ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.11, November 2007, pp.2766−2781
DOI: 10.1360/jos182766

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

DNA序列数据挖掘技术*

朱扬勇[1,2+], 熊 赟[1]

[1](复旦大学 计算机与信息技术系,上海 200433)

[2](上海生物信息技术研究中心,上海 201203)

DNA Sequence Data Mining Technique

ZHU Yang-Yong[1,2+], XIONG Yun[1]

[1](Department of Computer and Information Technology, Fudan University, Shanghai 200433, China)

[2](Shanghai Center for Bioinformation Technology, Shanghai 201203, China)

+ Corresponding author: Phn: +86-21-65642831, Fax: +86-21-65642219, E-mail: yunx@fudan.edu.cn, http://www.dmgroup.org.cn

Zhu YY, Xiong Y. DNA sequence data mining technique. *Journal of Software*, 2007,18(11):2766−2781.
http://www.jos.org.cn/1000-9825/18/2766.htm

Abstract: DNA sequence is one of the basic and important data among biological data. Researching DNA sequence data and then comprehending life essential is a necessary task in post-genomic era. At present, data mining technique is one of the most efficient data analysis means, which finds out information hidden in data. It has also become main data analysis technique adopted in Bioinformatics. It has been applied in DNA sequence analysis, which has got wide attention and rapid development. And considerable research achievements have emerged. Provides an overview of research progress in DNA sequence data mining field. In more detail, it proposes three research phases including statistics-based data mining methods application, general data mining methods application, and specialized DNA sequence-oriented data mining methods design, and then elaborates that sequence similarity is foundation of DNA sequence data mining technique. It also analyzes and comments some key techniques in this field by combining with biological background, such as DNA sequential pattern, association, clustering, classification and outlier mining. Finally, future work and open issues are given, including the research of a novel storage model and index methods, the design of data mining algorithm based on biological domain knowledge.

Key words: DNA sequence; data mining; bioinformatics; sequential pattern; sequence similarity

摘 要: DNA序列数据是一类重要的生物数据.研究DNA序列数据解读其含义是后基因组时代的主要研究任务.数据挖掘是目前最有效的数据分析手段之一,用于发现大量数据所隐含的各种规律,也是生物信息学采用的主要数据分析技术.将数据挖掘技术用于DNA序列数据分析,已得到了广泛关注和快速发展,并取得了许多研究成果.综述了DNA序列数据挖掘领域的研究状况和进展,提出了3个研究阶段:基于统计的挖掘方法应用阶段、一般化挖掘方法应用阶段和专门的DNA序列数据挖掘方法设计阶段.阐述了DNA序列数据挖掘的基础是序列相似性,评述了

* Supported by the National Natural Science Foundation of China under Grant No.60573093 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA02Z329 (国家高技术研究发展计划(863))

Received 2007-01-23; Accepted 2007-04-25

DNA 序列数据挖掘领域所采用的关键技术,包括 DNA 序列模式、关联、聚类、分类和异常挖掘等,分析讨论了其相应的生物应用背景和意义.最后给出 DNA 序列数据挖掘进一步研究的热点问题,包括 DNA 序列数据新的存储和索引机制的设计、根据生物领域知识的数据挖掘新模型和算法的设计等.

关键词: DNA 序列;数据挖掘;生物信息学;序列模式;序列相似性

中图法分类号: TP311 **文献标识码:** A

生物信息学(Bioinformatics)是生命科学、计算机科学、信息科学和数学等学科交汇融合所形成的一门交叉学科[1].DNA序列数据是生物信息学的主要研究对象之一.通过分析DNA序列,科学家不仅能够解已有的序列,而且能够更好地研究新的序列及其功能,解读序列在生物体中充当的角色,进而理解生命本质.当前,对DNA序列数据研究的方法主要是从DNA序列数据出发,分析序列中所包含的结构与功能的生物信息,所涉及的研究主题包括基因组注释、编码区和非编码区识别、基因序列功能预测等等.

随着生物信息学、分子生物学实验技术的发展,大量的各类生物数据不断产生.据统计,1982 年的第一个核酸序列数据库GenBank仅有 606 条序列,规模为 680 338bp;而 2006 年 8 月发布的数据显示,有 61 132 599 条序列,规模为65 369 091 950bp[2].如何有效地分析这些数据并发现规律以指导生物学研究和实验,是当今生物信息学研究的重要内容.

数据挖掘是目前最有效的数据分析手段,用于发现大量数据所隐含的各种规律.在 DNA 序列分析中,数据挖掘技术有着非常广阔的前景,对于提高数据处理能力、产生有价值的生物学知识起着重要作用.

自DNA序列数据库建立以来,研究者开始采用统计学方法分析DNA序列[3−5],虽然这与数据挖掘技术在实现手段和研究范围上存在差异,但当它被写成计算机程序并用于大规模DNA序列数据分析时,则成为DNA序列数据挖掘技术的雏形.但是,这类方法所需的计算量相当大.幸而,此时数据挖掘技术已有较大发展,于是人们将现有的挖掘方法直接用于DNA序列分析[6−16],这是一般化数据挖掘方法的应用阶段.但是,这些方法虽然在效率上有一定程度的提高,却并没有完全满足生物学家的需求,因为挖掘结果的可解释性和准确率可能偏离实际生物意义.因此,研究者提出应该结合DNA序列特点及实际应用背景发展专门面向DNA序列的数据挖掘方法[17−22],这标志着DNA序列数据挖掘跨入第 3 个阶段,并且这一领域的研究仍处于快速发展之中.本文对DNA序列数据挖掘技术的研究状况进行分析和综述,完整地提出了DNA序列数据挖掘的 3 个研究阶段的观点,阐述了DNA序列数据挖掘的基础是序列相似性,评述了DNA序列数据挖掘领域所采用的关键技术.最后,给出DNA序列数据挖掘进一步研究的热点问题.

本文第 1 节阐述 DNA 序列数据挖掘技术的基础是序列相似性.第 2 节给出 DNA 序列数据挖掘的 3 个研究阶段及其发展动因.第 3 节详细评述这一领域所采用的一系列关键技术,包括 DNA 序列模式挖掘、关联、聚类、分类和异常挖掘等.第 4 节提出 DNA 序列数据挖掘的未来发展趋势和研究热点.

1 序列相似性是 DNA 序列数据挖掘的基础

数据挖掘是一项通用的技术,对于 DNA 序列挖掘来说,结合 DNA 序列的特性设计挖掘算法,是 DNA 序列数据挖掘技术研究的关键和难点,也是推动 DNA 序列数据挖掘技术研究发展的主要因素之一.

