

计算科学的新领域:DNA 计算(I)

李 燕

(浙江师范大学信息科学与工程学院 金华 321004)

摘 要 DNA 计算是应用分子生物技术进行计算的新方法。从理论上研究 DNA 计算方法,有利于推动理论计算科学的发展。本系列文章应用形式语言及自动机理论技术,系统地探讨了 DNA 分子的可计算性及其计算能力。本文主要介绍常用 DNA 分子操作方法,并根据 DNA 分子的结构及特点,给出了 DNA 分子的形式化描述。

关键词 DNA 计算, DNA 分子操作方法, 形式化描述

A New Field of Compute Science: DNA Computing (I)

LI Yan

(College of Information Science and Engineering, ZheJiang Normal University, Jinhua 321004)

Abstract DNA computing is a new method for computation using the technology in molecular biology. The study of DNA computing theory will be of benefit to computing science theory. The series papers systematic discuss the computability and the computational capacity of DNA molecular using the formal language and automata theory. In this paper, we focus on common DNA manipulation, and provide the formal description of DNA molecular based on it's structure and characteristic.

Keywords DNA computing, DNA molecular manipulation, Formal description

1 引言

DNA 是一种高分子化合物,称为脱氧核糖核酸^[5](简称为核苷酸)。DNA 分子由不同的核苷酸排列而成,遗传信息就存在于其中。每一个构成 DNA 的核苷酸是由三部分组成的:①一个脱氧核糖分子 S;②一个磷酸基团 P;③一个含氮碱基(A、T、G、C)。脱氧核糖有 5 个碳原子,为了避免糖上的碳与碱基上的碳相混淆,将糖上的碳编码为 1' 到 5'。组成核苷酸的含氮碱基有四种,它们是腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)、胸腺嘧啶(T)。核苷酸可以通过两种不同的方式连接,一种是磷酸二脂键的绑定,这种连接是有方向性的。磷酸二脂键的连接是非常稳固的,除非用限制性内切酶进行切割才能断开。另一种是氢键绑定,氢键绑定比磷酸二脂键的绑定要弱,特别是单个碱基对之间的连接是非常弱的,它需要较长的碱基延伸才能保持绑定,这种绑定通过一定的分子操作方法可以断开。DNA 大分子是由两条头尾方向相反的脱氧核糖核苷酸长链,以右手螺旋的方式盘绕着同一中心轴而形成的。两条链上的碱基通过氢键连接起来,碱基对的连接严格依据 Watson-Crick 碱基互补配对原则。即腺嘌呤(A)通过两个氢键与胸腺嘧啶(T)配对,鸟嘌呤(G)通过三个氢键与胞嘧啶(C)配对。反过来也必定是 T 与 A 配对, C 与 G 配对。

DNA 分子计算是计算科学的新领域^[1~3]。DNA 分子链可以看作是由四个不同的符号 A、T、C、G 连成的一个串,所以可使用四个字母的集合 $\Sigma = \{A, T, C, G\}$,按照 DNA 的双螺旋结构和 Watson-Crick 碱基互补配对原则对问题进行编码。如果把 T 看作 0, C 看作 1, G 看作 2, A 看作 3, 一个核苷酸链就成了一个由四进制数组成的串。电子计算机中“0”和“1”编码可代表的信息,自然也可以用四进制数来表示。酶可以对 DNA 序列进行操作,不同的酶相当于 DNA 串上的不同算子^[6~8]。关于 DNA 计算的研究,已受到科学领域内许多

学者们的关注与兴趣^[9~16]。本文主要介绍常用 DNA 分子操作方法,并根据 DNA 分子的结构及特点,给出了 DNA 分子的形式化描述,以便从理论上探讨 DNA 分子计算。

2 DNA 分子的基本操作方法

常用 DNA 分子操作技术和生物酶,是处理 DNA 分子的基本工具箱。了解生物技术对 DNA 信息链的操作处理能力,对于研究 DNA 计算方法是相当重要的。下面主要介绍几种常用的分子操作方法和生物酶^[5]。

