

基于频繁子树挖掘的 DNA 重复序列识别方法

周溜溜, 业 宁, 徐 昇, 严敏利

(南京林业大学 信息技术学院, 江苏 南京 210037)

摘 要: 提出了一种基于频繁子树挖掘策略的 DNA 重复序列识别方法. 绕开了传统的序列比对方式, 将序列按照后缀树结构方式进行组织, 再对后缀树形式做了约减改进, 使其更加适合子树挖掘操作, 最后利用频繁子树挖掘的方法对其进行学习. 算法可以直接识别出满足设定阈值的重复序列, 避免了由短重复体拼接所造成的时间浪费, 设计的“二次识别技术”使得算法对模糊重复体也有着很好的识别效果, 提高了识别完整度. 实验证明: 算法在识别效率性能方面较升, 尤其当识别较长重复体时, 优势体现的更为明显, 同时在识别完整度方面也高度可比.

关键词: DNA 序列; 重复体识别; 频繁子树挖掘

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2011)09-0193-04

Algorithm of Identification the DNA Repeat Sequence Based on Frequent Subtree Mining

ZHOU Liu-liu, YE Ning, XU Sheng, YAN Min-li

(College of Computer Science and Technology, Nanjing Forestry University, Nanjing 210037, China)

Abstract: The proposed algorithm is based on the thinking of the frequent subtree mining repetitive DNA sequences in the body identified. The organization of DNA sequences in the new algorithm is different from with the others; organized a sequence as a tree, so we could avoid alignment as those traditional methods, then improved trees more simple that could be operating by frequent subtree mining, used a kind of algorithm for mining frequent subtree to learn these trees. This new algorithm could find out the repeated sequences which meet the threshold set directly, avoid the wasting of time result of splicing the short sequences. Designed the new technology “secondary identification”, which could find out the fuzzy repetitive sequences, also improved integrity of identification. Experiment show that our method improved the time efficiency compared with mainstream algorithms, especially learning to find out some long sequences and highly comparable on the integrity of identification.

Key words: DNA sequences; repeat sequence identification; frequent-subtree mining

1 引言

21 世纪以来随着生命科学的发展, 科学家们对特定生物的基因序列需求越来越多. DNA 测序工作变的尤为重要, 目前主流测序法是“鸟枪”法^[1]. 由于 DNA 序列存在大量重复体, 容易导致原序列中串行的重复序列被误以为并行的克隆体, 影响拼接精确度. 在生物学中, 重复片段是重要的遗传信息, 它已经被广泛的运用于肿瘤研究, 生物进化和亲子鉴定

等各个方面^[2]. 所以如何让对序列重复体进行快速有效的识别是一项极为重要且有意义的工作. 针对重复体中的散在重复^[3]识别虽然已有算法问世, 但识别精准度、减低冗余方面仍有着较大的发展空间; 且现今主流方法大都基于序列比对, 因此识别效率普遍不高, 因此如何绕开序列比对, 便成为了一种新的算法设计思路. 设想如果把 DNA 重复体视为一种特殊的频繁模式, 将每条 DNA 碎片组织成树型结构, 便可以利用频繁子树挖掘策略, 直接发现满足

收稿日期: 2011-01-15; 修回日期: 2011-02-13

基金项目: 国家自然科学基金项目(30671639); 江苏省自然科学基金项目(BK2009393); 江苏省青蓝工程学术带头人项目

压缩树进行学习. 现通过一个具体实例说明算法的挖掘过程, 例如: 有一待测 DNA 序列为“ATTCG-GCATATTCGAC”, 要求识别出所存在重复体. “鸟枪”法获得两条克隆体, 并将其平均打碎为 3 条碎片, 便可以得出 6 条序列碎片, 假设分别是“ATTCG”、“GCAT”、“ATTCGAC”、“ATTCGGC”、“ATATTCG”、“AC”, 并依次将这些碎片编号为 a, b, c, d, e, f; 按照规则将这些碎片组织为压缩树结构, 如图 3 所示.