1.1 DNA序列数据挖掘

生物学研究表明,DNA序列不是随机的、杂乱无章的字符串[23],它被看作是组成DNA的 4 种核苷酸A(adenine),G(guanine),C(cytosine),T(thymine)的线性排列,其中,不同排列顺序的DNA区段构成特定的功能单位——基因(gene).不同基因的功能各异,各自分布在DNA的一定区域中.与一般数据不同,DNA序列数据有其自身的特点:

① DNA 序列表示所用的符号集合很小,仅由 A,C,T,G 字符构成;

② 由非数值型的字符组成;

③ 长短差异大,有的很短,只有几十个字符,而有的却很长,甚至几百兆以上;

④ DNA 序列数据本身可能存在噪声;

⑤ 特别地,DNA 序列数据的重要特性还在于它具有特定的生物学意义.

一般地,用A,C,T,G等字符的排列顺序表示一条DNA序列,将DNA序列看作是字符集Σ(Σ={A,C,T,G})上的非空字符串,记为S,$S=\langle s_1,s_2,\ldots,s_i,\ldots,s_n\rangle$($i$=1,2,…,$n$,$s_i\in\Sigma$),其中,$n=|S|>0$,$|S|$表示DNA序列$S$的长度.多个DNA序列组成的有限序列集称为DNA序列集.

DNA 序列数据挖掘是从大量的 DNA 序列数据中寻找其规律的技术,目前,主要研究内容包括:

- DNA 序列模式挖掘:寻找某一指定 DNA 序列中(如某个指定病人)的重复模式或多条 DNA 序列组成的序列集合中(如具有同一病例特性的病人组合)的保守模式.
- DNA 序列关联挖掘:寻找两个或多个 DNA 序列或序列模式间的一种关联关系以及密切程度,通常用关联规则的形式描述,用置信度(confidence)和支持度(support)来评估.
- DNA 序列聚类挖掘:将 DNA 序列数据集划分成若干个簇,使得每个簇中的序列之间尽可能相似,而与其他簇中的序列尽可能地不相似.
- DNA 序列分类挖掘:给定一个 DNA 序列数据集和一个类标号集合,定义一个 DNA 序列数据集到类集合的映射关系,为未知类标号的 DNA 序列指定其所属的类.
- DNA 序列异常挖掘:在 DNA 序列数据集中寻找产生机制明显不同于其他序列的序列.

这些挖掘方法都离不开序列间的相似性分析,其中,DNA 序列模式挖掘目的是寻找在序列中出现次数频繁的模式,这些重复的模式之间可以是精确相等匹配或是模糊相似匹配的,序列模式之间的这种匹配程度需要借助模式间的相似程度来度量.DNA 序列聚类和分类挖掘的目的是将具有相似特点的序列划分到相同类(或簇)中,这样的类(或簇)中的 DNA 序列具有共同的特性,如相同的结构或功能等.如何判断序列之间是否具有相似性是实现聚类和分类的关键,同样也是 DNA 序列异常挖掘的关键,因为 DNA 序列异常前提是要求定义“偏差”或“异常”.

在生物学意义上,序列相似是指两个序列中存在相同和相似的位点,而序列同源性是指两个序列具有相同的祖先,是描述序列之间进化相关性时更通用的术语[24].生物学中通常认为,如果序列之间的相似性超过 30%,它们很可能是同源的[24].因而,如果两个序列有足够的相似性,则可以推测二者可能是经过序列内碱基的替换或序列片段的缺失以及序列重组等遗传变异过程演化而来的,它们可能有共同的进化祖先.所以,从其他已知功能的序列中找到与某未知序列具有高度相似性的部分,可以预测该序列的功能[25].基于序列相似、功能相似这个假设,利用DNA序列数据挖掘技术,可用于识别基因、发现功能区(如启动子、增强子等),从而确定其功能.

综上,序列相似性是 DNA 序列数据挖掘的基础,相似性研究是 DNA 数据挖掘研究的核心内容.

1.2 DNA序列相似性研究技术

DNA序列相似性研究中的一个主要应用问题是,给定一条DNA序列,在序列数据库中查询与其相似程度大于一定阈值的序列.序列比对是最常用的方法,它根据给定的相似矩阵(PAM250[26],BLOSUM62[27]等),同时考虑可能的插入、删除和突变找出序列间的最优联配,一般分为全局比对和局部比对.全局比对是对序列的全长进行比对,适用于全局水平上相似性程度较高的序列,典型的算法有Needleman-Wunsch算法[28]等.然而在实际应用中,用户提交的查询序列可能很短,需要查询的是与其相似的子序列,即整体上不具有相似性,而在一些较小的区域上存在局部相似性的序列,因此,研究局部相似性比全局相似性更有意义.于是,研究者提出了局部比对策略,寻找序列间相似性最大的子序列.典型的局部比对算法有基于动态规划思想的Smith-Waterman算法[29]以及启发式的两序列比对数据库相似性搜索算法FASTA[30]和BLAST(basic local alignment search tool)[31]等.Genbank等DNA序列数据库提供的相似性搜索服务是以序列两两比对为基础的.

多序列比对是将一组序列同时进行比对,发现序列间的相似程度.从理论上来说,两序列的动态规划比对算法可以推广到多序列比对中,但由于算法性能会随着序列的增多明显降低,因此,多序列比对大多采用启发式算法,其中具有代表性的两类主要是渐进比对和迭代比对方法.渐进比对的基本思想是,要比对的序列是进化相关

的,因此可以按照序列的进化顺序,由近及远将序列或子比对结果按两两比对算法逐步比对,重复这一过程直到所有序列都加入为止.在序列两两比对的基础上逐步优化多序列的比对结果,在序列比对后,将比对结果作进一步的处理,如构建序列谱、将序列簇构成分子进化树等.这类算法的主要优点是简单、快速,缺点在于比对初期引进的空位插入错误无法在比对后期因加入其他序列而改正,易于陷入局部最优解.CLUSTALW[32],DIALIGN (diagonal alignment)[33]等是典型的渐近比对算法.迭代比对基于一个可产生比对的算法,并通过迭代方法细化多序列比对,直到比对结果不再改进为止.这类算法不能提供获得优化比对结果的保证,但却具有鲁棒性和对比序列个数不敏感等特性.典型的算法有基于遗传算法的多序列比对SAGA(sequence alignment by genetic algorithm)算法[34]、MUSLE(multiple sequence alignment)算法[35]等.

评价生物序列比对算法的标准主要是算法的效率以及获得最佳比对结果的敏感性.两序列比对的Smith-Waterman算法敏感性强,但其复杂度高;FASTA和BLAST是以预测的敏感性的下降来换取速度上的提高;多序列比对算法中最为通用、有效的是CLUSTALW算法.序列比对中存在的主要问题在于,对于分歧较大的序列,比对的敏感性以及算法效率的提高[24].