2.1 DNA 分子的长度及其测量方法

一个单链 DNA 分子的长度是组成分子的核苷酸单个碱基的数量。例如,如果一个单链 DNA 分子是由 16 个核苷酸单体组成的聚合体,则称它为 16mer。一个双链 DNA 分子的长度是指组成分子的核苷酸碱基对的数量。如果一个 16mer 的 DNA 单链通过生物操作组成一个双链 DNA 分子,这个双链分子的长度是 16 碱基对,表示为 16bp。用凝胶电泳技术可以使 DNA 链按长度分离,测出 DNA 链的长度。凝胶电泳如图 1 所示。

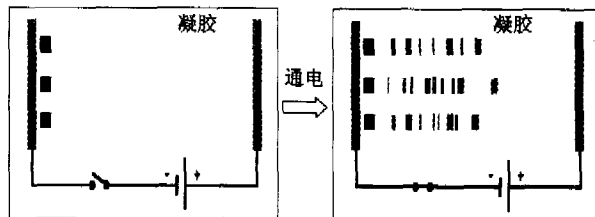


图 1 凝胶电泳

2.2 DNA 分子的双链分离和结合

DNA 分子中碱基互补的氢键连接比链中连续的核苷酸之间的磷酸连接要弱,在不破坏 DNA 单链的情况下,通

过加热、有机溶剂等破坏碱基间氢键的技术,使得组成双链 DNA 分子的两个单链分开,这个过程叫做 DNA 分子的变性。DNA 变性如图 2 所示。常用的方法是将 DNA 溶液加热到 80~100℃,使得双链 DNA 分子的两条链分开。

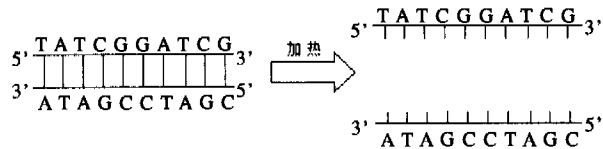


图 2 DNA 变性

如果把加热的溶液缓慢冷却,分开的单链可以按碱基互补关系重新恢复双螺旋结构,这是与变性相逆的过程,这个过程叫做 DNA 的复性或退火。当参与复性的核酸来源不同时,这个过程就称为杂交。

2.3 选择需要的 DNA 分子

如果准备从包含有双链 DNA 分子的溶液中选择出需要的目标分子,可以先将溶液加热,使双链分子溶解为单链分子,再使用探针、磁珠分离或者亲和层析(过滤析取)等技术找到目标分子,然后加入聚合酶,将这个单链分子聚合成双链分子。亲和层析如图 3 所示。

