

计算科学的新领域:DNA 计算(Ⅲ)

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 DNA 剪接计算模型的文法结构和剪接计算方法,探讨了不同 DNA 剪接计算模型的计算能力,证明了所有图灵机可计算的函数理论上都可以通过 DNA 剪接计算模型来计算。

关键词 DNA 分子重组, DNA 剪接操作, DNA 剪接计算模型

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

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 mainly introduce the grammar structures and the computation methods of DNA splicing model, discuss the computational capacity of several DNA splicing models, and prove the functions which can be calculated by Turning machine will also be work out by DNA splicing model in theory.

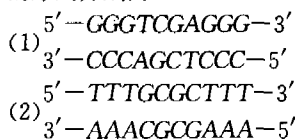
Keywords DNA recombination, DNA splicing operation, DNA splicing model

1 引言

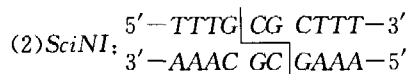
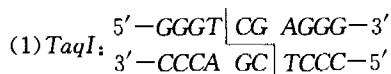
DNA 粘接计算模型是根据 DNA 分子的结构及其粘接特性抽象出的计算模型,由于 DNA 粘接计算模型只能通过粘接操作来完成计算,因此粘接计算模型的计算能力是有限的^[3,4,8]。应用 DNA 分子重组技术,可以在适当的位置对 DNA 分子进行切割,然后重新粘接组合成新的 DNA 分子,这一技术增强了对 DNA 分子的操作能力。DNA 分子剪接操作是抽象地模拟试管中的 DNA 分子进行重组的操作方法^[1,2,5,9,10]。DNA 剪接计算模型是模拟 DNA 分子重组技术抽象出的计算模型,是基于 DNA 分子剪接操作的计算模型。DNA 剪接计算模型的计算过程就是 DNA 分子进行重组的过程^[11~16]。

2 DNA 分子重组技术

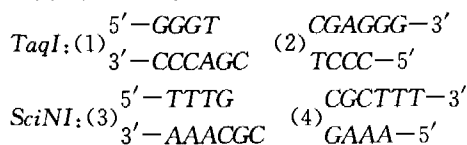
DNA 分子重组包括两个过程:(1)两个 DNA 双链分子在限制性核酸内切酶识别的位置上被分别切割。(2)在连接酶的作用下重组粘接,形成新的 DNA 分子。假设两个 DNA 分子的序列分别为:



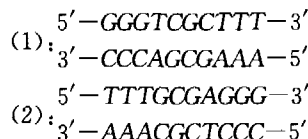
TaqI 和 SciNI 是限制性核酸内切酶。对于粘性末端“CG”的切割,TaqI、SciNI 的切割点都在 5'端,它们的切割位置分别为:



切割后分别形成以下四种具有粘性末端的 DNA 短链:



从切割后的粘性末端可以看出,短链(1)与短链(2)或者短链(1)与短链(4)按照碱基互补原则可以形成 DNA 分子。短链(3)与短链(2)或者短链(3)与短链(4)也可以形成 DNA 分子。这样,剪接操作之后,试管中 DNA 分子的形式,除了原来的两种之外,还可能增加以下两种:



剪接操作是 DNA 分子重组的数学抽象。由于 DNA 分子的碱基互补特性,分子的上下链是一一对应的,为了简便,剪接操作只表示分子上链的变换,通过互补关系便可推出下链的分子序列。引入符号:字母表 Σ 以及 Σ 上的两个字符串 x, y (分别表示 DNA 分子 x 和 DNA 分子 y),剪接规则 r 是 Σ^* 上的形式为 $\alpha_1 \# \beta_1 \$ \alpha_2 \# \beta_2$ 的字符串,其中 $\alpha_1, \beta_1, \alpha_2, \beta_2 \in \Sigma^*$, $\$, \# \notin \Sigma$, $\$$ 和 $\#$ 仅作为标记符号, $\$$ 表示两个字符串的分隔符, $\#$ 表示字符串的切割位置。

依照剪接规则 r ,剪接字符串 x 和 y 的过程分为两步:

