

文章编号:1001-9081(2006)03-0627-03

## 基于粘贴 DNA 计算模型的数据存储技术

王延峰<sup>1,2</sup>, 强小利<sup>1</sup>, 崔光照<sup>2</sup>

(1. 华中科技大学 控制科学与工程系, 湖北 武汉 430074;

2. 郑州轻工业学院 电气信息工程学院, 河南 郑州 450002)

(wangyf@mail.hust.edu.cn)

**摘 要:**提出了一种新的基于粘贴 DNA 计算模型的数据存储技术的实现方法。该方法以重组 DNA 技术作为实现 DNA 数据存储的技术基础,以 DNA 计算理论研究中的粘贴模型作为信息编码工具。具体实现过程包括选择 DNA 载体,选择受体细胞,通过创建粘贴 DNA 计算模型的 ASC II 字符编码进行信息编码,创建数据索引,最后实现数据的存储与检索。

**关键词:**DNA 重组; 数据存储; 信息编码; 数据索引

**中图分类号:** TP311.12 **文献标识码:** A

## Data storage technology based on sticking model of DNA computing

WANG Yan-feng<sup>1,2</sup>, QIANG Xiao-li<sup>1</sup>, CUI Guang-zhao<sup>1,2</sup>

(1. Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan Hubei 430074, China;

2. College of Electric & Information Engineering, Zhengzhou University of Light Industry, Zhengzhou Henan 450002, China)

**Abstract:** A new method to store data based on DNA molecules was proposed. In this method, recombinant DNA technology was used to realize data storage, and the sticking model was used to encode information. In addition, a new ASCII code based on the sticking model was proposed for encoding information. In order to realize data storage and search, various biotechniques were touched on, such as selecting DNA carrier and receptor cell, creating data index, etc.

**Key words:** DNA recombinant; data storage; information encoding; data index

## 0 引言

目前,信息存储技术的主要发展方向是超大存储容量、数据传输率和较高系统可用性。其中超大存储容量的研究主要集中于:1)现有存储技术的更新换代,如高密度光存储技术研究和超高密度磁性存储介质技术研究;2)寻找新的存储介质(材料),如英国 Polight 公司将近期推出全息存储介质——“Holonide”无机类材料以及本文所要讨论的基于 DNA 的数据存储技术。

早在 30 多年前,人们就已经开始进行 DNA 数据存储技术的可行性研究<sup>[1,2]</sup>。其中 Hoch<sup>[1]</sup>和 Wong<sup>[3]</sup>两个研究小组的试验研究成果最具代表性。

Hoch 研究小组的试验过程可简述为:1)编码原始信息并合成人工 DNA 序列;2)将该 DNA 序列隐藏在打印文档中;3)将该文档密封后通过邮局投递;4)所包含的信息在实验室里得以完整恢复。试验证明,DNA 序列可以像纸张一样作为信息存储介质并能够在投递过程的恶劣环境中得以保存。不足之处在于它不能像现有存储设备一样长久地保存数据。Wong 研究小组则在实验室内使用基本的生物工程技术,在 7 个细菌体内合成了 7 个带有异源载体信息(57~99 个碱基对)的 DNA 片断,成功地实现了数据的存储和提取。这两个试验在一定程度上验证了基于 DNA 的数据存储技术的可行性。