没有建立和利用索引机制的序列比对相似性查询算法属于非索引方法,由于执行过程中需要遍历整个序列数据库,因而其性能会受到序列数量增长的影响[36].为了弥补克服非索引方法的不足,研究者提出了基于索引技术的方法,通过访问索引结构过滤不相关记录,减少数据库访问次数,主要有MRS(multi-esolution string)索引结构[37]、后缀树索引结构[38]等.针对现有索引方法存在的一些问题,Wang等人提出一种具有更高过滤能力的基于二分频率变换的 2-PFT序列相似性查询处理技术,能够处理任意长度序列的查询[36].

选择合适的序列相似性分析方法,采用相应的索引技术,并根据实际应用需求和生物背景加以改进,是DNA 序列数据挖掘的基础和关键,并有助于提高挖掘结果的准确度.

2 DNA 序列数据挖掘技术的研究阶段

DNA序列数据分析的数学理论和基本算法可追溯到 1960 年,此后,许多统计模型、算法和计算技术开始应用于DNA序列数据分析[39].实验和应用表明,早期的DNA序列分析方法在大规模数据集中效率往往不高.数据挖掘在自身发展过程中吸收了数理统计、数据库和人工智能中的大量技术,应用于DNA序列分析,经历了 3 个研究阶段(见表 1).注:这里提到的专门的DNA序列数据挖掘方法不是指所有的算法仅是针对DNA序列设计的,而是表示它是一种适合DNA序列的挖掘方法,如CLUSEQ(efficient and effective sequence clustering)等,CLUSEQ算法是针对蛋白质序列数据处理提出的,但是其算法策略可以用于DNA序列数据.

Table 1 Research phases and features of DNA sequence data mining technique

表 1 DNA 序列数据挖掘的研究阶段及其特点

Phases	Features
Statistics-Based data mining methods application	Statistics-Based data mining methods have inherent ability to capture sequential constraints present in the data, but it is not effective and efficient for a large-scale data set and the result is lack of interpretability.
General data mining methods application	To apply general data mining methods, DNA sequence data need to be mapped to a suitable form. But because dependent relationship among elements in DNA sequence is overlooked, the accuracy is reduced.
Specialized DNA sequences-oriented data mining methods design	They can capture sequential characteristics, and are scalable for large data sets. Especially, they can deal with DNA sequences containing noise, gap and fault to some extent and define some effective new similarity measure to fit DNA sequence mining.

2.1 基于统计技术的数据挖掘方法的应用阶段

统计技术和数据挖掘的共同目的是发现数据中的规律,许多数据挖掘方法运用了数理统计的算法或模型.虽然统计技术和数据挖掘技术关注的对象、实现的手段和研究的范围存在差异,但当统计算法或模型写成计算机程序并用于大规模数据分析时,它就成为数据挖掘技术.由于一些数理统计的算法或模型能够抓住数据中的序列特征,基于统计的方法开始用于分析DNA序列数据.具有代表性的方法有:Chauhuri等人采用基于统计DNA词频的方法聚类DNA序列数据[3,4];Porikli提出一种使用HMM和特征向量分解方法的可变长序列聚类算法[5].

然而,DNA 序列数据高速增长,基于统计的数据挖掘方法在分析 DNA 序列数据时,由于计算量大、结果可解释性差等因素影响了它的应用范围.于是,研究者开始将具有处理大规模数据能力的一般化数据挖掘方法引入 DNA 序列数据分析.

2.2 一般化数据挖掘方法的应用阶段

这个阶段的DNA序列数据挖掘方法是先将DNA序列数据映射到适合的形式后,再采用一般化的数据挖掘算法挖掘DNA序列数据,如类Apriori、基于前缀投影等序列模式挖掘方法[6−10]、基于划分的*k*-mediod[11]、基于层次的单连接(single-link)[12]等聚类挖掘方法,基于序列比对记分机制计算序列间相似度的KNN(*K*-nearest neighbor)分类方法[13]以及决策树[14]、神经网络[15]、支持向量机[16]等分类DNA序列数据的方法等等.

经过对 DNA 序列数据进行一定的映射变换,一般化数据挖掘方法能够实现 DNA 序列数据分析的目的.然而,一般化数据挖掘方法常常只考虑 DNA 序列元素的值,忽略 DNA 序列元素间的顺序,认为 DNA 序列中 A,C,T,G 的连续位置间是独立的、无依赖关系的,造成挖掘结果缺乏准确性,例如,DNA 序列中的元素连续位置间不是独立的,相互之间是存在依赖关系的;挖掘的结果 DNA 序列模式是具有约束的;DNA 序列数据间的相似性评估与生物序列数据分析需求相关等等.因此需要开发专门的 DNA 序列数据挖掘技术.

2.3 专门面向DNA序列的数据挖掘技术的设计阶段

由于DNA序列数据本身的特殊性,一般化数据挖掘算法难以直接应用于DNA序列数据,研究者设计了针对DNA序列数据的专门数据挖掘技术.典型的算法有:2004 年,Ester等人提出的ToMMSA(top-down method for mining most specific frequent patterns)算法[17]采用自顶向下搜索策略,能够有效挖掘长模式,在一定程度上提高了挖掘DNA长序列模式的效率;2004 年,Wang等人提出两阶段算法挖掘包含任意可变长度间隔的DNA序列模式[18];2005 年,Wexler等人提出ATRHunter能挖掘DNA序列中多种类型的近似串联重复片段ATR (approximate tandem repeats)[19];2002 年,Wang等人定义了新的序列间以及序列与簇间的相似度度量,克服了已有的相似度度量局限,并提出了CLUSEQ聚类算法[20,21];Lesh等人提出了FEATUREMINE算法[22],将序列模式挖掘技术与分类技术相结合,提高了DNA序列分类分析的准确度.国内也开展了相关研究,如复旦大学在专门面向DNA序列的数据挖掘研究方面也取得了一定的进展[40−48],在自主研制开发的数据挖掘应用平台DMiner[41]的基础上,针对生物数据特点,设计专门的生物数据挖掘算法,实现了基因数据挖掘应用平台[42];东南大学生物电子学国家重点实验室研究并发展了基于特征的基因组分析方法和技术,开发了针对基因功能、基因与疾病关系的挖掘软件[49];北京大学生物信息中心在基因预测领域的研究项目中取得了相当大的研究成果[50];中国科学院上海生命科学研究院生物信息中心完成了生物信息学数据的整合与挖掘[51],成为国内第一个生物信息学数据仓库.