(1)按照剪接规则 r 规定的切割位置分别切割字符串 x, y 。

(2)将 x 的前缀与 y 的后缀粘接,同样可以将 y 的前缀与 x 的后缀粘接,形成两个新的字符串。

例如,两个字符串 $x = x_1 a_2 \beta_1 x_2$ 与 $y = y_1 a_2 \beta_2 y_2$,按照剪接规则 $r = a_1 \# \beta_1 \$ a_2 \# \beta_2$,产生了两个新的字符串 $w = x_1 a_1 \beta_2 y_2$, $z = y_1 a_2 \beta_1 x_2$ 。字符串 x 和 y 的剪接操作表示为 $(x, y) \Rightarrow (w, z)$ 。如果剪接模型是单输出的(只需要一个剪接结果),则默认剪接结果为 x 的前缀与 y 的后缀粘接所产生的新字符串,表示为 $(x, y) \mapsto w$,也可以直接表示 $(x_1 a_1 | \beta_1 x_2, y_1 a_2 | \beta_2 y_2) \mapsto x_1 a_1 \beta_2 y_2$ 。如果不是特别指出,本文所讨论的剪接计算模型是单输出的。剪接操作如图1所示。

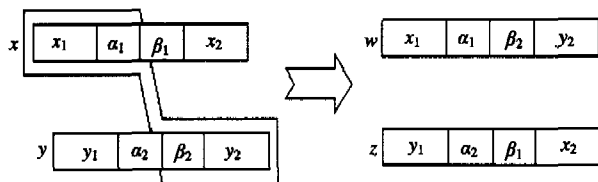


图1 DNA剪接操作

两个字符串 (x, y) 的单输出剪接操作的结果表示为:
 $\varphi((x, y), R) = \{w \in \Sigma^* \mid (x, y) \mapsto w, r \in R, x, y \in \Sigma^*\}$ 。
 其中, R 是剪接规则的集合。

3 DNA剪接计算模型

DNA剪接计算模型是基于剪接操作的文法产生系统。

定义1 DNA剪接计算模型为四联体 $\gamma = (\Sigma, T, A, R)$, 其中:

- (1) Σ 是字母表;
- (2) $T \subseteq \Sigma$ 是终结符号;
- (3) $A \subseteq \Sigma^*$ 是开始符号的集合;
- (4) $R \subseteq \Sigma^* \# \Sigma^* \$ \Sigma^* \# \Sigma^*$ 是剪接规则集合。

根据 Chomsky 谱系的文法类型来确定剪接规则的类型,相应地也就确定了剪接操作的类型。比如, $R \in REG$ 用作剪接规则(REG 表示正规文法产生的语言族),则剪接操作就称为是 REG 剪接或正规剪接类型。设两个语言族 FL_1 、 FL_2 ,语言族 FL_1 在 FL_2 类型的剪接规则下产生的语言表示为: $\Psi(FL_1, FL_2) = \{L(\gamma) \mid \gamma = (\Sigma, T, A, R)\}$,其中, $A \in FL_1$, $R \in FL_2$ 。如果语言族 FL_1 在 FL_2 类型的剪接规则下的封闭的,可表示为: $\Psi(FL_1, FL_2) \subseteq FL_1$ 。

4 简单剪接计算模型

简单剪接计算模型是在理想条件下发生一次剪接操作的单输出系统。

定理1 如果语言族 $FL_1 \subseteq FL'_1$, $FL_2 \subseteq FL'_2$,那么 $\Psi(FL_1, FL_2) \subseteq \Psi(FL'_1, FL'_2)$ 。证明:从简单剪接计算模型的定义,可以直接得到这个结论。

定理2 如果在语言族 FL_1 中所有语言与符号的连接都是封闭的,那么存在语言族 FL_2 ,使得 $FL_1 \subseteq \Psi(FL_1, FL_2)$ 。

证明:设 $L \subseteq \Sigma^*$, $x \in L$, $L \in FL_1$,符号 $a \notin \Sigma$ 。

令 $L' = L\{a\}$, $xa \in L'$ 。

由已知条件中语言的封闭性可知: $L' \in FL_1$ 。

构造剪接系统 $\gamma = (\Sigma', T, A, R)$ 。其中 $\Sigma' = \Sigma \cup \{a\}$, $A \in FL_1$, $T \subseteq \Sigma'$, $R = \{\# a \$ a \# \} \in FL_2$ 。

将字符串按照剪接规则 $\# a \$ a \#$ 进行剪接,剪接结果为: $(x|a, ya|) \mapsto x$ 。即 $L(\gamma) = \{x \mid xa \in L', L' \in \Sigma'\}$ 。