基于 DNA 的数据存储技术是研究 DNA 计算机体系结构的重要一环。本文拟从 DNA 重组技术着手,以 Wong 的工作

为基础,对目前 DNA 计算中最具发展前景的粘贴模型的数据存储技术作初步探讨。

## 1 DNA 分子结构及存储优点

DNA 组成的基本单位是脱氧核苷酸。每个脱氧核苷酸由一分子磷酸、一分子脱氧核糖和一分子含氮碱基组成。含氮碱基有四种:腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)和胸腺嘧啶(T)。碱基之间按碱基互补配对原则进行配对,即腺嘌呤(A)一定与胸腺嘧啶(T)配对;鸟嘌呤(G)一定与胞嘧啶(C)配对。一串用 4 个字母 A,T,C 和 G(代表碱基)所形成的一个单链称为一个寡核苷酸。它与其互补链配对形成双链 DNA。所以,从编码角度看,这意味着任何一条信息都可以表示成一个 DNA 单链码,即 DNA 分子可以作为信息的携带者。由碱基对所组成的特殊分子可以被形象地认为是用于计算的处理单元。即 DNA 的碱基互补配对特性可被用于执行辅助型存储器<sup>[4]</sup>,从而为实现更广泛意义上的数据的存储与提取提供了可能。

作为存储介质,DNA 具有以下突出优点<sup>[5]</sup>:1)海量信息可置于其基序列中。在 DNA 分子中,核苷酸之间的间隙为 0.35nm,从而使 DNA 具有了近 7Mbits/cm 的不同寻常的数据密度。在二维空间中,如果假设每 nm<sup>2</sup> 有一个核苷酸的话,则其数据存储密度可以超过  $1.5 \times 10^5$  Gbits/cm<sup>2</sup>(典型的高性能硬盘的数据存储密度仅为 1.1Gbits/cm<sup>2</sup>)。另外,更重要的是,对于相同长度的序列,DNA 序列所携带的信息也远比二

收稿日期:2005-09-07 收稿日期:2005-11-16

基金项目:国家自然科学基金资助项目(60573190, 30370356, 60574041);河南省自然科学基金资助项目(511011600, 0211050900)

作者简介:王延峰(1973-),男,河南南阳人,讲师,博士研究生,主要研究方向:DNA 计算、基因网络; 强小利(1979-),女,陕西延安人,博士研究生,主要研究方向:DNA 计算; 崔光照(1957-),男,河南洛阳人,教授,博士,主要研究方向:DNA 计算、基因网络。

进制数字丰富。

2) 超强自纠错能力。DNA 序列的互补性使 DNA 成为一种独特的计算结构,这种性质可以多种方式加以利用。典型例子就是自纠错能力。由于种种原因,DNA 或酶在生化反应过程中均可能出现错误。如果错误仅仅发生在双链 DNA 的某一段上,则修复酶利用补序列串作为参考,即可恢复正确的 DNA 序列。在生物系统中,这种纠错能力意味着极低的错误率。例如,在 DNA 复制时,每复制  $10^9$  个核苷酸仅发生一次错误,即错误率仅为  $10^{-9}$ 。

3) 在生命体中,DNA 可循迹记录并且具有自我复制功能。作为生命体的遗传物质,DNA 的主要功能是存储和传递遗传信息。遗传物质所必须具备的稳定性特点和能够精确的自我复制,从而使亲代与子代间保持遗传的连续性的特点仿佛暗示着 DNA 是大自然提供给人类的天然的、完美的存储介质。

4) 容易得到妥善保护。这得益于生物长期进化所形成的自我保护功能,DNA 以生命体作为自然屏障,既可适应于自然的或极端的(如果需要)环境条件,又可通过遗传保持信息的连续传递。

## 2 重组 DNA 技术

重组 DNA 技术是指利用酶学的方法,在体外将各种来源的遗传物质(同源的或异源的,原核的或真核的,天然的或人工的)与载体 DNA 接合成一个具有自我复制能力的 DNA 分子——复制子,然后通过转化或转染宿主细胞,筛选出含有目的基因的转化子细胞,再进行扩增提取获得大量同一 DNA 分子。也称基因克隆<sup>[6]</sup>。

DNA 重组技术操作过程可以形象地归纳为:分(分离目的基因),切(用限制酶剪切目的基因和载体),接(拼接重组体),转(转入受体菌)和筛(筛选重组体)。DNA 重组技术操作的主要步骤如图 1 所示。