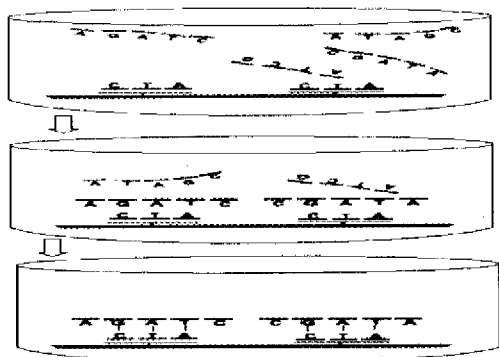


图 3 亲和层析

2.4 DNA 分子的延长

如果组成 DNA 分子的两条链的碱基数量不一样多,则称为具有粘性末端的 DNA 分子。如果两条链的碱基数量一样多,并且一一对应,称为平端 DNA 分子。

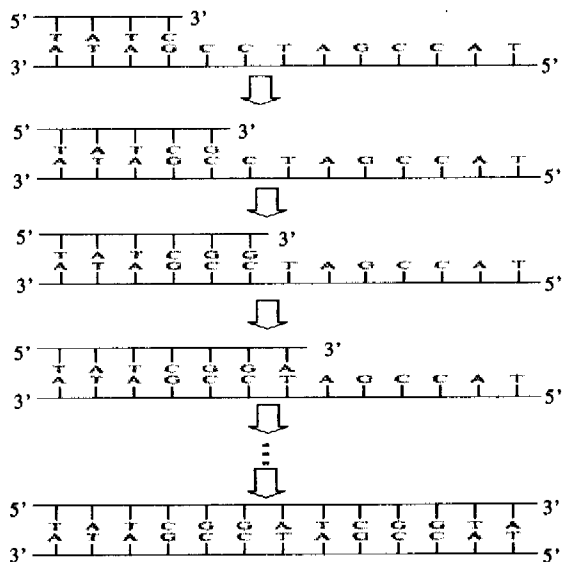


图 4 聚合酶链式反应

聚合酶是一种能够在 DNA 模板上催化寡核苷酸聚合的酶。通过聚合酶链式反应可以延长 DNA 分子。聚合酶链式反应(Polymerase Chain Reaction, 简称为 PCR)法是 1985 年由 K. B. Mullis 等研制成功的。如图 4 所示,具有粘性末端的 DNA 分子在聚合酶的作用下,溶液中的寡核苷酸按照碱基互补原则,聚合到模板 3'-ATAGCCTAGC CAT-5'上,使得 5'-TATC-3'沿着 5'→3'的方向延伸。

2.5 DNA 分子的连接(粘结)

两个具有粘性末端的 DNA 分子,如果它们的末端符合碱基互补原则,通过退火反应,这两个分子可以连接在一起。连接酶可以将 DNA 链的 5'端与 3'端连接到一起,也可以将 DNA 链的 3'端与 5'端连接到一起,防止重组分子变松动。在连接酶的作用下,还可以将单链 DNA 分子连接到平端分子中的一条链上,形成具有粘性末端的 DNA 分子,如图 5 所示。

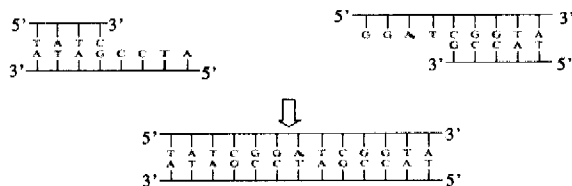


图 5 互补粘性末端的粘结

2.6 DNA 分子的复制(扩充)

复制是指以原来的 DNA 分子为模板合成相同分子的过程。DNA 的复制包括起始、延伸、终止三个步骤。进行 DNA 分子的复制,首先要将双链 DNA 分子溶解成两条 DNA 单链,这两条单链中的任何一条链都可以作为模板;第二步,在包含 DNA 单链及寡核苷酸的试管中加入引物,引物与模板中的互补序列发生退火;第三步,加入聚合酶,引物在聚合酶的作用下,按照 A 与 T、G 与 C 相互配对的原则,沿着 5'→3'的方向延伸,形成双链 DNA 分子。图 6 所示为 DNA 单链的复制过程。

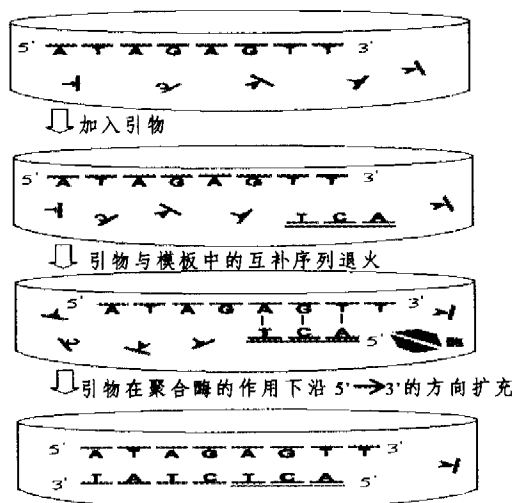


图 6 DNA 复制

2.7 DNA 分子的切割

凡是识别和切割双链 DNA 分子内特殊核苷酸序列的酶称为限制性核酸内切酶(简称限制酶)。限制酶首先在 DNA 分子链上找到切割位置,然后再进行切割,如图 7 所示。不同