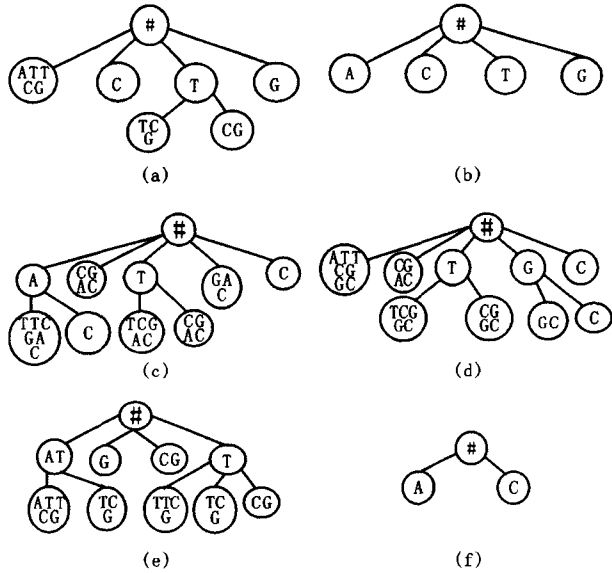


图 3 6 条碎片的压缩树表示

由于原始序列的克隆数为 2, 所以重复体所构成子树在以上六颗压缩树中的支持度阈值至少为 $2 \times 2 = 4$, 利用 FFSM^[15] 算法对 6 棵压缩树做频繁子树挖掘可以获得其所有满足阈值条件 ($m = 2, l = 4, p = 0$) 的重复体分别为“ATTC”、“TTCG”、“ATTCG”.

3.3 ISSM 算法伪代码

ISSM 算法思想分为两部分, 首先是将序列碎片集转化组织为压缩树结构, 然后利用改进后的频繁子树挖掘算法对这些压缩树进行操作得出重复序列集. 下面给出 ISSM 算法的主要步骤如下:

算法 1 ISSM(RD, l, m, p).

输入: 碎片集 RD, 支持度阈值 m , 重复序列最短长度 l , 错误率 p ;

输出: 重复序列集 Res.

- (1) 将 RD 组织成深度直接后缀结构数据 TD
- (2) 再将 TD 进化至压缩树数据集 CTD
- (3) 从 CTD 中每个个体的根结点扫描各条路径一次
- (4) 删除路径长度小于 l 的路径
- (5) 挖掘出支持度不小于 $n * m$ 的频繁子树 FT

- (6) n 为原始序列的克隆数
- (7) 以 FT 作为重复序列字典
- (8) for each $s \in FT$
- (9) 挖掘误差字符数 $\leq \text{length}(s) * p$ 的频繁子树 FuT
- (10) for each $\text{str} \in \text{FuT}$
- (11) 挖掘误差字符数 $\leq \text{length}(\text{str}) * p$ 的频繁子树 FFuT
- (12) end
- (13) end
- (14) 合并 FT, FFuT 为 Res

算法第 1~2 行将碎片集组织成压缩树结构数据, 3~4 行初步删除不满足阈值条件的子树, 5~7 行初步识别出“精确”重复体, 8~13 行以已识别的“精确”重复体做为“字典”, 识别出满足阈值条件的“模糊”重复体, 最后合并两类重复体得出识别结果.