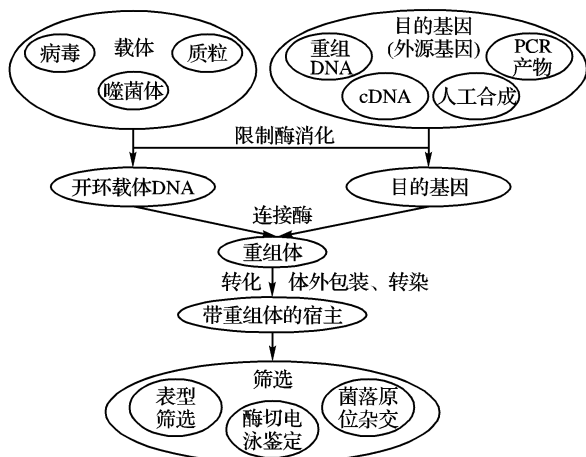


图1 DNA 重组技术操作的主要步骤

## 3 DNA 数据存储

利用 DNA 重组技术,以 DNA 作为存储介质的数据存储操作需要以下 5 个步骤:1) 选择 DNA 载体;2) 选择受体细胞;3) 信息编码;4) 创建数据索引;5) 数据的存储与检索。

### 3.1 DNA 载体的选择

Murray 于 1983 年首次成功构建酵母人工染色体(Yeast Artificial Chromosome, YAC)<sup>[7]</sup>,Burke 于 1987 年构建 YAC 克隆库<sup>[8]</sup>。酵母染色体的控制系统包括 3 部分:1) 着丝粒(Centromere, CEN)。它的作用是使染色体的附着粒与有丝分

裂的纺锤丝相连,保证染色体在细胞分裂过程中正确分配到子代细胞中。2) 端粒(Telomere, TEL)。位于染色体两个末端,它的功能是保护染色体两端,保证染色体的正常复制,防止染色体 DNA 复制过程中两端序列的丢失。3) 酵母自主复制序列。其功能与酵母细胞复制有关。如图 2 所示,图中,URA3 代表尿嘧啶核苷酸合成酶基因 3;NotI 代表限制性酶切位点。

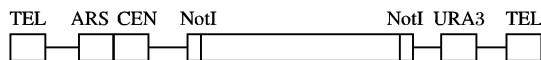


图2 酵母人工染色体(YAC)

简单地讲,酵母人工染色体克隆载体就是将酵母染色体 DNA 的端、DNA 复制起点(Autonomously Replicating Sequence, ARS)和丝粒以及必要的选择标志(HIS4 和 TRP1)基因序列克隆到大肠杆菌质粒 pBR322 中,从而构建 YAC 克隆载体。所以,YAC 克隆载体既含有质粒克隆载体所必备的第一受体(大肠杆菌)源质粒复制起始位点(ori),还含有第二受体(酵母菌)染色体 DNA 的着丝点、端粒和复制起始位点的序列,以及合适的选择标志基因。

YAC 克隆外源 DNA 能力非常大,一个 YAC 可插入长达  $10^6$  个碱基以上 DNA 片段,因此,YAC 可以保证所插入外源基因结构的完整性。目前,YAC 已成为构建高等真核生物基因库的重要载体,并在人类基因组的研究中起着重要作用。所以,选择酵母人工染色体作为 DNA 载体。

### 3.2 受体细胞的选择

受体细胞(Receptor Cell)又称为宿主细胞或寄主细胞(Host Cell),是指能摄取外源并使其稳定维持的有应用价值和理论研究价值的细胞。

恐兽球菌(Deinococcus Radiodurans)最早由 Anderson 于 1956 年发现,是迄今为止地球上发现的最具抗辐射功能的生物之一。具体表现为对电离辐射、紫外线、干燥、强氧化剂和一些化学诱变剂具有惊人的抗性。实验证明,该细菌细胞能够在几十小时内准确无误地修复由辐射所引起的几百个 DNA 双链断裂(DSBs)片断。最近通过电镜观察到致密有序的 DSBs 环状堆积结构在 DNA 的重组修复起着十分重要的作用。另外,恐兽球菌作为受体细胞,符合受体细胞选择的基本原则<sup>[6]</sup>。所以,选择恐兽球菌作为受体细胞。