专门面向 DNA 序列的数据挖掘技术较之一般化序列数据挖掘技术而言,在 DNA 序列分析上的优点在于其分析结果更具有生物学意义,因为这一阶段的方法是从 DNA 序列数据所具有的特征出发,结合了挖掘的生物学背景,在搜索策略、算法效率、准确度和有效性方面进行了改进,它们的开发极大地提高了 DNA 序列分析的效率并可满足更多的应用需求.但是,复杂的 DNA 序列数据特点、多样的生物学分析需求以及激增的生物数据规模对专门面向 DNA 序列的数据挖掘算法的研究提出了更高的要求,这一阶段的研究还需要做大量的工作.

3 DNA 序列数据挖掘的关键技术

3.1 DNA序列模式挖掘

DNA序列模式在进化过程中具有良好的保守性,对生物体的生存具有至关重要的意义.因此,识别这些模式是DNA序列数据分析的重要内容,有助于预测序列功能和解释序列间的进化关系[52].

在生物信息学领域,出现了大量的序列模式发现方法,Brazma[53]和Brona Brejova[52]分别在 1998 年和 2000 年作出了详细的归纳(见表 2),这些方法[54−62]大多在处理迅速增长的大规模DNA序列数据时面临效率难题.

Table 2 Classification, feature and representative algorithms of DNA pattern discovery methods in earlier years

表 2 早期 DNA 序列模式发现方法归类、特点及代表算法

Catagories of methods	Features	Representative algorithms
Based on exhaustive search. Enumerate all possible patterns.	Enumerate all possible patterns satisfying constraints given by user, for each pattern find its occurrences. Algorithms may run in exponential time in the worst case, therefore they may be suitable for finding short and simple patterns, especially when allowing patterns containing gaps. The advantage is that it is guaranteed to find the best pattern.	MOTIF[54], etc.
Using prune strategy in the process of search	It is difficult to find longer or more ambiguous patterns using straightforward exhaustive search. Sophisticated pruning techniques can be used to make the search feasible for typical input data. The search strategy is that starts from short patterns and then extends them until the support does not go below a certain threshold. Branches which can not yield any supported patterns are pruned. Its main advantage is that it allow to search for longer and more complicated patterns than simple exhaustive search. However the theoretical worst-case time still remains exponential.	Pratt[55], TEIRESIAS[56], etc.
Iterative heuristic method	It is not necessary to find the best pattern, but may converge to a local maximum, the strategy which limit length of pattern or gap of pattern is used, it may be lost some useful patterns, but it is guaranteed to find a pattern almost as good as possible.	Gibbs[57], COPIA[59], etc.
Methods constructing stochastic model	Some patterns cannot be well described by a simple deterministic pattern and can be expressed in form of stochastic model, such as Hidden Markov model, or position weight matrix. This kind of models does not necessarily converge to global maximum.	EM[60], MEME[61], etc.
Methods using additional information	Using information from sequence alignment, or global properties of a sequence. The results are more suitable for biological meaning and application because using more information available about sequences.	EMOTIF[62], etc.

在数据挖掘领域,研究者提出了适用于大规模数据的更高效的序列模式挖掘算法,并受到广泛关注.序列模式挖掘问题最早由Agrawal和Srikant于1995年在分析交易序列数据的基础上定义[6],他们指出:给定一个序列集和用户指定的支持度阈值,序列模式挖掘就是找到所有的频繁子序列,即在序列集中出现次数不低于最小支持度阈值的子序列[6].1996年,Srikant等人提出基于Apriori思想的GSP(generalized sequential patterns mining)算法[7].算法引入时间和概念层次约束,采用自底向上宽度优先策略挖掘所有频繁模式,但它的主要问题在于,当序列数据库规模很大时产生大量的候选模式,需要频繁扫描序列数据库,尤其是在序列模式长度较长的情况下,其对应的短模式规模太大,导致算法难以处理,整体效率不高;为了解决这个问题,2000年,Pei等人提出了基于模式增长的PrefixSpan算法[8].算法采用分治思想,不断产生序列数据库的多个更小的投影数据库,然后在各个投影数据库上进行序列模式挖掘,算法无须产生候选模式,大幅度缩减了搜索空间,其主要开销在于投影数据库的构造,性能优于类Apriori算法.

然而,DNA序列不同于交易序列等序列数据,直接将GSP[7],PrefixSpan[8],FreeSpan[9],SPADE(sequential pattern discovery using equivalence classes)[10]等序列模式挖掘算法应用到DNA序列数据时,实验证实在可伸缩性等方面存在问题,如某些数据结构或剪枝策略(Hash树、idlist/bitmap表示等)不能有效适用于DNA序列等.因此,研究者设计了专门的DNA序列模式挖掘算法.

本文从以下几个方面综述专门面向DNA序列的模式挖掘技术的研究状况:

① 搜索策略.自底向上的模式搜索策略是从短的频繁模式开始,不断扩展直到产生不频繁模式.这是序列模式挖掘中的一种常用方法.但是,由于频繁模式的子模式也是频繁的,该方法结果集中会包含一定数量长度较短、在使用时可能无意义的模式,在一些情况下不适合长的DNA序列模式的挖掘.在处理DNA序列数据时,可以采取一些额扩展或剪枝方法.针对DNA序列中长模式的挖掘,另一种策略是自顶向下的搜索方法,这种方法在一定程度上能够减少大量短模式的分析,如,2004年,Ester等人在文献[17]中提出一种采用自顶向下的模式搜索策略的算法ToMMSA.首先,为了不丢失某些可能的特定模式,算法设计一个自底向上的搜索函数,用于确定频繁模式的最大长度值;然后,从所有具有指定最大长度的子序列开始对每个模式进行缩减,直到其频繁为止.缩减的方法是减少子序列中的一个模式,或是利用概念图对其进行概化.ToMMSA方法生成的候选模式最大数为$n\times l\times(l-1)/2$,即$O(nl^2)$(其中,l是序列数据库中序列的最大长度,n是序列数目),远远少于自底向上方法.而自底向上方法最坏情况下产生的候选模式数目为$O(|\Sigma|^m)$(其中,Σ表示DNA序列字符个数,m是频繁模式最大长度).

② 具有约束限制的模糊匹配的生物序列模式.由于生物序列可能包含任意长度间隔的模式,模式间的匹配往往是包含间隔的模糊匹配.现有的一些带有约束的序列模式挖掘算法没有考虑生物序列模式的各种约束特征.由于DNA序列可能包含任意长度间隔的模式,模式间的匹配往往是包含间隔的模糊匹配.2004年,Wang等

人提出了一种有效挖掘包含任意长度间隔DNA序列模式的两阶段算法[18],其基本思想是,在第 1 阶段(segment phase),搜索所有称为segment的不包含间隔的短频繁模式X_i;第 2 阶段(pattern phase),使用第 1 阶段得到的频繁模式X_i生成包含被可变长度间隔分割的多个segment连成的长模式$X_1\times\ldots\times X_k$,虽然这个阶段是耗时的,但不同于以往的方法,它每次扩展一个segment而非一项.两阶段算法的优点在于,能够利用从第 1 阶段获得的局部模式信息X_i降低第 2 阶段全局模式$X_1\times\ldots\times X_k$搜索的时间.这种策略适合挖掘包含任意长度间隔的模式.然而,在生物序列模式中的约束条件是十分复杂的,此具有约束的生物序列模式挖掘方法的研究仍是当前生物序列模式挖掘研究中的一个重点和难点.

