第 27 卷 第 1 期 华侨大学学报(自然科学版) Vol. 27 No. 1
2006 年 1 月 Journal of Huaqiao University (Natural Science) Jan. 2006

文章编号 1000-5013(2006)01-0105-03

BYSJ 软件设计及在生物信息中的应用

张光亚 郑兰香 方柏山

(华侨大学材料科学与工程学院,福建 泉州 362021)

摘要 针对生物信息处理过程中难于找到相关程序进行序列分析的问题,自行设计一套数据处理软件. 经实际应用,软件可用于处理生物数据库 Transterm 中获得的海量原始数据. 文中简介 Transterm 数据库中所需信息的格式,并说明 BYSJ 软件的设计目标、实现原理、源代码,以及软件的功能及使用.

关键词 BYSJ 软件,数据处理,Transterm 数据库,生物信息学

中图分类号 Q 811.4; TP 311.1 **文献标识码** A

生物信息学以生物大分子为研究对象,以计算机为工具,运用数学和信息学的观点、理论和方法,研究生命现象和分析呈指数级增长的生物信息数据的一门科学[1,2]. 目前,生物信息处理的基本方法是通过计算机网络完成对数据库的搜索,然后应用相关软件对获取的数据进行分析. 对于特定的质询,有时很难找到相关的应用程序. 所以对于大量分析工作而言,自己有针对性地设计一些小软件就显得更有效和十分有必要了. 本文针对从 Transterm 数据库[3]里获取的基因序列信息设计程序,用来完成对该序列的分析及相关操作. 同时设计一个人机交互式界面,将序列提交给各个程序,并将结果整理清楚后交给最终用户.

1 编制原理和源代码设计

BYSJ 软件的编制和运行的软硬件环境要求不高,只要是 32 位以上的 Windows 系统都可运行,且要求系统安装有 Visual C^{++} 6.0 软件. 程序编制语言为中文版 Visual C^{++} 6.0.

1.1 设计目标

(1) 按照 N_c 值大小将所有基因序列进行排序,以区分高表达和低表达基因. N_c 是有效密码子数量(Effective Number of Codons),即基因中使用的有效密码子数量. 它反映了基因中密码子使用偏移的程度,表达水平高的基因被认为是有更大密码子使用偏移[4]. 排序完成后,位于两端的基因分别是高表达和低表达基因,然后根据取样多少选择取两端基因序列. (2) 统计起始密码子和终止密码子侧翼序列各位点的 4 种核苷酸的数量. (3) 计算各位点核苷酸出现的频率 P_i. 然后计算 W 值,有

$$W = \Sigma (P(i) - 0.25)^2 / 0.25. \quad (1)$$

在式(1)中,$i = A, T, C, G$; $P(i)$ 表示 4 种核苷酸出现的频率,参数 W 代表保守性.

1.2 设计要点

(1) 要处理的.txt 文件里,每条记录包含 Locus # CDS, Acc_No, Initiation Start, Term. Stop, Len, GC3, N_c, CMB 和 GI 共 9 个参数,如图 1 所示. 分别定义 5 个数组 3 个变量,用来存储这 9 个参数. 即 struct bct{ char Locus[15]; char Acc_No[11]; char start[34]; char Term[24]; long int Len; float GC3; float N_c; float CMB; char GI[15];}; bct Aquiaeol[40 000];使每条记录以结构体的形式存储,并指定 Aquiaeol[ ]为 bct 类型的结构数组,用来存储每条记录. (2) 数组需要预先分配静态存储空间. 目前所知的生物中,基因序列数目最多不超过 4 万条,所以预先为结构数组 Aquiaeol[ ]分配 40 000 ×size of

收稿日期 2005-10-09

作者简介 张光亚(1975-),男,讲师,博士研究生,主要从事酶与生物信息学的研究. E-mail:zhgyghh@hqu.edu.cn

基金项目 国家自然科学基金资助项目(20276026);华侨大学科研基金资助项目(03HZR7)