的限制酶的切割位点及切口是不同的。有的限制酶的切割结果是具有粘性末端的 DNA 分子,有的限制酶的切割结果是具有平端的 DNA 分子。

以上介绍的是一些与 DNA 计算相关的常用分子操作技术及生物酶。随着 DNA 分子计算方法的深入研究以及生物技术的发展,操作方法及生物酶还会得到改进和补充。

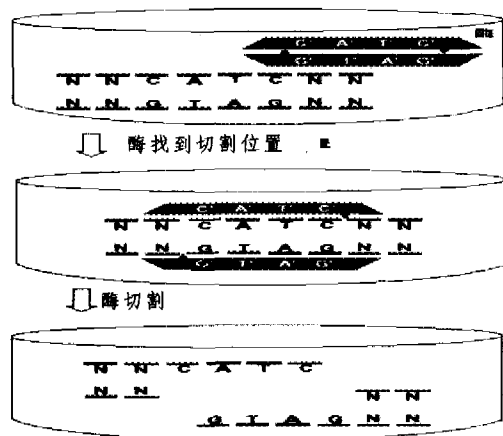


图 7 DNA 切割

3 DNA 分子的形式化描述

用形式化的方法描述 DNA 分子的结构是研究分子计算理论的基础^[17,18]。DNA 分子的双螺旋链提供了二维的数据信息。符号 Σ 表示组成 DNA 分子的字母表 $\{A, T, C, G\}$, 奇异的表示为 Σ^* 。字母表 Σ 上的互补对应关系表示为 $\rho: \Sigma \rightarrow \Sigma$, 对于所有的 $b \in E$, 存在 $d \in \Sigma (b \neq d)$, 使得 $\rho(b) = d$ 。为了描述方便, 用符号 $\bar{b}$ 表示 $\rho(b)$ 。这样集合 ρ 中的基本元素为: $\{(A, T), (T, A), (C, G), (G, C)\}$, 其中, $\bar{A} = T, \bar{T} = A, \bar{C} = G, \bar{G} = C$ 。DNA 分子由具有碱基互补关系的上下两条链(字符串序列)组成, 分别称为上链和下链。集合 $FZ_\rho(\Sigma) = \left[\begin{smallmatrix} \Sigma \\ \Sigma \end{smallmatrix} \right]_\rho^*$ 表示在字母表 Σ 上具有碱基互补关系 ρ 的 DNA 分子。其中, $\left[\begin{smallmatrix} \Sigma \\ \Sigma \end{smallmatrix} \right]_\rho = \left\{ \left[\begin{smallmatrix} a \\ b \end{smallmatrix} \right] \mid a, b \in \Sigma, (a, b) \in \rho \right\}$ 。

对于 DNA 分子的形式描述, 符号 $\begin{pmatrix} a \\ b \end{pmatrix}$ 和 $\left[\begin{smallmatrix} a \\ b \end{smallmatrix} \right]$ 是有区别的: $\begin{pmatrix} a \\ b \end{pmatrix}$ 中的 a 和 b 所代表的是没有互补关系的一对符号(a, b); 而 $\left[\begin{smallmatrix} a \\ b \end{smallmatrix} \right]$ 中的 a 和 b , 代表的是字母表 Σ 上按照碱基互补关系 ρ 精确绑定的一对符号(a, b)。集合 $\left(\begin{smallmatrix} \Sigma^* \\ \Sigma^* \end{smallmatrix} \right)$ 中的空元素 $\begin{pmatrix} \epsilon \\ \epsilon \end{pmatrix}$ 一般写成 ϵ , 当没有意义时可省略。 $\begin{pmatrix} u \\ \epsilon \end{pmatrix}$ 与 $\left[\begin{smallmatrix} a \\ b \end{smallmatrix} \right]$ 的连接表示为 $\begin{pmatrix} u \\ \epsilon \end{pmatrix} \left[\begin{smallmatrix} a \\ b \end{smallmatrix} \right]$, 不能用 $\begin{pmatrix} ua \\ b \end{pmatrix}$ 代替, 否则就失去了 a, b 的碱基互补关系。

集合 $FZ_\rho(\Sigma)$ 中的元素 $\left[\begin{smallmatrix} a_1 \\ b_1 \end{smallmatrix} \right] \left[\begin{smallmatrix} a_2 \\ b_2 \end{smallmatrix} \right] \cdots \left[\begin{smallmatrix} a_n \\ b_n \end{smallmatrix} \right] \in FZ_\rho(\Sigma)$ 可写成 $\left[\begin{smallmatrix} w_1 \\ w_2 \end{smallmatrix} \right]$ 的形式, 其中, $w_1 = a_1 a_2 \cdots a_n, w_2 = b_1 b_2 \cdots b_n$ 。集合