4 实验与分析

实验运行环境为 CPU 主频为 3.2GHz 的 Intel Pentium 4 多线程处理器, 内存 1GB, 硬盘为 250GB, 系统是 Windows XP Professional SP3. 实验采用半仿真数据集, 首先从 Genbank 数据库(<http://www.ncbi.nlm.nih.gov/Genbank/index.html>)中取得三条访问号为 AANG01002301、BAAF01004188、CH689725 的 DNA 序列; 然后随机产生一个指定长度(文中值取为 100bp)的序列 S_0 (各碱基服从均匀分布), 并分别从三条序列随机取得 M 个位置(文中取值为 80)并保证相邻位置间距离不小于 110, 将子串 S_0 按选定位置插入, 同时令每条序列的克隆数为 $n = 3$. 最后利用这 3 条完整序列对 ISSM 算法做性能测试. 文中实验方法主要通过分别调节支持度以及待识别重复序列长度两个主要参数观察算法挖掘完整度和识别效率两方面性能表现, 并与采用拼接的投影拼接算法和基于比对策略的 p -cloud 方法做了比较.

图 4 中通过固定算法中两个参数取值, 改变剩余一个参数值的方法(以下均采用此方式)考察了 ISSM 算法对 DNA 序列 AANG01002031 的挖掘效果. 图 4 的(a)、(b)部分通过改变 p 值进行观察; (c)、(d)部分通过改变 l 值进行观察. 从实验数据得出: ISSM 算法在对重复序列识别效率上要明显优于投影拼接算法和基于比对策略的 p -cloud; 图 4 的(a)、(c)在某些参数点处三种算法识别时间有交叉, 但整体趋势仍然优于另两种方法, 造成交叉现象的原因是序列 AANG01002031 中重复体长度较短, 使得 ISSM 算法省去序列拼接步骤这一重要优势没

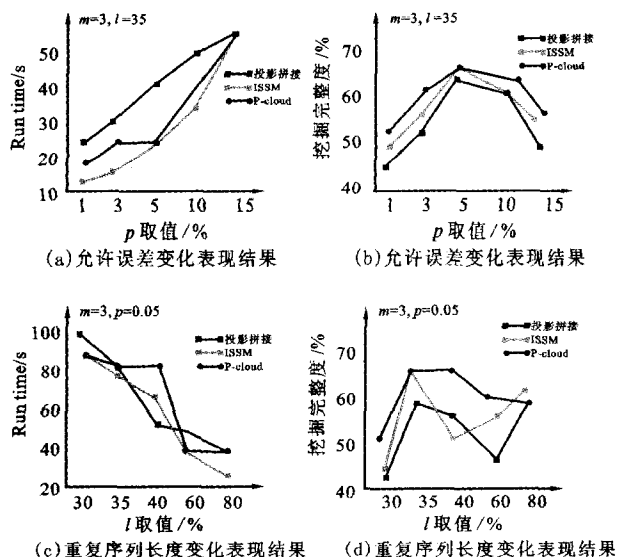


图 4 序列 AANG01002031 的挖掘性能结果

有得到充分体现,而这些交叉点的存在恰恰反映了各种不同重复体识别算法的各自应用范围;同样在图 4 的(b)、(d)中可以得出:ISSM 算法的识别完整度要明显优于投影拼接算法,这是由于采用了“二次识别技术”可以使得遗漏达到尽可能最小,尽管在某些特殊点低于 p-cloud 方法(取值在 40~60 时),但就整体挖掘结果而言,与 p-cloud 方法仍然高度可比。

对序列 BAAF01004188 的实验结果如图 5 所示。采用的实验方法仍然是固定参数法(如上述)。从实验结果分析得出,当重复序列的平均长度达到一定阈值时,ISSM 的挖掘性能与其他算法之间的区分度增大(交叉点较少),如图 5(a)、(c)所示的时间性能图充分体现了分别因略去序列对比和拼接的优势。(b)、(d)仍然给出了挖掘完整度比较结果,ISSM 优于投影拼接同时与 p-cloud 高度可比。综合图 4、图 5 的(b)、(d)部分观察发现三种算法对不同序列操作的完整度达到峰值时 p 、 l 取值有所不同,这是因为重复体平均长度不同而造成的,这也为如何判断 DNA 序列中重复体平均长度提供了新思路。