Locus # CDS	Acc _ No	Initiation Start				
AB000095 # 2	AB000095	CCCCCCTGGGGAGGAAGGCGATGGCCCCTGCGA				
AB000099 # 2	AB000099	TAACTCACCCATGTGTGTCCATGTCGTTAATCA				
AB000114 # 2	AB000114	ACAGGAAAAAAAAAAAGAAGATGGGTTTTTTAA				
AB000115 # 2	AB000115	ATGGAGGTAGCATTGAAGATATGGTTGAAAGAT				
Term	Stop	Len	GC3	N_c	CMB	GI0
CCGGCCCCTCTGAGCCTGGGTCT		1 542	0.785	40.1	0.840	GI:2924601
CCCATCAACCTGAAGGCATAAAC		357	0.461	52.5	0.743	GI:3090432
AAATCAAGAATAGCAAGAAACTA		1 266	0.277	49.7	0.694	GI:1769800
GCCCTGCATTTGAGATAAGTTGC		1 242	0.339	52.0	0.701	GI:1769802

图 1 9 个参数记录格式

(bct)字节的存储空间,即 Aquiaeol[40 000]. (3) 利用文件操作流将文件读入,并存储到结构数组 Aquiaeol[ ]里. 用目前效率较高的快速排序算法将 N_c 值按从小到大进行排序,并输出排序后各个 N_c 所对应的记录,利用文件操作流将结果输出到自定义的.txt 文件里. (4) 利用文件操作流将已排序的文件读入,并存储到数组结构 Aquiaeol[ ]里. 输出前总记录的 5 %条记录,并统计起始密码子和终止密码子侧翼序列每位点碱基 A, T, C, G 的数量,并计算碱基出现的频率值 P_a, P_t, P_c, P_g (即碱基占总数($A + C + T + G$)的百分数). 代入式(1)及用同样方法,计算出 W 值并将结果统一输出到自定义的文件里.

源代码包括 Program1, Program2, Program3 共 3 个程序,限于篇幅,未能给出.

2 功能及输出结果

"BYSJ"软件是针对从 Transterm 数据库下载得到的基因序列按要求进行排序,实现基因表达水平从高到低排列,并进行相关统计和研究而设计的. 其界面如图 2 所示. (1) 运行辅助文件. 将文件调入,统计文件里的记录数,并输出统计结果. (2) 运行文件. 重新将文件调入,对文件进行排序,并输出到自己命名的文件里. (3) 运行结果文件. 将排序输出的文件调入,统计文件中前 5 %的记录和后 5 %记录的序列中 A, T, C, G 个数,并进行相关计算. 该软件输出的结果比较整齐,容易将其数据读入到 Exel 文件中,进行作图处理. 该程序输出的部分结果,需要说明的是,结果中 l 值表示包含于原始的基因序列中除 4 种核苷酸以外的其它符号(主要是空格). 因为并不是所有的基因都包含完整的起始和终止密码子的侧翼序列,在缺乏的区域,Transterm 数据库以空格代替,故统计的结果中 l 值一般不为 0. 但在计算 4 种核苷酸出现的频率时,却并不参与计算. 因此对计算结果没有影响.

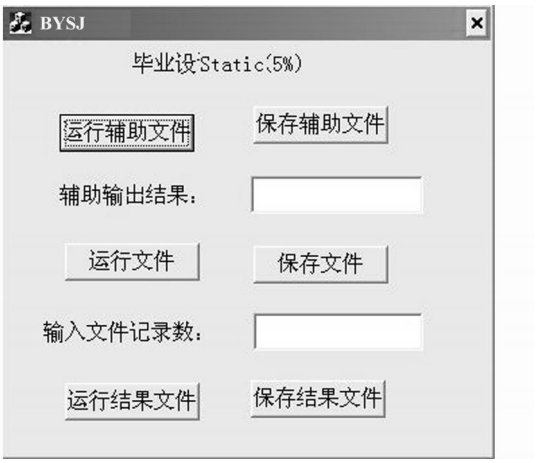

图 2 BYSJ 软件的运行界面

BYSJ 软件的使用应具备 2 个前提条件. (1) 安装有 VC^{++} 软件的 Pentium 微机. (2) 只能运行后缀为(.txt)的文件,而且文件里记录的格式要求完全一致. 有出现空格的地方,要补 0;若文件头有其他字符或文件尾有换行符,要尽可能删掉.

