DnaReSM:一个基于多支持度的DNA重复序列挖掘算法*)

熊 赟 陈 越 朱扬勇
(复旦大学计算机与信息技术系 上海 200433)

摘 要 DNA序列分析研究是生物信息学的重要内容之一。基因组的基因相关区域和基因外区域中含有大量重复序列,尽管目前大多数重复序列的功能还没能肯定,但它们在遗传分析中已起重要作用。挖掘DNA重复序列成为DNA序列分析的关键。自底向上的挖掘算法中间过程产生很多短的、甚至单字符的模式,使得挖掘效率降低;另一方面,目前序列模式挖掘算法在多序列挖掘中表现出高效性,但由于单支持度定义的局限导致无法在挖掘过程中同时找到单条DNA序列中的重复序列,因此不能很好地适用于DNA重复序列挖掘。本文基于新的多支持度序列模式挖掘框架,提出了一种融合自底向上和自顶向下策略挖掘DNA重复序列的新算法DnaReSM,其结果为生物学相关实验提供基础。实验结果表明,DnaReSM探测算法能有效挖掘DNA重复序列。

关键词 数据挖掘,DNA序列,重复序列,序列模式

DnaReSM: A Multi-Supports-based DNA Repetitive Sequences Mining Algorithm

XIONG Yun CHEN Yue ZHU Yang-Yong
(Department of Computer and Information Technology, Fudan University, Shanghai 200433)

Abstract Research on DNA sequence analysis is one of important subjects in Bioinformatics. There exist a great number of repeats in DNA sequences. Presently, most of them have largely unknown functions, but they have played important roles in genetics. Mining DNA repeats is a promising task. The methods of bottom-up pattern generation in mining sequential pattern, which produce considerable short patterns, will reduce efficiency. Furthermore, present algorithms based on the definition of single support are difficult to find DNA repeats, therefore a novel method which combined bottom-up with top-down named DnaReSM based on multiple supports framework is presented to mine DNA repetitive sequences. Our experimental results demonstrate that DnaReSM is efficient.

Keywords Data mining, DNA sequence, Repetitive sequence, Sequential pattern

1 引言

生物信息学(Bioinformatics)是生命科学、计算机科学、信息科学和数学等交汇融合所形成的一门交叉学科[1],成为当今研究热点。DNA序列分析研究是生物信息学的重要内容之一。生物由于进化等目的对基因进行复制,产生大量重复序列,在遗传分析中起重要作用。过去一直认为被称为"垃圾DNA"的非基因序列没有功能,近年来的研究揭示它们具有重要的功能,并且其中很大部分是重复序列[2]。2003年5月,Makalowski在《Science》上发表文章表示重复序列在进化过程中可以用于帮助形成新基因[3],International Human Genome Sequencing Consortium在《Nature》上的文章中提到很多人类疾病是由重复序列的突变所引起的,如Williams综合症、Charcot-Marie-Tooth病(肌骨肌萎缩症)等[4]。研究和寻找DNA序列中的重复序列对于了解基因的演化、进化历史和基因变异的原因有重要意义。采用数据挖掘技术,发现并分析DNA重复序列是必要的。

本文主要贡献在于在新的多支持度框架定义[5]下,提出一种挖掘DNA重复序列的有效算法DnaReSM,其结果为生物学实验提供一定的理论基础。本文第2节介绍DNA重复序列挖掘的相关研究;第3节提出算法;第4节采用来自GenBank的真实DNA序列进行实验,证明算法挖掘结果的正确性和有效性;最后是对全文的总结和进一步研究。

2 重复序列挖掘相关研究

与一般序列数据不同,DNA序列所用的符号集合很小,长短差异大。DNA重复序列是指组成生物的DNA序列中,重复出现的序列片段。Beason的TRFinder是最有影响力的串联重复序列发现算法[6],TEIRESIAS[7]、CONSENSUS[8]、WINNOWER[9]等方法限于查找DNA序列中长度较短的重复序列;RepeatMasker[10a]调用已知重复序列的RepBase数据库,实现字符串匹配算法来检查用户提交的序列中所包含的重复序列;Dotter程序采用点阵法在同一序列中查找重复序列,具有形象化优点,但也只适用于不太长的序列[10b];Kurtz等提出的REPuter[11]基于后缀树算法克服了输入序列大小限制,但它基于子序列两两比对,难以找到DNA序列中出现次数较高的重复序列。