由此可得 $L = L(\gamma)$ 所以有 $L \in \Psi(FL_1, FL_2)$,即 $FL_1 \subseteq \Psi(FL_1, FL_2)$ 。□

5 迭代剪接计算模型

在试验过程中,试管中第一次剪接的结果有可能在限制内核酶和连接酶的作用下继续进行剪接操作,称为有迭代的剪接操作。迭代剪接操作表示为 $\tau = (\Sigma, R)$,其中 Σ 是字母表, $R \subseteq \Sigma^* \# \Sigma^* \$ \Sigma^* \# \Sigma^*$ 是规则集合。

语言 $L \subseteq \Sigma^*$,迭代剪接操作 $\tau = (\Sigma, R)$ 的结果为:

- (1) $\tau^0(L) = L$;
- (2) $\tau^{i+1}(L) = \tau^i(L) \cup \tau(\tau^i(L))$, $i \geq 0$;
- (3) $\tau^*(L) = \bigcup_{i \geq 0} \tau^i(L)$ 。

$\tau^0(L)$ 可直接表示为 $\tau(L)$,由迭代剪接操作结果可知 $\tau^1(L) = L \cup \tau(L)$ 。语言 L 在迭代剪接操作 $\tau = (\Sigma, R)$ 下是封闭的,剪接结果 $\tau^*(L) \subseteq L$ 。

迭代剪接计算模型是基于单输出迭代剪接操作的文法产生系统。

定义2 迭代剪接计算模型为四联体 $\gamma_{di} = (\Sigma, T, A, R)$,其中:

- (1) Σ 是字母表;
- (2) $T \subseteq \Sigma$ 是终结符号;
- (3) $A \subseteq \Sigma^*$ 是开始符号的集合;
- (4) $R \subseteq \Sigma^* \# \Sigma^* \$ \Sigma^* \# \Sigma^*$ 是规则集合。

由迭代剪接计算模型产生的语言记为 $L(\gamma_{di}) = \tau^*(A) \cap T^*$ 。

设两个语言族 FL_1 与 FL_2 ,语言族 FL_1 在 FL_2 类型的迭代剪接规则下产生的语言族表示为: $\Psi_{di}(FL_1, FL_2) = \{L(\gamma_{di}) \mid \gamma_{di} = (\Sigma, T, A, R)\}$ 其中, $A \in FL_1$, $R \in FL_2$,称这种迭代剪接计算模型为 (FL_1, FL_2) 型的。

定理3 任意一个递归可枚举型语言(简称为 RE), $L \in RE$, $L \subseteq T^*$,可写成 $L = \Psi_{di}(FIN, REG) \cap T^*$ 的形式,其中, FIN 表示有限语言, REG 表示规语言。

证明:设 $G = (N, T, S, P)$ 为 0 型文法。

构造迭代剪接系统 $\gamma_{di} = (\Sigma, T, A, R)$,其中 $\Sigma = N \cup T \cup \{B, B', S_0, E, C\} \cup \{E_a \mid a \in N \cup T \cup \{S_0\}\}$ 。符号 B, B' 作为剪接字符串的最左端的标记。 E, E_a 符号作为剪接字符串的最右端的标记。符号 S_0 用于标记文法 G 中被模拟推导的字符串的开始位置。

$A = \{BS_0SE, BC, CE\}$
 $\cup \{CE_a, B'aC \mid a \in N \cup T \cup \{S_0\}\}$
 $\cup \{CvE \mid u \rightarrow v \in P\}$

规则集 R 包含以下规则:

1. 模拟规则:模拟产生式 $u \rightarrow v \in P$ 。
- (1) $Bw \# uE \$ C \# vE$,其中 $w \in (N \cup T \cup \{S_0\})^*$ 。
2. 转换规则:将字符串中的 u 移到右端。
- (2) $Bw \# aE \$ C \# E_a$,其中 $a \in N \cup T \cup \{S_0\}$, $w \in (N \cup T \cup \{S_0\})^*$ 。
- (3) $B'a \# C \$ B \# wE_a$,其中 $a \in N \cup T \cup \{S_0\}$, $w \in (N \cup T \cup \{S_0\})^*$ 。
- (4) $B'w \# E_a \$ C \# E$,其中 $a \in N \cup T \cup \{S_0\}$, $w \in (N \cup T \cup \{S_0\})^*$ 。
- (5) $B \# C \$ B' \# wE$,其中 $w \in (N \cup T \cup \{S_0\})^*$ 。

(下转第184页)