③ 挖掘结果模式的压缩和优化.规模序列集中包含的序列模式数量非常大,并存在许多冗余模式,需要耗费大量时间和空间,并且使用者难以理解.为了减少结果模式的数量,去除冗余,常常取包含大量一般模式的特定频繁子序列,如输出最大(maximal)序列模式或闭合(closed)序列模式.闭合序列模式挖掘能够大量减少生成模式的数量,并且这种压缩是无损的.Yan等人提出一种基于PrefixSpan算法的闭合序列模式挖掘算法:CloSpan算法[63],定义闭合序列模式为:一个序列模式s是闭合的,如果在数据库中不存在任何与s具有相同支持度的s的超序列. CloSpan算法采用的基本思想是,在已挖掘的候选闭合序列模式集上进行候选-测试(candidate maintenance-and- test).使用这个集合剪枝搜索空间,并判断是否新的序列模式可能是闭合模式.由于大量闭合模式(或候选模式)占据内存并需要大量空间检测新的模式,因此使用CloSpan算法处理长序列数据时,所需的代价相当大.之后,Wang等人提出BIDE(bi-directional extension based frequent closed sequence mining)闭合序列模式算法[64],采用一种称为BI-Directional Extension的闭合检测策略,算法不需要维护候选模式.文献[64]中实验证实,BIDE比CloSpan更为有效.相关的研究还有Tzvetkov等人提出Top-K闭合序列模式挖掘算法TSP(top-k closed sequential patterns mining)[65]以及Cong等人提出并行闭合序列模式挖掘算法Par-CSP(parallel closed sequential pattern mining)[66]等.这些方法的基本思想和算法策略可以根据生物序列数据的特点进行改进并应用于生物序列模式挖掘研究中.对结果模式的压缩可在一定程度上有效去除冗余,提高模式集的可用性.

④ 特殊类型的序列模式——串联重复序列(tandem repeats).生物由于进化等目的对基因进行复制产生大量重复序列,研究表明,重复序列并不是没有任何功能的,只是更多的功能目前尚未发现,并且是非常重要的[67].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[68],International Human Genome Sequencing Consortium在《Nature》上的文章中提到,很多人类疾病由重复序列的突变所引起,如Williams综合症、Charcot-Marie-Tooth病(肌骨肌萎缩症)等[69].因此,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等都有重要意义,为今后研究基因或非基因序列的未知功能奠定了基础.采用数据挖掘技术,发现并分析这些高频度、长长度的DNA重复序列是非常必要的.串联重复序列挖掘是其中的一个主要内容,已用于DNA指纹印记、基因定位、比较基因组学以及进化研究等,具有重要的生物意义.串联重复序列是在某一DNA序列中重复出现次数超过一定数量的、重复单位呈串状、首尾相连排列的一段序列[70].早期的研究主要是针对完全串联重复序列PTR(perfect tandem repeat)[71,72],国内也有相当多的研究成果.例如 2005 年,Wang等人提出了针对精确查找的新的重复片段的定义——最大模式重复LPR(largest pattern repetition),并设计了索引结构后继数组(succeeding unit array,简称SUA)完成LPR的查找[73].然而,由于DNA序列存在突变等情况,导致这些重复成为非精确相等的.已有的PTR发现,算法无法发现这样的近似串联重复序列ATR.于是,出现了大量ATR发现算法[19,70,74−81],如Beason设计的TRFinder算法是最有影响力的串联重复序列发现算法[70];Kurtz等人提出的REPuter[78]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;2005 年,Li等人提出了一种新的投影拼接算法有效识别重复序列[82];2005 年,Ydo Wexel等人给出了关于评分函数的 3 种类型ATRs的定义,给定一个评分函数Φ,序列片段$T=T_1T_2\ldots T_r$,当存在一个子序列T_*,满足$\Phi(T_i,T_*)\geq\eta$(其中,i=1,…,r),则T为简单ATR(simple ATR);当满足$\Phi(T_i,T_{i+1})\geq\eta$(其中,$i$=1,…,$r$−1),$T$为邻近ATR(neighboring ATR);当满足$\Phi(T_i,T_j)\geq\eta_{ij}$时,$T$为成对ATR(pairwise ATR)[19].文中提出一种由扫描和候选验证两阶段组成的ATR发现算法,扫描阶段采用可变大小的滑动窗口检测相邻子串是否相似,生成可能包含ATRs的候选域列表,再由第 2 阶段验证这些候选ATRs,验证过程是首先将含

有长度为t模式的候选ATR与下一个长度为t子串比对,计算评分是否大于给定阈值;然后扩展ATR.算法还设计了一个统计框架以调整阈值、评分函数参数以及模式间间隔等;2007 年,Wang等人研究了相似性重复片段的查找问题[83,84],针对重复片段查询的索引结构所需空间过大问题,引入了新的相似性标准以及SATR(segment-similarity based approximate tandem repeats)的概念,并在后继数组(SUA)的基础上设计了SATR的查找算法.算法在查找过程中不需要限制模式长度,在同样相似度要求下,对于相同的待查序列,算法查找时间低于其他同类方法.然而,目前的序列模式挖掘算法将支持度定义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,关于串联重复序列模式的挖掘不同于KDD领域中以往的序列模式挖掘算法,是一项值得研究的新内容.

以上从搜索策略、模式的约束、结果模式选择、模式类型等方面析了现有的生物序列模式挖掘方法,们各自有相应的生物学背景.处理实际生物序列模式挖掘问题时,对不同的生物学需求,要根据相应的领域知识和所处理的序列的特征,顾多方面因素设计合适的生物序列模式挖掘算法(见表 3).