序列模式挖掘问题最早由Agrawal和Srikant[12]在分析交易序列数据的基础上提出,他们指出:给定序列集和一个用户指定的支持度阈值,序列模式挖掘就是找到在序列集中出现次数不低于最小支持度阈值的子序列,算法[12~15]中定义支持度为包含序列模式的交易个数(或百分比),对于一条序列,支持度仅为1,因此基于单支持度的序列模式挖掘算法不适用于分析序列数据集中各序列中的频繁模式,或当序列数据集中仅单条序列的情况。文[5]中提出了一种基于多支持度的序列模式挖掘框架,可同时挖掘多序列中的多种序列模式(存在型序列模式、局部序列模式、总体序列模式),更合适生

*)基金项目:国家自然科学基金项目(60573093)。熊 赟 博士研究生,主要研究方向为数据挖掘,数据库,生物信息学;陈 越 博士研究生,讲师,主要研究方向为数据挖掘,数据库,生物信息学;朱扬勇 博士,教授,博士生导师,主要方向为数据挖掘,数据库,生物信息学等。

物领域中 DNA 重复序列挖掘、提供个性化服务的 Web 访问模式挖掘等。另一方面，算法[12~15]采用自底向上方法，生成大量短的、甚至单字符模式，不适于长 DNA 重复序列挖掘。文[16]中提出了一种自顶向下的 ToMMS 方法，采用概念图概化短模式思想，先用简化算法从长度为 1，再成倍增长的检测方式决定频繁模式的最大长度，再用 ToMMS 算法找到所有的频繁模式。本文结合 ToMMS 算法思想，提出一种基于多支持度的探测挖掘算法 DnaReSM (Multi-Supports-Based Dna Repetitive Sequence Mining Algorithm)。

3 基于多支持度的 DNA 重复序列挖掘算法

文[5]中定义了三种有意义的序列模式：在一条指定序列中频繁出现的局部序列模式、在所有序列中具有足够出现频率的总体序列模式和在足够多序列中出现的存在型序列模式，本节介绍基于多支持度挖掘单 DNA 重复序列的算法 DnaReSM，算法可扩展到挖掘多个 DNA 序列中的其他序列模式。算法目的挖掘单 DNA 序列中的重复序列，即局部序列模式[5]。

3.1 问题定义

定义 1(单 DNA 序列) 单 DNA 序列可看作是字符集 Σ (一般地，$\Sigma=\{A,C,T,G\}$) 上的字符串，记为 $((X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n))$，其中，$(X_i,Y_i)(1\leqslant i\leqslant n)$ 是 DNA 序列中的元素，称为 DNA 序元；X_i 是 DNA 序元的序，表示每个字符在 DNA 序列中的位置信息；Y_i 是 DNA 序元的值，$Y_i\in\Sigma$。n 是 DNA 序列的长度。M 条同类单 DNA 序列组成的有限序列集称为 M-DNA 序列。

定义 2(子序列) 一条序列 $T=((x_1,t_1),(x_2,t_2),\cdots,(x_m,t_m))$ 是 M-序列中某一序列 $S=((X_{i1},Y_{i1}),(X_{i2},Y_{i2}),\cdots,(X_{mi},Y_{mi}))$ 的子序列($m\leqslant n_i$)，满足下面条件：对每一个 $j,1\leqslant j\leqslant m-1$，有 $x_j<x_{(j+1)}$ 且对于每一个 $j,1\leqslant j\leqslant m$，存在 $1\leqslant l_1<l_2<\cdots<l_n\leqslant m$，使得 $t_1\subseteq Y_{il1},t_2\subseteq Y_{il2},\cdots,t_m\subseteq Y_{ilm}$，则称序列 T 为 S 的子序列，S 为 T 的超序列。

