

技术与方法

核酸序列特征区域的计算机判定

张懋弧 郭礼和

(中国科学院上海细胞生物学研究所)

提 要

本文叙述了一个称作“SPECIAL REGION”的计算机核酸序列分析程序, 该程序具有判定核酸序列中的诸如“回交”、“对称”、“重复”、“丰富”等特征区域的功能。程序用 Apple Pascal 语言编写, 在 Apple II_e 计算机上运行。

随着计算机技术应用的的发展和核酸序列分析工作的不断深入, 人们已经开发了用于各种目的的分析核酸蛋白质等的计算机程序^[1-5]。这其中包含有同源分析、限制性内切酶位点的判定、各个相邻片段间的自动连接、核苷酸序列和氨基酸序列之间的自动翻译, 以及 DNA、RNA 分子二级结构的预测等。

我们在已拥有上述计算机分析程序基础上, 考虑到判定 DNA、RNA 序列中的“回文”、“对称”、“重复”、“丰富”等某些具有特征的区域, 是核酸序列数据分析的一个重要方面, 它既能从已知序列的特征区域寻找。总结出相对应的生物功能, 又能从预期的生物功能出发, 预测和改造序列。为此我们拟开发一个名为“Special Region”的分析程序, 以便形成一套更为完整的核酸序列分析系统。

系统配置与编程原则

本程序针对 APPLE II_e 微型计算机而设计。用 APPLE PASCAL 语言编写, APPLE UCSD PASCAL 1.1 版本的编译器编译。

一、配置

APPLE II_e (6502 CPU, 64kb RAM) 微

型计算机;

APPLE PASCAL 系统主磁盘;

Z-80 卡;

80 列/ 64kb RAM 卡;

两只 5"(1/4) 软磁盘驱动器和驱动器接口卡;

EPSON FX-80 打印机和打印机接口卡。

二、方法

程序的开发遵循下述原则:

1. 尽可能给程序使用者以方便

1) 程序功能以“菜单”(MENU) 形式给出 (见图 1), 使用者可按需要随意选择;

2) 增加“会话”功能, 在程序执行过程中, 给使用者以适当的指导与提示;

3) 经分析而得到的数据输出去向 (监视器, 打印机或数据软磁盘), 可自行选择; 数据输出格式, 尽可能简单易读;

4) 程序运行中所用参数, 既可由使用者自行设定, 亦可用在程序中预先设定的参数值。

2. 以大写字母 (A、G、C、T、U) 表示已确知的核苷酸符号, 以小写字母 (a, g, c, t, u) 表示尚未确知的符号。数据经分析输出时, 仍

* : * : * : * : * : * : * : * : * : * : * : * : * : * : * : *	
FUNCTION	RESPONSE
Palindrome	P
Symmetry	S
A-T Rich Region	A
C-G Rich Region	C
Repeated Subsequence (Forward)	F
Repeated Subsequence (Reverse)	R
Exit	E

Press The Key According To The Function Selected

图 1 程序功能以“菜单”形式给出

以相应的符号表示之。

3. 既考虑核苷酸的直链形序列，亦考虑其环状形序列；既考虑 DNA，亦考虑 RNA。

4. 有些程序^[2,4]，专门编写一个数据输入、修改过程 (PROCEDURE)，它虽然强调了数据文件 (DATA FILE) 的格式化，却限制了组织数据文件时的灵活性。考虑到 APPLE PASCAL 系统的文本编辑器 (EDITOR) 具有极强的功能，它包括了一整套一般文本编辑所常用的诸

如插入、删除、修改、寻找、存盘等指令，操作灵活，使用方便。决定借助编辑器通过计算机键盘来完成核酸序列的数据输入。目前市场上已有一种名为 SONIC DIGITIZER 的核酸序列自动输入装置，它与计算机相连，可直接从凝胶电泳片上，通过一枝声笔 (SONIC PEN) 自动完成核酸序列的输入，以形成数据文件。图 2 表示了核酸序列数据文件和分析程序间的相互关系。

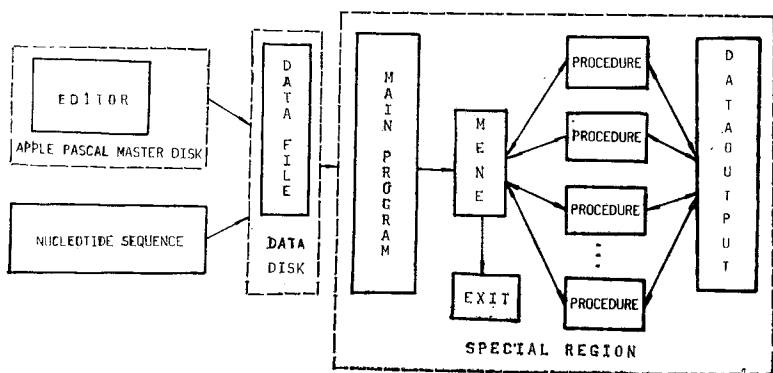


图 2 数据文件和分析程序各过程间的相互关系

程序功能和分析结果

一、程序功能

本程序的编写充分体现了 PASCAL 语言的特点——模块结构化。每一个过程描述一个与之相对应功能的分析算法。允许各过程间的相互递归调用。

1. 回文结构 (PALINDROME)

