文章编号：2095-2163(2019)01-0211-03 中图分类号：TP311.13 文献标志码：A

植入(*l*, *d*)模体发现若干算法的实现与比较

胡宏涛，龚逸文
(西安石油大学 计算机学院，西安 710065)

摘 要：模体发现是生物信息学的核心问题之一，对于研究基因表达的调控机制有着极为重要的生物意义。植入(*l*, *d*) 模体发现(Planted (*l*, *d*) motif search, PMS)是模体发现领域中一个广为接受的问题模型。本文主要研究了 4 个基础的算法解决模体发现问题，这些算法可以帮助人们理解模体发现问题。4 个精确算法主要包括：(1) 实现基于候选模体实例字符串深度优先搜索+剪枝思想解决的位点比对的 PMS 问题。(2) 实现基于候选模体字符深度优先搜索+剪枝思想解决的位点比对的 PMS 问题。(3) 实现基于候选模体字符广度优先搜索+剪枝思想解决的位点比对的 PMS 问题。(4) 实现 PMSP 算法。

关键词：模体发现；生物信息学；算法

Implementation and comparison of some algorithms for implantable(*l*,*d*) module discovery

HU hongtao, GONG Yiwen
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 Module discovery is one of the core problems of bioinformatics, which is very important to study the regulation mechanism of gene expression. Implant able (*l.d*) discovery(planted (*l.d*) motif search, PMS) is a widely accepted problem model in the field of body discovery. This paper mainly studies four basic problems of algorithm module discovery. These algorithms can help to understand the problem of template discovery. Four of these precise algorithms mainly include :(1)Implementation of string depth priority search based on candidate template instance + PMS problem of site alignment in cutting technique.(2)Implementation of candidate mode character depth priority search+ PMS problem of site alignment in cutting technique.(3) Implementation of priority search based on candidate template character breadth + PMS problem of site alignment in cutting technique.(4) Implementation of PMSP algorithm.

【Key words】 template discovery; bioinformatics; algorithm

0 引 言

模体发现可以形式化地定义为植入 (*l*, *d*) 模体发现问题(Plant (*l*, *d*) motif search, PMS)。其中，PMS 将模体定义为一个长为 *l* 的模式串 *m*，其以最多 *d* 个位置失配出现于 *t* 条输入序列中，*m* 和其在序列中的出现分别称为一个(*l*, *d*) 模体和实例。给定 *t* 条长为 *n* 的 DNA 序列集合 $D = \{s_1, s_2, \cdots, s_t\}$、*l* 和 *d*，qPMS 的目标是找出 *D* 中所有的(*l*, *d*) 模体。

1 方法描述

在描述模体发现算法之前，首先对本文中要使用到的符号做一个基本的定义，用来规定文中将要使用到的符号。论文中常用的符号对照见表 1。

表 1 文中常用符号对照表

Tab.1 The symbols and corresponding meaning in the paper

符号	名称
t	数据集中序列的数量
n	数据集中每条序列的长度
l	模体的长度
d	模体的最大退化程度
l - mer	长为 *l* 的字符串
D	输入序列的集合，$D = \{s_1, s_2, \cdots, s_{\|D\|}\}$；
pms_1	基于候选模体实例字符串深度优先搜索的 PMS 算法
pms_2	基于候选模体字符深度优先搜索 PMS 算法
pms_3	基于候选模体字符广度优先搜索 PMS 算法
PMS	植入(*l*, *d*) 模体搜索

1.1 基于候选模体实例字符串深度优先搜索的 PMS 算法

首先，对基于候选模体实例字符串深度优先搜

作者简介：胡宏涛(1965-)，男，硕士，教授，主要研究方向：信息系统、计算机网络、人工智能；龚逸文(1994-)，男，硕士研究生，主要研究方向：人工智能、信息系统。

收稿日期：2018-10-17

索的 PMS 算法进行分析。在设计算法的时候用到了深度优先的方式分别对长为 l 的字符串进行操作,最终输出符合要求的模体和模体实例。下面对算法进行描述。

在此算法中需要建立一个分支节点为 $n-l+1$ 的树,树的深度为 t。建立该树之后可以使用深度优先搜索的方式对建立的树进行初始化。在初始化时需注意,在每次深度优先搜索时都需要先对下一条序列中长为 l 的字符串进行判断,看其与前面树中已经存在的字符串是否满足海明距离 $\leqslant 2d$, 如果满足,则证明这些长为 l 的字符串可能是属于同一个模体的模体实例。相反如果不满足条件,则证明这些字符串绝对不可能是属于同一个模体的模体实例,这时就需要采取剪枝的策略来降低程序的时间和空间复杂度,直接减去不需要的字符串并且不再继续遍历。当程序遍历到深度为 t 的时候,证明已经找到了 t 条模体实例,二者之间满足相互之间的海明距离 $\leqslant 2d$, 此时,只需要进一步用位点比对的方式找到候选模体,并对 t 条 $l-$mer 进行判断,如果符合模体发现问题的定义,则对模体和模体实例进行输出。