定义 3(局部支持度) 在 M-DNA 序列 D 的某一 DNA 序列中子序列出现的次数称为局部支持度。子序列 T 的局部支持度是 S 中包含子序列 T 的数目。注，由于序列长度不同，局部支持度更合适的描述是百分比形式。

定义 4(局部序列模式) 是在 M-DNA 序列中的某一指定序列中出现次数超过一定局部支持度阈值的模式。给定正整数 $\zeta1$ 为局部支持度阈值，如果子序列 T 的局部支持度大于等于 $\zeta1$，则子序列 T 是频繁的，并称 T 是局部序列模式。

DNA 序列中重复序列挖掘问题即是在 M-DNA 序列中的某一 DNA 序列中找出所有局部序列模式。

3.2 算法 DnaReSM

从短模式开始，不断扩展的至底向上的挖掘方法生成的候选模式数量非常大，且结果频繁模式中存在许多短的甚至仅一个原始字符的模式，在使用时可能是无意义的，大大限制了其在生物领域中的应用。本文设计了一种基于多支持度的探测算法 DnaReSM，它融合了自底向上和自顶向下两种策略，基本思想是选择一定长度的模式作为探测模式，如果探测模式频繁，则自连接，向右扩展，继续探测是否还有更长的模式频繁，直到找到不再频繁的模式为止；如果探测模式非频繁，则向左缩减，检查更短的模式是否频繁，直到找到频繁模式。

根据 Aprior 性质[17]，长度为 length 的模式是频繁的，其子集模式均是频繁的，因此若探测模式选择足够适当，则可避免扫描大量更短的模式，将一定程度提高长度较长的重复序列的挖掘效率。算法关键在于探测模式和探测模式长度的选择，DNA 序列中长度为 probelen 的所有不同模式数目应是字符个数的(probelen)次方，但对于单 DNA 序列，最差情况(即每个可能的模式都不重复)下，仅有 n−probelen+1 个探测模式(n 为 DNA 序列长度)，因此探测次数最差情况是(n−probelen+1)，并非指数级。由于 DNA 序列中存在大量高频度重复序列，因此探测模式个数会随模式的重复次数大幅减少，探测模式长度范围根据生物学家的领域专家知识是可定的。

算法主要包括自底向上向右扩展探测模式(rBroaden 函数)、自顶向下向左缩减探测模式(lShorten 函数)、探测模式生成(probePatGen 函数)、模式频繁判断等几部分(图 1)。

```
输入：DNA 序列 dnaSeq；探测模式长度 probelen；
最小局部支持度阈值 min_loc_sup
输出：DNA 序列中的最大重复序列集 MaxRepeatPatternSet
1) DnaReSM (dnaSeq, probelen, min_loc_sup){
2)   probePatternSet ← probePatGen(dnaSeq,probelen);
3)   foreach 探测模式 probePatten∈probePatternSet do
4)     if (supCount(probePatten) >= min_loc_sup)
         Fre ← Fre ∪ {P|P = probePatten};
5)     else InFre ← InFre ∪ {P|P = probePatten};
6)   repeatPatternSet ← rbroaden(Fre, probelen, min_loc_sup);
7)   repeatPatternSet ← repeatPatternSet ∪
     lshorten(InFre, probelen, min_loc_sup);
8)   MaxRepeatPatternSet ← MaxRePatGen (repeatPatternSet);}
```

图 1 DnaReSM 算法

步骤 2)采用探测模式生成策略 probePatGen，从 DNA 序列中每个序元位置开始，截取长为 probelen 的片段作为探测模式；步骤 3)～5)判断每个探测模式是否频繁，若频繁存入长为 probelen 的频繁模式集 Fre 中；否则存于 InFre 中。candidator 候选模式生成函数基本思想是将长度为 k 的频繁模式集 Fre_k 进行自连接，若模式 P 的后 $k-1$ 子模式与模式 Q 的前 $k-1$ 子模式相同，则模式 P 与 Q 的第 k 个相连形成长度为 $k+1$ 的候选模式存于 $CFre_{k+1}$ 中。最大重复序列生成函数 MaxRePatGen 从重复序列集 repeatPatternSet 中找到最大超集以生成最大重复序列。