3 实际应用

分别从 Transterm 数据库获取大肠杆菌和啤酒酵母中 9 293 条、6 357 条 mRNA 起始和终止密码子密码子侧翼序列(从 - 20 位到 + 13 位)、基因长度、N_c 和 GC3s 值. 利用上述程序计算后,获得的高表达和低表达基因在起始和终止密码子侧翼序列的核苷酸组成特征与文[5]报道的结果相似,存在差异的地方可能与 mRNA 序列的来源不一样有关. 说明本软件计算结果可信度和准确性高,出错几率很低. 同时,可根据需要对任何从该数据库获得 mRNA 序列进行处理,以获取有关生物基因表达调控、非编码区保守性及其功能等方面的数据信息. 程序运算后部分结果,如图 3 所示. a, t, c, g 分别为各列中 A, T, C,

*G*的个数.

```
1  row is:  a = 147  t = 80  c = 49  g = 36  l = 5  #  Pa = 0.471  Pt = 0.256  Pc = 0.157  Pg = 0.115 0  #  W = 0.303
2  row is:  a = 154  t = 60  c = 67  g = 31  l = 5  #  Pa = 0.494  Pt = 0.192  Pc = 0.215  Pg = 0.099 4  #  W = 0.346
3  row is:  a = 132  t = 73  c = 75  g = 32  l = 5  #  Pa = 0.423  Pt = 0.234  Pc = 0.240  Pg = 0.103 0  #  W = 0.208
4  row is:  a = 162  t = 63  c = 58  g = 29  l = 5  #  Pa = 0.519  Pt = 0.202  Pc = 0.186  Pg = 0.092 9  #  W = 0.414
5  row is:  a = 146  t = 71  c = 60  g = 35  l = 5  #  Pa = 0.468  Pt = 0.228  Pc = 0.192  Pg = 0.112 0  #  W = 0.281
6  row is:  a = 164  t = 71  c = 50  g = 27  l = 5  #  Pa = 0.526  Pt = 0.228  Pc = 0.160  Pg = 0.086 5  #  W = 0.445
7  row is:  a = 164  t = 61  c = 58  g = 29  l = 5  #  Pa = 0.526  Pt = 0.196  Pc = 0.186  Pg = 0.092 9  #  W = 0.431
```

图 3　程序运算后的部分输出结果

4　结束语

我们曾利用从 Transterm 数据库获取的敏捷气热菌的 mRNA 序列的信息(以.txt 格式保存),将其通过"复制"并"粘贴"到 Word 文件中.然后,利用"文字转换为表格"的命令,将其转换为表格文件,并"粘贴"到 Exel 文件中. 利用排序命令实现按照 N_c 值排序,并手工统计其结果,然后计算. 此过程比较繁琐[6].用该软件对上述数据进行处理后,效率大为提高,而计算的结果与原文一致.尽管如此,但本软件对文件记录的要求很严格,目前只能用于对从 Transterm 数据库里获得的格式如上所示的基因序列的进行操作.一旦格式发生改变,BYSJ 软件也得进行相应的改变才能应用.同时,在很多方面的功能仍需要进一步加强,如对起始和终止密码子侧翼序列密码子和氨基酸的统计和更为复杂的计算公式的实现.此外,软件操作界面不甚美观,要求使用的计算机安装过 Visual C^{++},等等.

参　考　文　献

1　邹承鲁. 21 世纪的生命科学[J].生物化学与生物物理进展,2000,27(1):3～5

2　王　玲.基于知识发现的生物信息学[J].生物工程进展,2000,20(3):27～29

3　Grant H J, Oliver R, Peter A S. Transterm: A database of mRNAs and translational control elements[J]. Nucleic Acids Res, 2002, 30(1):310～311

4　钱　韦. 两种植物病源黄单胞菌基因组同义密码子使用分析[J].植物病理学报,2004,34(2):97～106

5　陈颖丽,李前忠. *E. coli* 和 *Yeast* 基因起始与终止密码子邻近序列碱基保守性、关联性的对比研究[J].内蒙古大学学报(自然科学版),2000,31(2):164～167

6　张光亚,方柏山. 敏捷气热菌密码子及 AUG 侧翼序列保守性分析[J]. 华侨大学学报(自然科学版), 2004, 25(2): 192～196

The Design of BYSJ Software and Its Application to Bioinformation Processing

Zhang Guangya　Zheng Lanxiang　Fang Baishan

(College of Material Science and Engineering, Huaqiao University, 362021, Quanzhou, China)

Abstract　In the course of bioinformation processing, it is necessary to design some small software for sequence analysis in answering specific inquiry. BYSJ is just the software designing for processing enormous initial data, which are obtained from Transterm database. The authors describe the format of the file in Transterm database and on this basis, explain the design objective, implementation principle, source code of BYSJ software and also its fanction and usage. This is a successful attempt of software development in the field of bioinformatics.

Keywords　BYST software, bioinformatics, Transterm database, data processing