图 6 所示的是序列 CH689725 的挖掘结果。其表现出的性能说明与图 4、图 5 相似,更进一步验证了 ISSM 相较于另两种算法优势的理论分析。同时可以看出,三张图对识别完整度的表现均非简单的单调关系,所以 ISSM 算法所识别的最佳有效长度视不同参数取值情况有待于进一步确定。从实验表现结果可以得出,除在个别的参数点外,其他各参数点处 ISSM 算法的识别效率要始终优于投影拼接 p-cloud 且识别完整度高于投影拼接算法,也与 p-

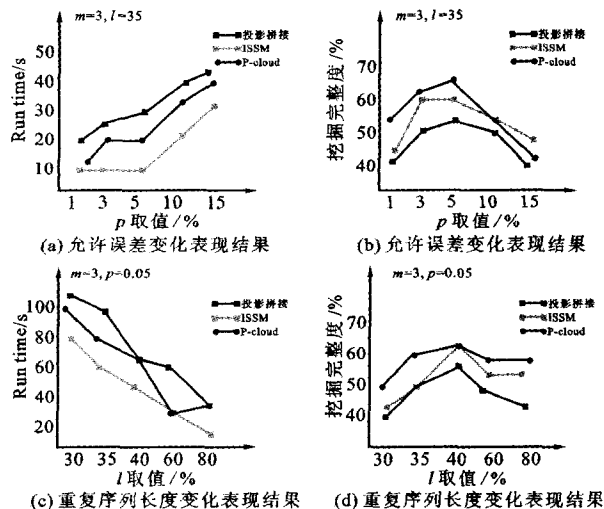


图 5 序列 BAAF01004188 的挖掘性能结果

cloud 高度可比,结果与分析表明文中提出的 ISSM 算法在生物信息学的重复序列识别这一基础研究中是一款有着较强实用性的方法。

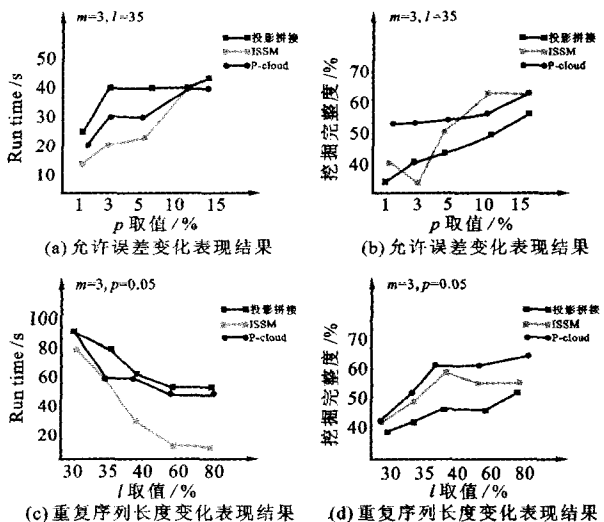


图 6 序列 CH689725 的挖掘性能表现结果

5 结束语

文中提出的 ISSM 算法首先将 DNA 序列碎片组织成压缩树,再利用频繁子树挖掘方法对数据集进行操作,识别出满足阈值条件的重复序列集,理论分析了其由于省略序列比对和拼接过程所以使得新算法在识别较长重复体时有着较大的效率优势,设计的“二次识别技术”保证了识别完整度。并利用 Genbank 数据库中三条 DNA 序列所形成的半仿真数据通过实验证明了理论分析。ISSM 识别完整度要全面优于投影拼接算法同时与 p-cloud 高度可比;但发现完整度曲线表现与变量并不是全局单调

(下转第 201 页)

此基础上提出了基于A*算法的动态组合方法,同时解决了A*算法在应用于Web Services组合时出现的多服务问题.通过实验分析确认该算法适用于多输入的Web Service动态合成情况,并且具有较好的算法处理时间.下一步的研究是进一步改善算法估价函数,将该算法生成的组合方案转换为可执行的代码并提高转换和执行效率.