1.2 基于候选模体字符深度优先搜索 PMS 算法

1.1 节主要是通过对 t 个模体实例进行操作,然后找出符合条件的候选模体,而在 1.2 节中,主要是通过遍历所有的模体(其中加入剪枝算法)找到符合条件的模体。第二种方法更加直接,也能直接找到所有符合条件的模体。

在此算法中,程序使用栈的方式进行模体的查找。首先建立一个深度为 l, 分支数为 4 的树。与1.1节相似,使用深度优先搜索的方式对模体进行遍历找到合适的模体。但与其不同的是,在基于候选模体字符深度优先搜索 PMS 算法中是直接对模体进行遍历而不是对模体实例中的字符串进行遍历。对模体进行遍历的好处在于当完成遍历时,就可以得到所有可能的模体集合而不会遗漏满足条件的模体。在对模体进行遍历的同时,应该特别注意要使用剪枝的方法去除不符合条件的模体。一个长度 $\leqslant l$ 的候选模体与初始序列进行对比, 如果在一个初始序列的某一列中不能满足候选模体与长度等同的初始序列海明距离 $\leqslant d$, 则表明候选模体不符合条件,进行剪枝,最终输出所有符合条件的候选模体。

1.3 基于候选模体字符广度优先搜索 PMS 算法

基于候选模体字符广度优先搜索 PMS 算法与 1.2 节中的 PMS 算法基本思路相同,不同的地方在于上节中采取的是深度优先搜索进行遍历,而本节主要采用广度优先搜索进行遍历。

在本节中,算法的主要思路可以借鉴 1.2 节的内容。需要注意的是,在遍历的过程中需定义一个队列对候选模体进行操作,当某一个候选模体符合海明距离 $\leqslant d$ 的条件时程序将以队列的形式对其进行一系列的操作。

1.4 PMSP 算法

PMSP 算法相比前面 3 种算法有着更高的效率,PMSP 思路如下:对于 s_1 中的每一个长为 l 的字符串 x,都可以生成一个模体集,在这个模体集合里面所有的模体与 $l-$mer x 的海明距离都 $\leqslant d$,这个集合就被称为候选模体集。用候选模体集中的每一个候选模体 y 去和 s_2-s_t 中的每一个 $l-$mer x' 进行比较,判断是否在 s_i 中存在一个与 y 的海明距离不大于 d 的字符串,如果在 s_2-s_t 中的每一个序列中都存在这一个字符串,则表明候选模体 y 是一个潜在的模体,而在 s_2-s_t 满足海明距离为的 $l-$mer 都是潜在的模体实例;否则,从 s_1 中选择下一个 $l-$mer,重新生成其候选模体集,再进行上述判断过程,直到遍历完 s_1 中的所有 $l-$mer 为止。

2 实验比较

通过上一节中对 4 种算法的描述,可以分别实现程序,并且对 4 种算法的运行时间进行记录,比较其运行时间并得出一定的结论。在实验比较的过程中,分别选用了(9, 1)、(11, 2)、(13, 3)、(15, 4)对程序进行测试统计运行时间。程序运行时间的记录数据见表 2。

表 2 模体发现 4 种算法比较

Tab. 2 Comparison of four algorithms for template discovery s

(l, d)	pms_1	pms_2	pms_3	PMSP 算法
(9, 1)	11	159	159	18
(11, 2)	61	2 749	2 812	28
(13, 3)	534	39 595	36 630	331
(15, 4)	-	-	-	13 002

可以看到,在模体发现问题的4种精确算法中,程序实现时间随着(l, d) 的变化算法有着比较大的差异。通过实验,可以得到产生这些变化的原因。

首先,在 pms_1 中对于每一个输入字符串,每次都要加入一个模体实例,如果判断出来一组模体实例符合模体实例互相 $\leqslant 2d$ 的条件,且这组模体实例的总数小于 t,则必须加入 $n-l+1$ 个模体实例继续进行深度优先遍历。通过对比可以得知,pms_1 和 PMSP 算法有着比较相近的时间和空间复杂度,

只是 pms_1 算法在处理相对比较大的问题上没有 PMSP 算法高效。在这里需要注意的是，pms_1 算法并不像其它 3 种算法那样能够完全遍历出所有的模体，可能会遗漏掉一些模体，但是 pms_1 输出的模体一般都是得分较高的模体。

