中,其形式如图 2(a)所示,容易导致产生伪解并且不容易去除.

2)当洞集中分布在 DNA 双链两端时,其形式如图 2(b)、(c)所示.

左右两端分别选取下链和上链,将包含洞的部分用外切酶切去,即出现类似移位杂交的形式. 对于这种形式,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,从而实现完全匹配,反复操作,直到所有这种形式的移位杂交全部实现完全匹配. 然后将这些完整双链固定在充满聚丙烯酰胺凝胶体的玻璃板上,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,再用 94℃缓冲液冲洗玻璃板,冲洗后留在玻璃板上的即为改良后的 DNA 单链分子,从而实现对编码的优化.

3)当洞集中分布在 DNA 双链一端时,其形式如图 2(d)、(e)所示.

在出现洞的一端时选取下链(因为聚合酶链式反应的方向为),将包含洞的部分用外切酶切去,同上述情况一样,加入游离的核苷酸和聚合酶进行聚合酶链式反应,从而实现完全匹配,剩余操作同 2)中所述,不再赘述.

4)当洞集中分布在 DNA 双链中间时,其形式如图 2(f)、(g)所示.

用限制性内切核酸酶对包含洞的部分进行切割,其切割可分为齐式切割和非齐式切割(如图 3),这 2 种切割形式可用如下例子进行说明.

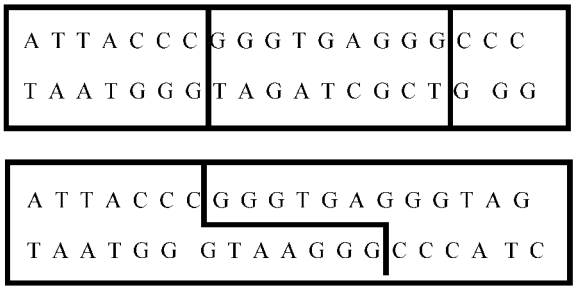

图 3　齐式切割与非齐式切割

Fig. 3　Blunt cuts and unblunt cuts

例 2　切割之后通过凝胶电泳实验将切割所得的 DNA 片断去除(因为在凝胶电泳中,由于较大的 DNA 片断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断移动得快,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,到达阳极的完整双链通过测定长度将长度小于给定长度的 DNA 双链去除),得到的即为理想的 DNA 双链分子,然后执行 2)中所述操作,得到的即为理想的 DNA 单链分子.

DNA 序列的不完全匹配包含以上 4 种情况,通过对其讨论可知,2)~4)这 3 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,可以从最大程度上

加以优化.

3 结束语

通过引入部分字及其洞的定义和性质,可以从最大程度上解决 DNA 编码中的不完全匹配问题. 众所周知,DNA 计算最主要和核心的反应为 DNA 分子间的杂交反应,其效率和精度直接影响到 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. 因此,DNA 编码中不完全匹配问题的解决可以在一定程度上避免假阳性的出现,为 DNA 编码理论的研究注入了活力. 21 世纪是生物科学的世纪,且随着晶体管的宽度逐渐接近极限,DNA 计算更成为人们普遍关注的一种计算模式. DNA 编码是 DNA 计算的基本问题之一,这方面的突破将有助于人们加深对 DNA 计算机的理解,促进 DNA 计算机解决目前电子计算机所无法解决或者很难解决的问题.

参考文献:

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

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

[3] BONEH D, DUNWORTH C, LIPTON R. Breaking DES using a molecular computer[C]//Proceedings 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]//Proceedings of the 3rd DIMACS Workshop on DNA Based Computers. [S. l.], 1997: 230-237.

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

[6] BAUM E B. DNA sequences useful for computation[C]// Proc Second Annual Meeting DNA-Based Computers. [S. l.], 1996: 223-227.

[7] GARZON M, DEATON R, NINO L F. A new metric for DNA computing[C]//Proceedings of the 2nd Annual Genetic Programming Conference GP-97. Morgan Kaufmann, USA, 1997: 472-487.

[8] GARZOR M, DEATON R, NINO L F, STEVENS S E, WITTER M. Genome encoding for DNA computing[C]// The Third DIMACS Workshop on DNA Based Computing. Pennsylvania, America, 1997: 230-273.

[9] FELDKAMP, STEVENTS S E, NINO L F. A DNA sequence complier[C]//Proceedings of 6th DIMACS Workshop on DNA Based Computing. Leidon, The Netherlands, 2000: 253.

[10] FRUTOS A G, BONEH D, DUNWORTH C. Demonstration of a word design strategy for DNA computing on surface[J]. Nucleic Acids Research, 1997, 25(23): 4748-4757.

[11] ARITA M, JOHNSON C, ROTHEMUND P W I. The power of sequence design in DNA computing[C]//The 4th International Conference on Computational Intelligence and Multimedia Applications. [S. l.], 2001: 163-167.

[12] BRAICH R S, JOHNSON C, ROTHEMUND P W K, ADLEMAN L M. Solution of a satisfy problem on a gel-based DNA computer[C]//The 6th International Workshop on DNA Based Computing. London, UK: Springer-verlag, 2001: 27-42.

[13] DAN C, TULPAN H H, CONDON H. Anne. Condon. Stochastic local search algorithms for DNA word design [C]//The 8th International Conference on Computational Intelligence and Multimedia Applications. [S. l.], 2003: 229-241.

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

[15] BLANCHET-SADRI F, HEGSTROM A. Partial words and a theorem of fine and Wilf[J]. Theoretical Computer Science, 2002, 270: 401-419.

[16] BLANCHET-SADRI F. Primitive partial words[C]//The 8th International Conference on Computational Intelligence and Multimedia Applications, [S. l.], 2003, 218: 135-141.

[17] LEUPOLD P. Partial words for DNA coding[C]// The 8th International Conference on Computational Intelligence and Multimedia Applications. [S. l.], 2005, 221: 224-234.

[19] 王淑栋,宋斐. DNA Golay 码的设计与分析[J]. 电子学报, 2009,37(7): 135-141.
WANG Shudong, SONG Tao. The design and analysis of DNA Golay codes[J]. Acta Electronica Sinica, 2009, 37 (7): 135-141.

[20] 陈鲁生, 沈世镒. 编码理论基础[M]. 北京:高等教育出版社, 2005: 168-221.

作者简介:

李珍,女,硕士. 主要研究方向为 DNA 计算.