### 3.3 信息编码

信息编码的过程实际上就是对原始信息进行编码,合成人工 DNA 序列。

根据粘贴计算模型二进制编码方法可以知道:粘贴模型有一个由存储合成物所构成的随机访问存储区。一个存储合成物是一个可以看作是二进制数的编码的部分双链的 DNA 串。其中,双链代表“1”,单链代表“0”。每个存储合成物由两种称为“存储链”和“粘贴链”的单链 DNA 分子形成。其中,一个存储链是一个单链的 DNA 分子,由  $l$  个碱基构成。一个存储链含有  $n$  个不重叠的子链,每个子链由  $m$  个碱基构成。取  $l = m \times n$ 。例如,下面是一个长度为  $l = 5 \times 6$  的存储链,其中, $m = 5, n = 6$ :

5'ATCGA TAGCA CCATG TAGAT CGCGT TTAAG 3'

要求在一个存储链中,每个粘贴链恰好与  $n$  个子链中的一个互补。存储链的每个子链被视为一个位元的位置。如果一个粘贴链被退火于存储链的匹配子链上,则这个特殊的子链为“开”(相当于数字逻辑电路中的高电平);否则为“关”(相当于数字逻辑电路中的低电平)。因此,存储合成物可用

来表示一个二进制数,其中子链为“开”表示该位元的数为“1”,子链为“关”表示该位元的数为“0”。图3是4个存储合成物的例子。

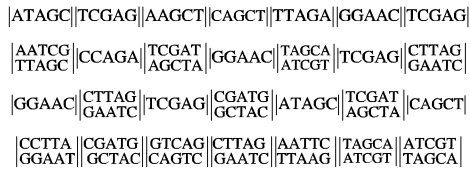


图3 表示4个二进制数的存储合成物

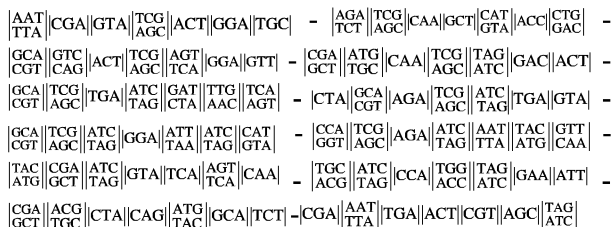
图3中所表示的四组二进制数分别为:0000000,1010101,0101010,1111111。

由此就可以列出粘贴DNA计算模型的ASCII字符编码,如表1所示,表中列出了其中96个(不包括32个通用控制字符)。表中将7位二进制编码用3位八进制编码表示。

八进制码	DNA序列图示	字符	八进制码	DNA序列图示	字符	八进制码	DNA序列图示	字符
040		SP	100		@	140		`
041		!	101		A	141		a
042		"	102		B	142		b
043		#	103		C	143		c
044		\$	104		D	144		d
045		%	105		E	145		e
046		&	106		F	146		f
047		'	107		G	147		g
050		(	110		H	150		h
051		)	111		I	151		i
052		*	112		J	152		j
053		+	113		K	153		k
054		,	114		L	154		l
055		-	115		M	155		m
056		.	116		N	156		n
057		/	117		O	157		o
060		0	120		P	160		p
061		1	121		Q	161		q
062		2	122		R	162		r
063		3	123		S	163		s
064		4	124		T	164		t
065		5	125		U	165		u
066		6	126		V	166		v
067		7	127		W	167		w
070		8	130		X	170		x
071		9	131		Y	171		y
072		:	132		Z	172		z
073		;	133		[	173		{
074		<	134		\	174		
075		=	135		]	175		}
076		>	136		??	176		~
077		?	137		?	177		DEL