$FZ_\rho(\Sigma)$ 中的元素 $\left[\begin{smallmatrix} w_1 \\ w_2 \end{smallmatrix} \right] \in FZ_\rho(\Sigma)$ 称为完备形式的双链序列或完备形式的分子, w_1 称为上链, w_2 称为下链。相应地, 集合 $FZ_\rho(\Sigma)$ 被称为是完备分子的集合。

DNA 分子除了上述描述的具有完备形式的之外, 还有非完备形式的 DNA 分子, 称为“准 DNA 分子”, 用符号 $Z_\rho(\Sigma)$ 表示准分子的集合。

准 DNA 分子的集合为: $Z_\rho(\Sigma) = L_\rho(\Sigma) \cup R_\rho(\Sigma) \cup LR_\rho(\Sigma)$, 其中,

$$L_\rho(\Sigma) = \left(\begin{pmatrix} \epsilon \\ \Sigma \end{pmatrix} \right) \cup \left(\begin{pmatrix} \Sigma^* \\ \epsilon \end{pmatrix} \right) \left[\begin{smallmatrix} \Sigma \\ \Sigma \end{smallmatrix} \right]_\rho^*,$$

$$R_\rho(\Sigma) = \left[\begin{smallmatrix} \Sigma \\ \Sigma \end{smallmatrix} \right]_\rho^* \left(\begin{pmatrix} \epsilon \\ \Sigma \end{pmatrix} \right) \cup \left(\begin{pmatrix} \Sigma^* \\ \epsilon \end{pmatrix} \right),$$

$$LR_\rho(\Sigma) = \left(\begin{pmatrix} \epsilon \\ \Sigma \end{pmatrix} \right) \cup \left(\begin{pmatrix} \Sigma^* \\ \epsilon \end{pmatrix} \right) \left[\begin{smallmatrix} \Sigma \\ \Sigma \end{smallmatrix} \right]_\rho^+ \left(\begin{pmatrix} \epsilon \\ \Sigma \end{pmatrix} \right) \cup \left(\begin{pmatrix} \Sigma^* \\ \epsilon \end{pmatrix} \right).$$

准 DNA 分子的结构形式如图 8 所示。

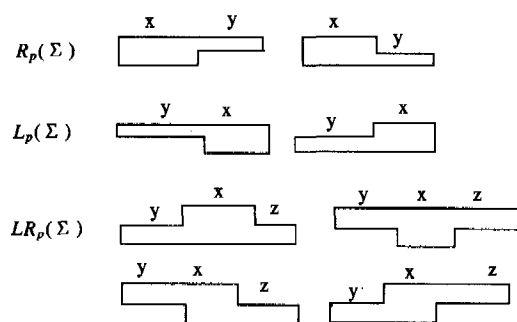


图 8 准 DNA 分子的结构

从图 8 中可以看出, 准分子序列中的 x 是具有完备形式的 DNA 双链, 在 x 的一端或两端分别延伸出单链序列 y, z , 这种延伸出的单链序列称为粘性末端。

集合 $L_\rho(\Sigma)$ 和 $R_\rho(\Sigma)$ 中的元素, 可以没有 x 序列, 只有单链序列 y , 这种形式的 DNA 分子被称为单链 DNA 分子。

集合 $LR_\rho(\Sigma)$ 中的元素, 因为 $x \in \left[\begin{smallmatrix} \Sigma \\ \Sigma \end{smallmatrix} \right]_\rho^+$, 所以 x 序列至少包含一个 $\left[\begin{smallmatrix} a \\ b \end{smallmatrix} \right], (a, b) \in \rho$ 。

集合 $LR_\rho(\Sigma)$ 中的元素可以没有粘性末端, 这个集合中没有粘性末端的元素就是具有完备形式的 DNA 分子。

小结 从理论上探讨 DNA 分子计算方法, 是理论计算科学的新领域。关于 DNA 分子计算理论的研究, 有利于推动理论计算科学的发展。本文所介绍的常用 DNA 分子操作方法以及 DNA 分子的形式语言描述, 是研究 DNA 分子计算理论的前提。