回文结构的定义是：无论将序列顺读或倒读，顺序相同，即形如“ABCDCBA”。在此，预置回文结构的长度需大于或等于 7 个核苷酸，而上限不定。凡符合此条件者，即调用输出子过程。

2. 对称结构 (SYMMETRY)

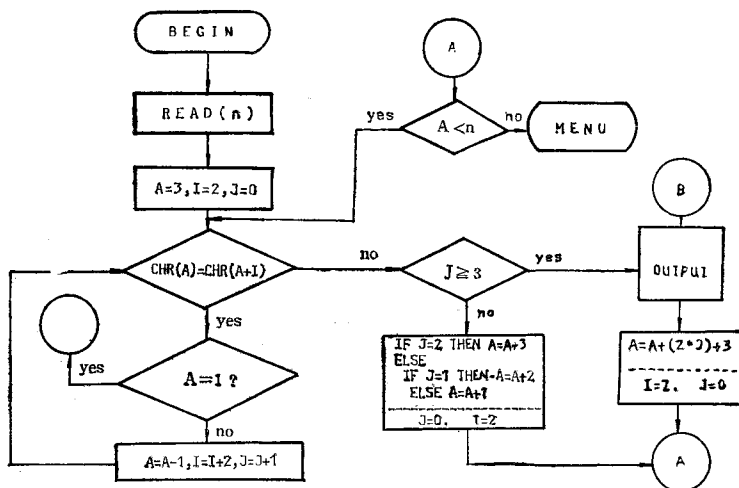


图 3 回文结构分析算法流程简图

N: 序列长度, A: 判别位置, I: 判别位置增量, J: 符合轴对称的次数

定义类似于回文结构,但在所读序列中缺少一“轴”,以“ABCCBA”的形式出现。长度规定为大于或等于 6 个核苷酸。输出格式同回文结构。

限于篇幅,图 3 仅以回文结构为例,给出此过程的分析算法流程简图。

3. 重复序列(REPEATED SUBSEQUENCE) 重复区域分为正向重复 (FORWARD) 和

反向重复 (REVERSE)。正向重复定义为在同一核苷酸序列中,凡长度在 7 个核苷酸以上,能两次或多次重复,且同向出现的相同子序列;而反向重复定义类似于正向重复,只不过当第二次或以后多次出现的重复子序列的读向与第一次读向恰好相反。需分析的起止位置等参数可自行设定。图 4 给出了反向重复序列的分析算法流程简图。

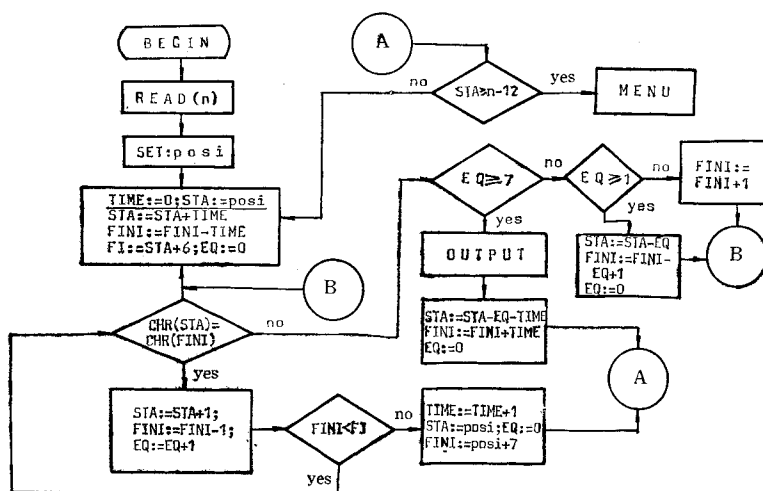


图 4 反向重复序列分析算法流程图

n: 序列长度; POSI: 开始判别位置; TIME: 判别次数; STA: 链首位置; FINI: 链尾位置; EQ: 相符次数

4. 丰富区域 (A-T/C-G RICH REGION)

丰富区域分为 A-T 丰富区和 C-G 丰富区。遵循 QUEEN 和 KORN^[6] 的原则, 在连续 8 个核苷酸中有 6 个或 6 个以上是 A 或 T (C 或 G) 的即为 A-T (C-G) 丰富区。也就是说, 丰富区的最小长度为 6 个, 而上限不定。同时, 在输出的 A-T (C-G) 丰富区结果中, 会自动地将第一个或最后一个是 C 或 G (A 或 T) 的核苷酸删除。通过计算还可分段求得 A-T 或 C-G 在整个核苷酸序列中所占的比例, 以了解核苷酸的分布情况。

5. 两个具有同源序列的格式化显示 (DISPLAY WITH FORMAT)

原有的同源性分析程序, 仅指出了同源区域, 而未能将其组织成一个格式化的排列显示。为此对这一点作了补充。在显示中, 同源区域以“*”表示, 缺省区域为“-”, 并可分段求得同源程度。