Table 3 The criterions of DNA sequential pattern mining algorithm

表 3 DNA 序列模式挖掘算法的考量指标

The criterions of DNA sequential pattern mining algorithm	Features	Scope of application	Representative algorithms
Search strategy	Traditional sequential pattern mining methods following the paradigm of bottom-up pattern generation, which lead to very large numbers of patterns. And they are typically too short to be meaningful. The method which performs top-down pattern enumeration can be used to mine long pattern efficiently, because it does not generate enormous meaningless short patterns. The results may be more suitable for biological meaning.	To find long and meaningful DNA sequential pattern.	ToMMSA[17], etc.
Scale of result pattern set	Since a long sequence always contains considerable subsequences, an explosive number of candidate sequential patterns will be generated, which is prohibitively expensive in both time and space. Mining closed sequential patterns can lead to not only more compact yet complete result set but also better efficiency.	Eliminating redundant patterns in order to get suitable result pattern set. This compressed set possesses improved explanation.	CloSpan[63], BIDE[64], etc.
Special pattern type	In earlier years, the aim of tandem repeats finding methods is to deal with exactly equal patterns. However, events such as mutations, translocations and reversal events will often make copies imperfect. These patterns consist in approximate and adjacent repetitions of a DNA sequence.	Finding approximate tandem repeats is an important and meaningful tasks in bioinformatics research. Tandem repeats detection is useful for genetic markers, such as for DNA fingerprinting, mapping genes, comparative genomics and for evolution studies.	ATRHunter[19], SATR finding algorithm[84], etc.

3.2 DNA序列关联分析

关联规则挖掘问题自 1993 年由Agrawal等人在文献[85]中引入后,开始受到广泛关注.事实和实验表明,大部分疾病并非仅由一种基因变化而激发,而是一组基因共同作用的结果[39].DNA序列关联分析能够发现两个或多个DNA序列之间存在的规律、描述序列间的密切度,有助于研究DNA序列间的相互关系和相互作用,如医学中并发症的研究等.

简单地说,DNA序列关联规则是形如S==>T的表达式,其中,S和T是DNA序列模式[86].DNA序列关联分析算法[87−89]包括两个步骤:第 1 步生成序列模式,这是关联分析中的关键步骤(方法采用第 3.1 节中提到的序列模式挖掘算法);第 2 步由得到的序列模式生成强关联规则S==>T,规则的有效性用支持度*support*(包含S和T的DNA序列数量的百分比)和置信度*confidence*(包含S和T的DNA序列数与只包含S的DNA序列数的比值)来评估.

DNA序列关联分析的另一个应用是检测在DNA序列数据集中冗余出现的序列[90],作为DNA序列其他挖掘操作的预处理过程.这是因为,相同的DNA序列常被提交到一个以上的数据库或被多次提交到同一数据库等因

素,从而造成DNA序列数据集出现冗余.

此外,与关联分析有相关性的一个DNA序列数据挖掘操作是演变分析(evolution analysis),其任务是找到序列在时间或阶段上的关联关系.大量的医学实验证实,引起疾病的基因可能在疾病的不同阶段起作用,因此,如果能够找到控制疾病发展的不同阶段的遗传因子,则有可能开发针对疾病不同阶段的治疗药物,从而取得更为有效的治疗效果.这是生物制药研究非常关心的问题,具有很好的实际价值,DNA 序列演变分析技术正是一种有效的手段.

3.3 DNA序列聚类分析

聚类分析是识别未知类别的 DNA 序列所属的类别、揭示序列之间相互关系,进而分析 DNA 序列功能的一种有效方法.

早期的聚类算法,如基于划分的*K*-medoid算法[11]、基于层次的全连接(complete-link)算法[12]等,在计算相似度时,由于需要进行成对比较,以致计算复杂度很大,因此,它们难以直接应用于大规模的DNA序列数据集中.

一种解决方法是使用"编辑距离(edit distance)"方法计算每对序列之间的距离[91].两个序列s_1和s_2之间的编辑距离被定义为将s_1映射到s_2所需操作({insert,delete,replace})的个数,然而"编辑距离"的方法只抓住序列对间的全局特征,却忽视序列的局部特征[20],如两个(或多个)序列所共享的序列模式可作为生成簇的重要特征.之后,文献[92,93]提出的块操作(block operation)在一定程度上弥补了这一不足,但这种方法仍不能完全解决这个问题,且计算是NP-Hard的.

Wang等人提出CLUSEQ算法[20]构建了基于序列统计属性的相似度度量方法,认为紧接着一个段(segment)的符号的条件概率分布能够更准确地描述这个序列的结构特性,有效克服了编辑距离方法的不足,并提出用概率后缀树PST(Probabilistic Suffix Tree)来组织获取一个簇中序列的条件概率分布,使得相似度的计算更加准确、有效.CLUSEQ算法能够处理不同长度的序列,并将序列聚类为可能重叠的簇的集合,而且簇的数量和异常点的数量可以在聚类过程中自动调节.

另一种解决方法是基于特征抽取进行聚类的方法.这类方法的基本思想是,DNA 序列数据中的某些序列模式常常能够决定序列的功能特点,对序列的聚类起决定性的作用,因此可以抽取能够表达序列特征的序列模式作为聚类分析的第 1 步,然后基于抽取的特征采用传统的聚类算法对 DNA 序列数据进行聚类.其中具有代表性的工作有:

Chaudhuri等人提出基于DNA序列中DNA词(words)分布的统计,使用DNA序列中词频的标准欧式距离作为相似度度量,再使用平均连接(average linkage)聚类方法进行聚类[3,4].该方法认为,DNA词频可以作为给定DNA序列的统计概貌(statistical summaries),将比较DNA序列间的相似性转化为比较DNA序列相关的给定大小的DNA词的频率分布,整个DNA序列被认为是完整分子结构的描述,而指定大小的词被认为是分子的特征,并设计了相应工具SWORDS.

更进一步地,Guralnik等人提出一种可伸缩的聚类序列数据算法[94].方法是,首先抽取能够表示序列特性的特征集,将序列投影到这些特征的新空间;然后运用传统的*k*-means聚类方法[11]找到转化空间的序列簇,解决了由于*k*-means算法难以计算序列数据质心而无法直接应用到DNA序列数据聚类的问题,并且该方法不使用动态规划计算相似度.实验证明了算法不仅具有好的可伸缩性,而且也得到了良好的聚类结果.

类似的方法还有一些,比如 Morzy 等人在文献[95]中也提出,为了提高大序列数据集中的层次聚类效率,他们使用频繁子序列作为描述序列特征,对共享共同子序列的序列组进行操作,用凝聚的层次聚类方法实现序列数据的聚类.该算法不仅能发现高质量的簇,而且能以组成簇中序列共享的频繁的形式提供簇的描述.

对DNA序列聚类还需要根据生物学领域知识,从生物意义角度出发,得出更具有实际应用价值的聚类结果.如目前基因芯片技术已为生物学研究提供了大量的基因表达数据,通过对基因表达数据进行聚类分析,可以得到在一定条件或环境影响下具有共调控功能的基因簇.进一步地,基于基因表达数据得到的聚类结果还可以对这些共表达基因的上游序列挖掘保守模式,以提高转录调控元件预测的准确程度,并能得到关于调控元件与相应的生理条件或环境信息间的关联.典型的基因表达数据聚类方法是双向聚类,如BiCluster[96]及其的改进算法

Op-Cluster[97],pCluster[98]等.