参考文献

- 1 Adleman L M. Molecular computation of solutions to combinational problems [J]. Science, 1994, 266(5187):1021~1023
- 2 Adleman L M. On Constructing a Molecular Computer [J]. In DNA based computers, American Mathematical Society, 1996 (27):1~21
- 3 Lipton R J. DNA solution of hard computational problems [J]. Science, 1995, 268(5210):542~545
- 4 邓少平, 等. DNA 计算的一些基本问题 [J]. 科学 (Scientific American 中文版), 1996(5):51~54

(下转第 222 页)

个递减函数且 $y'(1) = \frac{m}{n+m} - 1 < 0$, 由此可知在 $[1, +\infty]$ 内 y 是一个递减函数。

当 $n_0 + m \geq e^2$ 时, $y(1) \geq 2$, 因此只需再证明在 $(1, \text{poly}(n))$ 内存在一点 x_0 使得 $y(x_0) < 0$, 那么就可以推断在 $(1, \text{poly}(n))$ 内存在 y 的零点 x_0 , 也即 $\lg(n+x_0m) = x_0$, 于是取 $\delta = \lceil x_0 \rceil$ 即有 $2 \leq \delta \leq \lg(n+m\delta) < \delta+1$ 。

首先, 当 $2m \leq x$ 时, $\lg(n+xm) < \lg(2mx) \leq 2\lg(x)$; 其次, 当 $x \geq e^2$ 时, $2\lg x < x$ 。于是, 取 $x_0 = \max\{2m, e^2\}$ 即可, 显然 $x_0 \in (1, \text{poly}(n))$ 。证毕。□

现在大家普遍认为 $P \neq NP$, 在这个假设下, NP 难优化问题不存在多项式时间的完全算法。因此我们下面将讨论这个问题的近似算法。

3 近似算法

命题 1 对于带测度函数 f 的连通支配集问题实例 $G(V, E)$, 假设 G 的邻接矩阵为 M , 计算矩阵 $M' = M^{f(n)}$, 由 M' 生成新的图 $G'(V, E')$, 其中 $(u, v) \in E' \Leftrightarrow M'(u, v) \neq 0$, 那么 D 是 G 的带测度函数 f 的连通支配集当且仅当 D 是 G' 的带测度函数 $f' = 1$ 的连通支配集。

证明: 由邻接矩阵的定义易得。□

由命题 1, 首先由图 G 得到图 G' , 计算 G' 的连通支配集 D , 同时也就是 G 的带测度函数 f 的连通支配集。于是, 可以直接利用连通支配集的近似算法得到带测度函数连通支配集的近似算法。目前已知的连通支配集的最好近似度为 $\text{Ln}\Delta + 3^{[5]}$, 于是我们可以直接得到下面的算法。

算法 1 带测度函数 $f(n)$ 的连通支配集算法:

输入: 图 $G(V, E)$ 及测度函数 $f(n)$

输出: 图 G 的带测度函数 $f(n)$ 的连通支配集 D 。

步骤:

(1) 计算 G 的邻矩阵 M 及 $M' = M^{f(n)}$ 。

(2) 由 M' 构造图 $G'(V, E')$, 满足:

$(u, v) \in E' \Leftrightarrow M'(u, v) \neq 0$ 。

(3) 对图 G' 应用连通支配集的近似算法, 求 G' 的连通支配集的解 D' 。

(4) 输出 D' 作为 G 的带测度函数 $f(n)$ 的连通支配集的解。

由命题 1 可知, 算法 1 求出的解确实是满足 G 的带测度函数 $f(n)$ 的连通支配集, 并且仍然保持 $\text{Ln}\Delta + 3$ 的近似度。

定理 5 对于带测度函数的连通支配集问题存在近似度

为 $\text{Ln}\Delta + 3$ 的近似算法。

总结 本文提出了带测度函数的连通支配集问题, 这个问题目前还没有人研究过。本文对多种情况研究了它的 NP 完全性, 并给出近似算法。下一步的研究课题是对大于 n^c 的测度函数研究以及把测度函数的情况推广到其他支配集的变种问题。此外, 从实际应用的角度考虑, 对测度函数的不同定义方式也是可研究的课题。