rbroaden 向右扩展函数描述如图 2。

```
参数说明：频繁模式长度 k；存放长度为 k 的频繁模式集 Frek。
返回：长度 k+1 的频繁模式集 Frek+1
1) rbroaden (Frek, k, min_loc_sup){
2)   CFrek+1← candidator ( Frek, k);
3)   foreach 候选重复序列 CP ∈ CFrek+1 do
       if (supCount(CP)>= min_sup) Frek+1← Frek+1 ∪CP;
4)   if (Frek+1 ! = Φ) rbroaden (Frek+1, k+1, min_loc_sup,)
5)  Return Frek+1;}
```

图 2 rbroaden 向右扩展函数

lshorten 向左缩减函数如图 3：

```
参数说明：InFreprobelen：长度为 probelen 非频繁模式集。
返回：向左缩减后的最大频繁模式集
1) lshorten(InFre,probelen, min_loc_sup){
   LeftMaxPattern=Φ; k=probelen−1;
2) foreach 长度为 probelen 非频繁探测模式 P ∈ InFreprobelen do
3)   while(supCount(P)< min_loc_sup && k>=lenmin) do
4)     k = k−1; P ← LEFT(P, k); // 生成长度为 k−1 的模式
5)     if( k>=lenmin && P! ∈LeftMaxLocalPattern)
         LeftMaxLocalPattern ← LeftMaxLocalPattern ∪{P};
   end for
6) return LeftMaxLocalPattern;}
```

图 3 lshorten 向左扩展函数

DNA 序列中存在大量插入、删除、突变等，因此重复序列一般是高度近似而非精确匹配的，重复序列间的相似度度量定义是一个相当大的挑战[19]。

4 实验结果

为验证 DnaReSM 算法的有效性，我们应用一个真实数

(下转封四)

(上接第 212 页)

据集进行实验，具体的实验环境是 2.6GHz 的 P4CPU，1G 内存的 PC 机，用 C++ 实现了全部算法；实验数据是来自 GenBank 序列数据库[18] 的基因序列（索取号分别为：AANG01002031、BAAF01004188、CH689725、NZAAQE01000118、BC112498）。实验任务比较探测模式长度的选择对运行时间、最小局部支持度阈值的选择对挖掘结果（最大重复序列个数）的影响等。