图4 粘贴DNA计算模型的ASCII字符编码

如果每个存储链由21个碱基构成。每个存储链含有7个不重叠的子链。每个子链由3个碱基构成,那么“hello, world!”经编码后的DNA序列为:



### 3.4 创建数据索引



图5 作为创建数据索引蓝本的DNA序列

大肠杆菌的整个染色体组已经被完整地测序,并且可以通过www.tigr.org检索到。这里借鉴Wong的试验中所选用

的25个序列作为创建数据索引的蓝本<sup>[3]</sup>,如表2所示。序列中以ATG或GTG表示起始密码子,以TAG或TAA或TGA表示终止密码子,用于提示细菌已经到达载体DNA序列末端,应该停止翻译。

### 3.5 数据的存储与检索

在数据的存储与检索过程中,所需生化过程包括:产生补链;在开环载体DNA中插入人工合成DNA序列,通过连接酶合成重组体;转换序列成活性组织;组织生长和繁殖;从组织中萃取信息。

#### 3.5.1 产生补链

在存储数据之前,需要先产生两条互补链。每条链携带46个碱基,其中包括两个不同的20个碱基长的片断,中间通过6个碱基长的限制性核酸内切酶识别序列相连。这两条20个碱基长的寡核苷酸从图5中选择。限制性酶切位点是为了随后插入编码DNA片断。这两个46个碱基长的互补链形成一个双链DNA片断。这个DNA片断然后被克隆为重组质粒——体外DNA片断组合成一个环形DNA分子。因为这20个碱基长的寡核苷酸不属于载体的染色体组,所以可用于随后的数据索引的识别标记。

如果选择TTAGGGATGTGTGTAGTTAG和GGTTAGATGACTGTAGTTAG两条DNA序列作为数据索引蓝本,根据碱基互补配对原则,其对应的互补链分别为:AATCCCTACACACATCAATC和CCAATCTACTACATCAATC。设两条互补链中间所插入的6个碱基长的限制性核酸内切酶识别序列为: $\begin{bmatrix} \text{GAATTC} \\ \text{CTTAAG} \end{bmatrix}$ ,则包含数据索引序列和限制性核酸内切酶识别序列的每条链携带46个碱基的两条互补链为: $\begin{bmatrix} \text{TTAGGGATGTGTGTAGTTAG} & \text{GAATTC} & \text{GGTTAGATGACTGTAGTTAG} \\ \text{AATCCCTACACACATCAATC} & \text{CTTAAG} & \text{CCAATCTACTACATCAATC} \end{bmatrix}$ 。

#### 3.5.2 目的基因转入受体细胞

外源目的基因(环形DNA分子)在体外连接重组后形成重组DNA分子,该重组DNA分子必须导入适宜的受体细胞中才可以使外源目的基因得以大量扩增或表达,即在带重组体的宿主体内自我复制。结果带有编码DNA片断的数据就可以被转移入大肠杆菌体内,如图6所示。然后通过电穿孔(Electroporation)<sup>[11]</sup>法把宿主细胞与外源DNA混合并置于电击槽中。在高压电脉冲作用下,使细胞膜瞬时击穿出现微孔,外源DNA通过微孔进入细胞,允许带有编码宿主细胞繁殖以备后用。

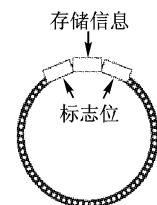


图6 两个DNA片断结合在一起形成的重组质粒

#### 3.5.3 转换序列成活性组织

载体和编码DNA然后被合并入恐兽球菌染色体组内形成带重组体的受体细胞,用于存取信息。当能够忍受高辐射、紫外线、干燥、强氧化剂等极端环境时,载体就能够为编码信息提供完美的保护。