二、结果

所开发的程序, 对 DNA 序列 A₃₆₄A 和 P_{CBYI} 分别作了分析。限于篇幅, 仅给出部分分析结果, 以说明分析输出格式。结果中指明了

表 1 DNA 序列 A₃₆₄A 和 P_{CBYI} 的部份分析结果

a. 回文结构分析例子

SPECIAL REGION		VERTION 10/1/85	

Sequence : A364A		Configuration:Linear	Length: 2001 bp
Type:	PALINDROME		
=====			
BEGIN	END	LENGTH	SEQUENCE
91	97	7	TACCCAT
161	171	11	TATAATAATAT
197	205	9	ATGTATGTa
251	257	7	TGCACGT
357	367	11	CGCGAGAGCGC

b. 正向重复序列输出例子

SPECIAL REGION			VERSION 10/1/85	

Sequence: A364A		Configuration: Linear		Length: 2001 bp
Type : REPEATED SEQUENCE (FORWARD)				
=====				
BEGIN	END	LENGTH	REPEATED POSITION	SEQUENCE
15	21	7	1527-----1533	TGATATT
			1647-----1653	
47	43	7	1152-----1158	GAGATAC
40	46	7	1378-----1384	ACACAGC
65	73	9	399-----407	TTTACAGA
			524-----532	
97	104	8	1799-----1806	TCATTGAA

c. A-T 丰富区输出例子

SPECIAL REGION		VERSION 10/1/85	

Sequence: A364A		Configuration: Linear	Length: 2001 bp
Type: A-T RICH REGION			
=====			
BEGIN	END	LENGTH	SEQUENCE
13	21	9	AATGATATT
99	108	10	ATTGAATTTT
103	111	9	AATTTTGAA
143	156	14	ATAGTATATTGAA
159	176	18	TGTATAATAATATAGT
200	208	9	TATGTAITTT

d. 两具有同源序列的格式化输出

TITLE:A364A/Pcby1

```

      10      20      30      40      50      60      70
CTGCAGCTTC TCAATGATAT TCGCCTACGC TTGAGGAGA TACAGCCTAA TATCCGACAA ACTGTTTAC
CTGCAGCTTC TCAATGACAT TCGAACACGC TTGAAGAGA TTCAGCCTGA TATTCGACAA ACTGTTTAC
*****
      80      90     100     110     120     130     140
AGATTACGA TCGTACTGT TACCCATCAT TGAATTTGA ACATCCGAAC CTGGGAGTTT TCCCTGAAAC
AGAATTACGG TCGTACTGGT TACCCATCAT TGGATTTTGA ACATCCGAAT TTGGGAGITT TACCTGAAAC
*****
     150     160     170     180     190     200     210
AGATAGTATA TTGAACCTG TATAATAATA TATAGTCTAG CGCTTIACGG AAGACAA-TG TATGTATTTC
GGATAGTATA TTGAATCTG TATAATAACA TATAGTCTAA CACTTIACGG AAAACAAATG CATGTGTCCA
*****

```

特征子序列的起、止位置，长度以及子序列本身。

结果表明，程序运行速度快，5kb 长度的核苷酸序列判定其回文结构仅需 1 分多钟。因具有会话功能加以提示，故无计算机实践经验者也可使用。同时，程序易于移植，在 APPLE II PLUS 和紫金 II 型机上运行，均获通过。

- [2] Roger Staden: *Nucl. Acids Res.*, 10, 4731, 1982.
- [3] Brian Fristensky, John Lis et al.: *Nucl. Acids. Res.*, 10, 6451, 1982.
- [4] Bruno Malthiery et al.: *Nucl. Acids Res.*, 12, (1), Part 2, 569, 1984.
- [5] 何东明等:《生物化学杂志》, 1, 81, 1985.
- [6] Queen, C. L. et al.: *Methods in Enzymology*, 65, 595, 1980.

[本文于 1986 年 6 月 4 日收到]

参 考 文 献

- [1] James Pustell et al.: *Nucl. Acids Res.*, 10, 51, 1982.

科技消息

肺癌病人呼气中的挥发性有机化合物

早期诊断肺癌用紫外线辅助纤维支气管内窥镜术和细胞学检查法。由于费用昂贵和系创伤性诊断法，难于用作大规模筛选。因此须寻求其他的早期检测方法。应用特殊的呼气收集技术及电子计算机辅助气相色谱/质谱法(GC/MC)，已证明肺癌病人呼气中丙酮、甲基乙基酮及正丙醇等若干挥发性有机化合物含量高。用 12 个肺癌病人的样品和 17 个正常人对样品 GC/MC 描绘图谱经电子计算机统计，发现具有显著

性差异。根据所测得 GC/MC 的数据表明，呼气分析是非创伤性早期诊断肺癌的可行方法，且费用较低，可用于大规模筛选。应作进一步的研究和改进。

[*Clin. Chem.*, 31(8), 1278, 1985.

安徽医科大学临床肿瘤研究所陈培恩摘译，
上海第二医科大学李立群校]