如何抽取不同长度 DNA 序列的序列特性,以及如何结合生物学应用问题设计有效的相似度度量,仍是目前 DNA 序列聚类分析方法面临的两个关键问题,并将直接影响到聚类 DNA 序列结果的准确性和算法的高效性,同时也是改进现有 DNA 序列聚类算法、发展新算法的前提和基础(见表 4).

Table 4 The criterions of DNA sequence clustering algorithm

表 4 DNA 序列聚类算法的考量指标

The criterions of DNA clustering algorithm	Scope of application	Representative algorithms
Features of sequence	Aimed at clustering sequences which share local common characteristic instead of global similar.	CLUSEQ[20], etc.
Sequence similarity measure	In DNA sequence set, some features which represent sequential nature of sequences can be exacted, for example, protein domain when DNA sequences be translated into protein sequences, because domain often refers to common function possessed by the protein family. Designing similarity measure based on such information may enhance accuracy.	SWORDS[3,4], Guralnik approach[94], etc.
Combining biological knowledge and background	Biological experimental data which related to DNA sequences can be provided, such as gene expression data, etc.	BiCluster[96], etc.

3.4 DNA序列分类分析

DNA序列分类分析目的是能够为未知类标号的DNA序列S指定其所属的类C,进而预测它的功能以及它与其他DNA序列间的相互关系,辅助DNA分子中的基因识别等[99,100].

1999 年,Wang等人将DNA序列分类技术划分为 3 类[101]:(i) 一致搜索(consensus search),找到类C的序列集并生成一个一致序列(consensus sequence),即在类C中经过多序列比对后生成的子序列,用这些子序列识别未标号DNA序列;(ii) 归纳学习/神经网络方法,将类C的序列集和非C的序列集,使用机器学习方法生成规则决定未标号序列S是否属于C;(iii) 序列比对方法,使用比对工具(如FASTA[30],BLAST[31])等将未标号的序列与C的成员进行比对,将具有最高比对得分的类标号C赋予序列S.

随着 DNA 序列数据规模的不断增大,关于 DNA 序列分类技术的研究也不断发展,根据目前的研究状况,我们认为 DNA 序列数据分类技术可分为以下几类:

① 基于序列比对的分类方法

使用序列比对工具分类DNA序列数据是较早使用的一种方法.其基本思想是,将未标号的序列S与类C_i(i=1,2,…,m,m是类别总数)中的序列进行比对,如果S具有最高比对得分,则将S归为类C_i.然而,当数据规模较大时,序列比对策略的计算量将是非常大的.

② 基于统计的分类方法

基于马尔可夫链(Markov chain)的分类分析方法是,根据每个序列的类标号划分训练序列,并对每个更小的数据集建立一个马尔可夫链.对每个测试序列S_r,通过计算由每个马尔可夫链生成的序列的概率,即条件概率$P(S_r|M)$,然后将具有最高概率的马尔可夫链相关的类标号赋予测试序列S_r[102].隐马尔可夫模型HMMs(hidden Markov model)是由马尔可夫链发展扩充而来的一种随机模型,HMM能够抓住序列数据的序列特性,有效地对序列数据进行建模,成为DNA序列分类分析中应用广泛的模型之一.

③ 传统机器学习分类方法

基于统计的方法对序列分类后的结果缺乏可解释性,而传统的机器学习分类方法对分类结果具有可解释性,因此易于提高生物学家对知识的理解[103].

在机器学习方法中,由于KNN具有简单且能捕获序列特性的特点[102],它不需要在训练集上建立分类器,因此是一种DNA序列分类分析的常用算法.KNN分类DNA序列数据的方法是,为分类某一测试序列S,首先找到k个与它最相似的训练序列,然后将这k个序列中最具代表性的类标号赋予测试序列.虽然KNN不需要任何训练时间,但在寻找k个最相似序列时,需要计算两个序列间相似性度量,通常采用的是序列比对记分,而这种方法的计算量是相当巨大的[102].

传统的机器学习分类方法(如决策树[14]、神经网络[15]、支持向量机SVM(support vector machine)[16]、贝叶斯分类[104]等)难以对具有序列特性的数据集建立分类模型.为了更加有效和准确地获得DNA序列数据的分类结果,DNA序列数据需要被映射到适合这些算法的形式(如建立Markov model).2002 年,Deshpande等人比较了基于KNN、基于SVM、基于Markov model等几种分类器[102],实验表明,SVM方法与其他两种方法相比,在分类生物数据时具有较高的准确率.

Maddouri等人提出通过生成知识基并使用知识基中的规则分类DNA序列的机器学习分类方法[103].方法首先使用KMR(Karp, Miller and Rosenberg algorithm)算法构造与每个类相关的特征描述(discriminant descriptor),然后用样本属性表编码这些类,再从表中抽取知识生成规则,最后使用获得的规则分类序列.算法中还使用punishment/award算法为每个规则设置权重,正确的规则增加权重,而错误的规则减少权重,并生成一个排序关系为分类提供规则选择.

基于机器学习的分类方法对分类结果具有很好的可解释性,然而,这类方法仅将单个核酸位置作为属性,在构造分类器时,单独考虑每个属性,认为核酸的连续位置间是无依赖关系的、独立的.但事实上,只认为特征值是重要的、而不考虑特征的顺序的方法,将造成这些分类方法在 DNA 序列分类上缺乏准确性.因此,一个有效分类 DNA 序列数据的分类器应该是考虑各种共享子序列或特征的相对位置信息.

④ 基于模式挖掘的特征抽取分类方法

传统的分类算法难以直接应用于DNA序列分类,但由于每个序列样本具有大量潜在的有用特征来描述它,研究者设计了一种基于模式挖掘的特征抽取的分类方法,即先抽取序列模式集作为DNA序列分类的预处理步骤,然后用贝叶斯、决策树等标准分类方法分类DNA序列.为了减少特征数量,通常采用 3 个剪枝标准:特征是频繁的;每个选择的特征要与至少 1 个类相关;特征集中不包含冗余特征.序列模式挖掘和分类挖掘两种技术相结合能够提高标准分类算法在分类序列数据时的分类精度.实现这一思想的一种方法是 1999 年Lesh等人提出的FEATUREMINE算法[22].该算法首先用序列模式挖掘方法(如SPADE[10])挖掘得到与目标类相关的序列模式,然后将其转换为适合标准分类算法的布尔特征集,再用贝叶斯等标准分类算法从不同特征学习权重对新样本进行分类.

在以上 4 类方法中,基于序列比对的分类方法的优点是简单,其不足在于当数据规模增大时,需要的计算量过大;基于统计的分类方法能够抓住序列数据的序列特性,但分类后,结果的可解释性弱于基于传统的机器学习的分类方法;而结合序列模式挖掘方法的基于特征抽取的分类方法在保持序列特性和分类精度目标方面都有理想的效果.