#### 3.5.4 信息的提取

当需要提取编码信息的时候,我们通过聚合酶链反应从细菌体内萃取出信息部分的DNA链;根据预先知道的片断两

的机会;当  $\theta_1 < \text{Sim}(x_{b_j}, x_{s_j}) \leq \theta_2$  时,判断卖方属于过渡类型  $a_h$ ,买方采取线性策略  $s_l$ ;当  $\theta_2 < \text{Sim}(x_{b_j}, x_{s_j}) \leq 1$  时,判断卖方属于亲我类型  $a_c$ ,买方采取让步策略  $s_c$ ,这样使双方达到一致的可能性最大。

#### 4 算法实现

```

REPEAT
  FOR  $j = 1$  to  $n$ 
    { IF  $t < T_{b_j}$  THEN
      send(  $x_{b_j}(t)$  ) to  $c$ ;
      send(  $x_{s_j}(t)$  ) to  $c$ ;
      send(  $s_j$  . action ) to  $c$ ;
    ELSE terminate (  $b_j, s_j$  );
    END IF; }
  //c Decision:
  IF  $s_j$  . action = "accept " THEN  $x_{b_j}(t) = x_{s_j}(t)$ ;
  IF  $s_j$  . action = "refuse " THEN  $x_{s_j}(t) = \text{null}$ ;
  //评估提议
   $\text{Sim}(x_{b_j}(t), x_{s_j}(t)) = \sum_{i \in I} h^i \cdot \text{Sim}^i(x_{b_j}^i(t), x_{s_j}^i(t))$ ;
  IF  $0 \leq \text{Sim}(x_{b_j}(t), x_{s_j}(t)) \leq \theta_1$  THEN
     $s_j$  . AT: =  $a_i$ ; //非亲我类型
     $b_j$  . ST: =  $s_i$ ; //强硬策略
  END IF;
  IF  $\theta_1 < \text{Sim}(x_{b_j}(t), x_{s_j}(t)) \leq \theta_2$  THEN
     $s_j$  . AT: =  $a_h$ ; //过渡类型
     $b_j$  . ST: =  $s_l$ ; //线性策略
  END IF;
  IF  $\theta_2 < \text{Sim}(x_{b_j}(t), x_{s_j}(t)) \leq 1$  THEN
     $s_j$  . AT: =  $a_c$ ; //亲我类型
     $b_j$  . ST: =  $s_c$ ; //让步策略
  END IF;
  send(  $b_j$  . ST ) to  $b_j$ ;
  //选择时间信念函数
  IF  $b_j$  . ST =  $s_t$  THEN  $p_{b_j \rightarrow s_j}(t) = t/T_b$ ;
  IF  $b_j$  . ST =  $s_l$  THEN  $p_{b_j \rightarrow s_j}(t) = 1/2$ ;
  IF  $b_j$  . ST =  $s_c$  THEN  $p_{b_j \rightarrow s_j}(t) = 1 - t/T_b$ ;
  //子买方  $b_j$  生成提议
   $x_{b_j} = \max(x_{b_j}) - \left( \sum_{k=1}^{T_b} p_{b_j \rightarrow s_j}(k) \cdot \gamma^{k-1} \cdot Q_{b_j}^T \right) /$ 
  (  $T_b - t + 1$  );

```

```

UNTIL  $t \geq T_b$  OR  $c$  . action = "confirm " END REPEAT;
IF  $t \geq T_b$  THEN terminate (  $b, s$  );
IF  $c$  . action = "confirm " THEN deal;