参考文献:

- [1] Yuri Babich. Composition of web services and Qos aspects[Z]. seminar;Data Communication and Distributed Systems 2003/2004.
- [2] Sheshagiri M, DesJardins M, Finin T. A planner for composing services described in DAML-S[C]// Proceedings of the AAMAS Workshop on Web Services and Agent-based Engineering. Baltimore:[s. n.], 2003: 267-279.
- [3] Lloyd Smith B, Kist A. A shortest paths in stochastic networks[J]. IEEE, 2004(4):492-496.
- [4] Seyyed Valhid Hasemian, Farhad Mavaddat. A Graph-based approach to web services composition[C]//Proc of the Symp on Applications and the Internet. 2005:183-189.
- [5] Liang Wen Yau, Huang ChunChe, Chuang HorngFu.

The design with object (DWO) Approach to Web services composition[J]. Computer Standards&Interfaces 2007,29(1):54-68.

- [6] 滕星全,李增智. 基于OWL-S的Web服务组合的验证研究[J]. 微电子学与计算机, 2007,24(12):62-69.
- [7] Juntao Cui, Jiamao Liu, Yujin Wu. An ontology modeling method in semantic composition of web services[C]// Proceedings of 2004 International Conference on e-Commerce Technology for Dynamic e_business. Washington, DC, USA; IEEE Computer Society, 2004: 102-109.
- [8] Liu Jiamao, fan Chenhui, Gu Ning. Web services automatic composition with minimal execution price[C]// Proceedings of the IEEE International Conference on Web Services (ICWS'05). Washington, DC, USA; IEEE Computer Society, 2005:302-309.
- [9] Cui Juntao, Zhu Yiwen, Gu Ning. An algorithm with best selection policy in semantic composition of web service[J]. Lecture Notes in Computer Science, 2005: 816-821.

作者简介:

蔡虹女,(1977-),硕士.研究方向为计算机软件及应用、数据库技术.

(上接第196页)

关系,所以接下来亟待解决的问题是如何确定识别完整度峰值时对应的各参数相应取值.

参考文献:

- [1] Weber J L, Myers E W. Human whole-genome shotgun sequencing[M]. New York: Cold spring Harbor Laboratory Press, 1997.
- [2] Beleza S, Alves C, Gonzales-Neira A. Extending STR markers in Y chromosome haplotypes[J]. International Journal of legal Medicine, 2003, 117(1):27-33.
- [3] Sonnhanner E L, Durbin R. A dot-matrix programme with dynamic threshold control suited for genomic DNA and protein sequence analysis[J]. Gene, 1995, 167(1):1-10.
- [4] 李冬冬,王正志,倪青山. 一种有效的重复序列识别算法[J]. 生物信息学, 2005, 63(4):163-166.
- [5] Gu Wanjun, Todd A Castoe, Dale J Hedges, et al. Identification of repeat structure in large genomes using repeat probability clouds[J]. Analytical Biochemistry, 2008, 380(1):77-83.

[6] Buhler J. Search algorithms for biosequences using random projection [D]. Washington, DC, USA; Ph D thesis University of Washington, 2001.

- [7] Morrison D R Patricia. Practical algorithm to retrieve information coded in alphanumeric[J]. Journal of the Association for computing Machinery, 1968.
- [8] Ukkonen E. A space-economical suffix tree construction algorithm[J]. Journal of ACM, 1976, 23(2):262-272.
- [9] Weiner P. Linear Pattern Matching Algorithms[C]// Proc 14 th IEEE Annual Symp on Switching and Automata Theory. Washington, DC, USA; IEEE Computer Society, 1973:1-11.
- [10] Han J, Wang W, Prins J. Efficient mining of frequent subgraphs in the presence of isomorphism[C]// Proceedings of the third IEEE International Conference on Data Mining. Washington, DC, USA; IEEE Society, 2003.

作者简介:

周溜溜男,(1987-),硕士.研究方向为数据挖掘、机器学习.