(1)探测模式长度与运行时间。

当探测模式长度选择适当时，所做的自连接及扩展操作相应减少，运行时间随之降低，克服了完全自底向上算法大量短候选模式的生成和判断（图 4）。

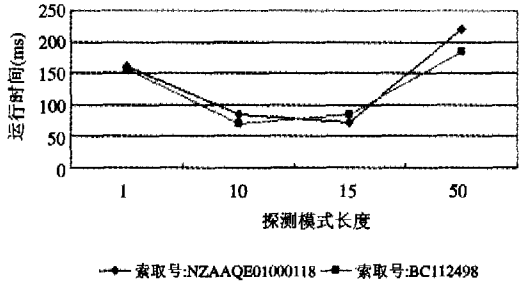

图 4　探测长度与运行时间

(2)最大重复序列个数根据最小局部支持度阈值变化。

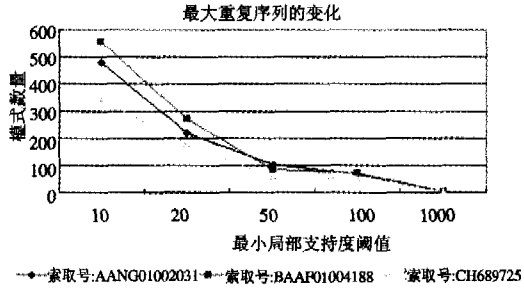

图 5　最小局部支持度阈值与最大重复序列个数的变化

结论和进一步研究　本文方法的主要思想是基于多支持度，融合自底向上和自顶向下策略挖掘 DNA 重复序列，与以往的单支持度、基于序列比对以及完全自底向上的挖掘重复序列的方法相比，DnaReSM 有更好的适用性。挖掘结果可进一步应用于 DNA 序列的分类和聚类等挖掘操作。重复序列的模糊相似度的度量[19] 是今后工作的一项挑战；另外，算法挖掘结果得到的是最大重复序列，而最大重复序列的子集的支持度可能存在较大差异，即只考虑了模式的表达，没有考虑支持度[20]，今后进一步改进以找到更合理的压缩频繁模式。最后，挖掘得到的重复序列是否具有一定的功能，有待于生物学家进一步研究鉴定。

参 考 文 献

1 Rashidi H, Buehler L. Bioinformatics Basic: Application in Biological Science and Medicine. CRC Press, 2000
2 http://wiki.cotch.net/index.php/Junk_DNA
3 Makalowski W. Not Junk After All. Science, May 2003,23
4 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, October 2004,431
5 Xiong Yun, Zhu Yang-yong. A Multi-Supports-Based Sequential Pattern Mining Algorithm. CIT, 2005. 170～174
6 Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acid Research, 1999, 27:573～580
7 Rigoutsos I, Floratos A. Combinatorial Pattern Discovery in Biological Sequences: the TEIRESIAS Algorithm. Bioinformatics, 1998
8 Hertz G Z, Stormo G D. Identifying DNA and protein patterns with statistically signicant alignments of multiple sequences. Bioinformatics, 1999
9 Pevzner P, Sze S. Combinatorial approaches to finding subtle signals in DNA sequences. the Eighth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, San Diego, USA,2000. 269～278
10 a) http://www.repeatmasker.org. b) ftp://ncbi.nlm.nih.gov(/pub/esr/dotter
11 Kurtz S, et al. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res, 2001, 29: 4633～4642
12 Agrawal R, Srikant R. Mining sequential patterns. In: Proceedings of the 11th International Conference on Data Engineering, 1995. 3～14
13 Srikant R, Agrawal R. Mining sequential patterns: generalization and perfomance improvements. In: Proceedings of the 15th International Conference on Extending Database Technology, 1996. 3～17
14 Pei J, Han J, Mortazavi-Asl B, et al. Prefixspan: Mining sequential patterns efficiently by prefix-projected growth. ICDE, 2001. 215～224
15 Han J, Pei J, Yan Xi-Feng. From sequential pattern mining to structured pattern mining: a pattern-growth approach. J. Comput. Sci. & Technol, 2004
16 Ester M, Zhang X. A Top-Down Method for Mining Most Specific Frequent Patterns in Biological Sequence Data, SDM'2004, 2004
17 Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. SIGMOD'93, 1993. 207～216
18 GenBank. http://www.ncbi.nih.gov/genbank/
19 Yang J, Wang W, Yu P S, Han J. Mining long sequential patterns in a noisy environment. 2002 ACM SIGMOD ACM Press, 2002. 406～417
20 Xin Dong, Han Jiawei, et al. Mining Compressed Frequent-Pattern Sets. VLDB,2005. 709～720

计算机科学

(1974 年 1 月创刊)
第 34 卷第 02 期（月刊）
2007 年 2 月 25 日出版

国际标准连续出版物号 ISSN 1002－137X
国内统一连续物出版号 CN50—1075/TP

定价：30.00 元　国外定价：5 美元
邮发代号：78—68
发行范围：国内外公开

主管单位：国家科学技术部
主办单位：国家科技部西南信息中心
编辑出版：《计算机科学》杂志社
重庆市渝中区胜利路 132 号　邮政编码：400013
电话：(023) 63500828　E-mail：jsjkx@swic.ac.cn
网址：www.jsjkx.com
社　长：牟炳林
总　编：彭　丹
主　编：朱宗元
主编助理：徐书令
印 刷 者：重庆科情印务有
总发行处：重庆市邮
订 购 处：全国各地邮
国外总发行：中国国际图书贸易总公司（北京
国外代号：6210—MO