接下来，通过比较 pms_2 和 pms_3 可以发现二者在运行时间上并没有较大的差异，但这并不意味着二者在算法上没有什么区别。在这 2 种算法中分别采用了深度和广度优先的方式对算法进行设计，其中在广度优先中由于程序需要额外在队列中存储比较多的元素，因此必然会占用更大的内存。过大占用内存也证明了 pms_3 算法效率较低。pms_2 和 pms_3 算法运行时占用内存见表 3。

表 3　pms_2 和 pms_3 占用内存对比

Tab. 3　The occupied memory cornparison of pms _ 2 and pms_ 3 memory occupied contrast

算法名称	CPU/s	占用内存/K
pms_2	45	2 188
pms_3	41	157 860

最后，通过对前面 3 种算法的实现，本文提出另一种相对于前面算法来说更加高效的 PMSP 算法。在 1.4 中已经比较详细地介绍了 PMSP 算法的基本思想，相对于前面的算法提出了一种比较新的思路来解决模体发现问题，在这种新的思路下，模体发现问题算法运行时间将大大减小。在 PMSP 算法中，程序可以解决前面算法很难处理的(15，4)问题。

3　结束语

本文主要实现并比较了模体发现的 4 种算法，通过比较可以发现，由于模体发现是生物信息学、计算生物学和计算机科学的挑战问题，因此算法的选择至关重要。对于同样一个问题，选择不同的模体发现算法其程序执行时间可能会出现比较大的差异。这就要求设计者在进行 PMS 算法设计时充分考虑到算法的时间和空间复杂度，从而设计出运行时间更短的 PMS 算法。

论文中存在的不足及改进方法如下：

(1)基于候选模体实例字符串深度优先搜索的 PMS 算法存在问题：当进行位点比对时，并不是只有当模体实例符合一致序列条件的时候得出的模体才满足模体的定义，有可能有一些候选模体不符合一致序列条件，但同样符合模体的条件。改进办法：遍历所有候选模体然后再输出所有满足条件的模体。

(2)PMSP 算法中存在问题。在 PMSP 中，程序主要使用了 1.4 节 PMSP 算法的思路和伪代码对其进行研究，但是由于在编码过程中缺乏对程序的优化，导致程序进行测试时运行的时间过长，在后续的研究中应加以改进。

参考文献

[1] DAVILA J, BALLA S, RAJASEKARAN S. Space and time efficient algorithms for planted motif search[C]// ICCS´06 Proceedings of the 6th international conference on Computational Science. Reading, UK: ACM, 2006:822-829.

[2] ZAMBELLI F, PESOLE G, PAVESI G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era[J]. Briefings in Bioinformatics, 2013, 14(2):225-238.

[3] MRÁZEK J. Finding sequence motifs in prokaryotic Genomes-a brief practical guide for a microbiologist[J]. Briefings in Bioinformatics, 2009, 10(5):525-536.

[4] 霍红卫，林帅，于强，等. 基于 MapReduce 的模体发现算法[J]. 中国科技论文，2012，7(7):487-502.

[5] D'HAESELEER P. What are DNA sequence motif[J]. Nature Biotechnology, 2006, 24(4):423-425.

(上接第 210 页)

3　结束语

本文提出了一种温室智能监测与调控方案，将 STC89C52RC 单片机、NRF24L01 无线发射模块、PT100 温度模块以及控制模块结合在一起，实现了温室的智能监测。在温室中使用本系统，减少了节点功耗，提高了数据监测的精确性，并且由于使用无线传输，扩大了覆盖区域，提高了通信效率和可靠度。为当前的温室大棚监测与调控提供了一种成本低廉的解决方案，为智慧农业的普及打下了基础，具有一定的推广价值。

参考文献

[1] 王纪章. 基于物联网的温室环境智能管理系统研究[D]. 镇江：江苏大学，2013.

[2] 钟新平. 基于单片机的温室大棚环境参数自动控制系统[D]. 南宁：广西大学，2011.

[3] 王磊，李桂香，王元麒. 基于 Pt100 热电阻的温度检测系统设计[J]. 中国仪器仪表，2014(12):33-35.

[4] 杜洋. A/D 转换芯片 ADC0832 的应用[J]. 电子制作，2006(1):44-46.

[5] 朱慧彦，林林. 基于 MCU 和 nRF24L01 的无线通信系统设计[J]. 电子科技，2012，25(4):81-83，91.

[6] 章昕，黄秋，汤彬，等. 智能温度报警系统的研制[J]. 自动化技术与应用，2009，28(7):103-105.