参考文献

- Garey M R, Johnson D S. Computers and Intractability: A guide to the theory of NP -completeness, Freeman, San Francisco, 1978
- Cormen T, Leiserson C, Rivest R. Introduction to Algorithms (Second Edition). the MIT Press, 2002
- Vazirani V V. Approximation Algorithms. Springer, 2001
- Chen Ning, Meng Jie, Zhu Hong. Approximation for Dominating Set Problem with Measure Functions
- Guha S, Khuller S. Approximation Algorithms for Connected Dominating Sets. In: European Symposium on Algorithms, 1996
- Johnson D S. Approximation algorithms for combinatorial problems, J. Comput. System Sci., 1974
- Raz R, Safra S. A sub-constant error-probability low-degree test, and sub-constant error-probability PCP characterization of NP . In: Proc. 29th Ann. ACM Symp. On Theory of Comp., ACM, 1997

• 附录 A

证明, 对正整数 $m, n (1 < n-1 \leq m \leq n^2)$, 存在正整数 δ 满足:

$$1 \leq \delta \leq (n+\delta m)^e < \delta+1, \text{ 且 } \delta \leq 2n^{\frac{2e}{1-e}}.$$

证明: 令函数 $y = (n+xm)^e - x$, 易得 $y' = \frac{me}{(n+xm)^{1-e}} - 1$, y' 是一个递减函数且有零点为 $\frac{(me)^{\frac{1}{1-e}} - n}{m}$, 由此可知 y 是一个凸函数。

又因为 $y(1) = (n+m)^e - 1 > 0$, 于是在 $(1, +\infty)$ 内有且只有 y 的一个零点, 且这个零点必出现在 y 的下降阶段。记这个点为 x_0 , 令 $\delta = \lceil x_0 \rceil$, 于是 $y(\delta) \geq 0$ 也即 $\delta \leq (n+\delta m)^e$ 。再由 $(n+\delta m)^e \leq (n+x_0m)^e = x_0 < \delta+1$ 可得证。

接着要证明 $\delta \leq 2n^{\frac{2e}{1-e}}$, 只需证明 $x_0 \leq 2n^{\frac{2e}{1-e}}$ 即可。

由于 $(n+x_0m)^e = x_0$, 变形得 $x_0^{\frac{1}{e}} - mx_0 = n$, 由 $x_0 \geq 1$ 可推出 $x_0^{\frac{1}{e}-1} - m \leq n$, 于是 $x_0 \leq (n+m)^{\frac{1}{1-e}} \leq 2n^{\frac{2e}{1-e}}$ 。

(上接第 204 页)

- 朱玉贤, 等. 现代分子生物学[M]. 北京: 高等教育出版社, 1997
- Liu Q, Liman W. Dna computing on surfaces[J]. Nature, 2000 (403): 175~179
- Rooß D, Wagner K W. On the Power of DNA- Computing[J]. Information and Computation, 1996, 131(2): 95~109
- Roweis, Sam, Erik W, et al. A sticker based model for DNA computation[J]. Journal of Computational Biology, 1998, 5(4): 615~629
- Wu H Y. An improved surface-based method for DNA computation[J]. Biosystems, 2001, 59(1): 1~5
- Gatterdam R W. Splicing systems and regularity[J]. International Journal of Computer Mathematics, 1998(31): 63~67
- Gibbons A, et al. DNA Computing[J]. Current Opinion in Biotechnology, 1997, 8(1): 103~106
- Head T, et al. Computing with DNA by operation on plasmids[J]. Biosystems, 2000, 57(2): 87~93

- Maley C, et al. DNA computation: Theory, practice and prospects[J]. Evolutionary Computation, 1998, 6(3): 201~229
- Sakakibara Y. DNA computers: A new computing paradigm[J]. Journal of Photopolymer Science and Technology, 1998, 11(4): 681~686
- Wu H Y. An improved surface-based method for DNA computation[J]. Biosystems, 2001, 59(1): 1~5
- Takenaka Y, Hashimoto A. A proposal of DNA computing on beads and its application to SAT problem. In: 7th International Meeting on DNA Based Computers, 2001. 331~339
- 邹海明, 等. 形式语言、自动机和语法分析[M]. 武昌: 华中工学院出版社, 1985
- Kari L, et al. At the Crossroads of DNA Computing and Formal Languages: Characterizing Recursively Enumerable Languages by Insertion-Deletion Systems[A]. In: Proc. of 3rd DIMACS Workshop on DNA- Based Computers, Philadelphia, June 1997. 318~333