```

#### 参考文献:

- [1] NGUYEN TD, JENNINGS NR. A Heuristic Model for Concurrent Bi-lateral Negotiations in Incomplete Information Settings[ A ]. Proceedings of 18th International Joint Conference on AI[ C ]. Mexico, 2003.
- [2] RAHWAN I, KOWALCZYK R, PHAM HH. Intelligent Agents for Automated One-to-Many E-Commerce Negotiation[ A ]. Twenty-Fifth Australian Computer Science Conference ( ACSC2002 ) [ C ]. Australian, 2002. 197 - 204.
- [3] ARAI S, SYCARA K, PAYNE T. Experience-based Reinforcement Learning to Acquire Effective Behavior in a Multi-agent Domain[ A ]. Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence[ C ], 2000.
- [4] ZENG D, SYCARA K. Bayesian Learning in Negotiation [ A ]. Working Notes for the AAAI Symposium on Adaptation, Co-evolution and Learning in Multiagent Systems[ C ]. Stanford University, CA, 1996.
- [5] EXCELENTE-TOLEDO CB, JENNINGS NR. Using reinforcement learning to coordinate better[ J ]. Computational Intelligence, 2005, 21 (3): 217 - 245.
- [6] OLIVER JR. A machine-learning approach to automated negotiation and prospects for electronic commerce[ J ]. Journal of Management Information Systems, 1997, 13(3): 83 - 112.
- [7] TAN M. Multi-Agent Reinforcement Learning: Independent vs. Co-operative Agents[ A ]. Proceedings of the Tenth International Conference on Machine Learning[ C ]. 1993. 330 - 337.
- [8] MITCHELL TM. Machine Learning[ M ]. Beijing: China Machine Press, 2003.
- [9] NAGAYUKI Y, ISHII S, DOYA K. Multi - agent reinforcement learning: An approach based on the other agent's internal model [ A ]. Proceedings on the Fourth International Conference on Multi-Agent Systems ( ICMAS-00 ) [ C ]. Boston, MA, 2000. 215 - 221.
- [10] BUI H, KIERONSKA D, VENKATESH S. Learning other agents' preferences in multiagent negotiation[ A ]. Proceedings of the Thirteenth National Conference on Artificial Intelligence[ C ]. Menlo Park, CA, AAAI Press, 1996. 114 - 119.

(上接第629页)

端的序列信息,通过一系列的加热和冷却循环来扩增 DNA 片段<sup>[11]</sup>;然后通过相应的序列读取装置读取序列;最后,转换序列为原始数据。

#### 参考文献:

- [1] HOCH JA, LOSICK R. Panspermia, spores and the bacillus subtilis genome[ J ]. Nature, 1997, 390: 237 - 238.
- [2] CLELLAND CT, RISCA V, BANCROFT C. Hiding messages in DNA microdots[ J ]. Nature, 1999, 399: 533 - 534.
- [3] WONG PC, WONG K-K, FOOTE H. Organic data memory using the DNA approach[ J ]. Communications of the ACM, 2003, 46 (1): 95 - 98.
- [4] WASIEWICZ P, MALINOWSKI A, NOWAK R, et al. DNA computing: implementation of data flow logical operations[ J ]. Future Generation Computer Systems, 2001, 17(4): 361 - 378.
- [5] JONATHAN P, COX L. Long-term data storage in DNA[ J ]. Trends

in biotechnology, 2001, 19(7): 247 - 250.

- [6] 楼士林, 杨盛昌, 龙敏南. 基因工程[ M ]. 北京: 科学出版社, 2002.
- [7] MURRAY AW, SZOSTAK JW. Construction of ratification chromosome in yeast[ J ]. Nature, 1983, 305: 189.
- [8] BURKE DT, CARLE GF, OLSON MV. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors[ J ]. Science, 1987, 236: 806.
- [9] 许进, 董亚非, 魏小朋. 粘贴 DNA 计算机模型( I ): 理论[ J ]. 科学通报, 2004, (3): 205 - 212.
- [10] 许进, 李三平, 董亚非, 等. 粘贴 DNA 计算机模型( II ): 应用[ J ]. 科学通报, 2004, (4): 299 - 307.
- [11] HARWOOD AJ. Basic DNA and RNA protocols[ M ]. 盛小禹, 等译. 北京: 科学出版社, 2002.