3.5 DNA序列异常分析

Hawkins于 1980 年给出了异常的本质性定义:异常是数据集中偏差较大的数据,它们的产生机制可能不同于其他数据[105].大多数聚类算法在一定程度上可以检测异常.然而,它们的主要目的不是研究异常本身价值,而是保证异常不干扰聚类过程从而优化聚类结果.

DNA序列异常分析可用于在DNA序列数据集中检测异常DNA序列,也常被作为DNA序列数据挖掘的数据预处理任务之一.目前,关于DNA序列异常挖掘技术的研究相对前面几种技术而言比较少,但已开始逐渐受到研究者的关注.2006 年,Sun等人提出了一种使用概率后缀树挖掘序列数据集中异常的方法[106],算法的关键点是,当序列集被组织在概率后缀树PST(probabilistic suffix trees)中时,仅通过检查与PST中根节点相近的点就能区分异常和非异常序列,因此,算法的优点在于仅需构造PST的一部分而非全部,由此降低运行时间和存储代价.文献还对比了两个相似度度量:SIM_N(normalized probability)和SIM_O(odds)度量,实验证明是更适合序列异常检测的相似度度量,并且用基于熵的信息论参数解释了normalized probability度量的可靠性.

当前,DNA 序列异常挖掘的研究还将面临很多挑战,这主要是因为,高维情况下的异常挖掘算法仍不够成熟和完善;缺乏对生物意义上的“异常 DNA 序列”的合理定义,目前仍处于研究的探索阶段.

4 总结和进一步工作

DNA 序列分析为探索生物间和生物体内的遗传变异提供了机会,DNA 序列数据的高速增长、人们对 DNA 序列分析需求的不断扩大,使得 DNA 序列数据挖掘技术面临新的挑战,同时又为其未来的发展创造了新的机遇.本文回顾了 DNA 序列数据挖掘领域的主要研究及最新成果,提出了 3 个研究阶段,重点讨论了这一领域的各种关键技术及其存在的问题,并给出了相应的生物应用背景.基于此研究,我们认为未来若要在该领域取得突破性进展,以下两方面的一些关键问题值得特别关注:

一方面是在 DNA 序列分析需求清楚的情况下,如何设计更高效的 DNA 序列数据挖掘算法.包括:

① DNA 序列数据新的存储和索引机制的研究.目前已出现大量 DNA 序列数据挖掘算法,但它们在分析 DNA 序列数据时仍表现出低效性或结果准确度低的一个原因是大多数 DNA 序列数据仍以文本文件形式存放,这种存储机制造成了 DNA 序列数据存储和访问效率低下,导致算法分析效率降低.因此,研究新的 DNA 序列数据模型、设计合理的存储和索引机制、加快对序列数据的访问,是提高 DNA 序列数据挖掘算法效率的必要手段.它将成为一个新的研究热点.

② 融入生物学背景知识作为先验知识设计新的数据挖掘模型和算法.现有的 DNA 序列数据挖掘算法在处理实际应用中的问题时可能具有高效的性能,但是挖掘结果的敏感性和特异性难以达到良好的生物学指标要求,甚至比生物学中一些传统工具准确率还要底.这是因为在设计模型和算法时没有结合相应的生物学知识,如基因表达数据的聚类可以得到同类基因序列.但是,仅仅考虑数据值之间的关系,而不考虑表达水平具有一致波动趋势的基因也可能是相似的,那么在设计算法时将会与实际生物意义相背离.目前,已有的一些成功的研究表明结合生物学背景知识为设计更高效的 DNA 序列数据挖掘算法提供良好的指导,例如,基因本体 Go(gene ontology)语义模型的引入;在分析基因表达谱数据的基础上进一步利用序列模式挖掘技术研究基因上游序列的转录调控元件等.因此,如何结合生物意义挖掘 DNA 序列数据是值得继续研究探索的难题.

另一方面的关键问题是,在 DNA 序列分析需求不明确,或者是没有生物学家的先验知识的情况下,如何发展新的挖掘算法,为生物学研究提供指导.包括:

① 近年来的研究表明,来自不同物种的控制区域并不是序列看起来相似就发挥相似功能[107].因此,传统的基于“序列相似、功能相似”假设的DNA序列分析方法在一定程度上限制了从DNA序列中寻找出更多的规律和知识.在突破这一假设的基础上分析DNA序列数据、开展生物信息学研究,数据挖掘技术将是最具潜力的.

② 研究人员已成功分析了很多被人们称为垃圾DNA序列(非编码DNA序列)的序列,鉴定出一些在控制基因功能方面起关键性作用的DNA序列区域,证实了“垃圾DNA序列可能并不是这么没有价值”[68,108,109].这表明,非编码区DNA序列同样隐藏着能为生物实验提供指导的信息,数据挖掘技术无疑是辅助发现这些隐藏信息的有效手段.

综上,本文详细综述了生物信息学领域的研究热点——DNA 序列数据挖掘技术,提出了未来 DNA 序列数据挖掘研究领域取得突破性进展所需关注的几个关键问题以及今后的一些研究方向和趋势.我们相信,数据挖掘技术和生物领域更加紧密结合,将得到更多有意义的挖掘结果,为未来的人类生命研究提供良好支持.

References:

[1] Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods Information in Medicine, 2001,40(4):346−58.

[2] GenBank. National Center for Biotechnology Information, 2007. http://www.ncbi.nih.gov/genbank/

[3] Chaudhuri P, Das S. SWORDS: A statistical tool for analyzing large DNA sequences. Journal of Biosciences, 2002,27(1):1−6.

[4] Chaudhuri P, Das S. Statistical analysis of large DNA sequences using distribution of DNA words. Current Science, 2001,80(9): 1161−1166.

[5] Porikli FM. Clustering variable length sequences by eigenvector decomposition using HMM. In: Ana LNF, Terry C, Robert PWD, Aurelio CC, Dick DR, eds. Proc. of the Int’l Workshop on Structural and Syntactic Pattern Recognition. LNCS 3138, London: Springer-Verlag, 2004. 352−360.

[6] Agrawal R, Srikant R. Mining sequential patterns. In: Yu PS, Chen ALP, eds. Proc. of the 11th Int’l Conf. on Data Engineering.

Taipei: IEEE Computer Society, 1995. 3−14.
[7] Srikant R, Agrawal R. Mining sequential patterns: Generalization and performance improvements. In: Apers PMG, Bouzeghoub M, Gardarin G, eds. Advances in Database Technology, Proc. of the 15th Int'l Conf. on Extending Database Technology. London: Springer-Verlag, 1996. 3−17.
[8] Pei J, Han JW, Mortazavi-Asl B, Pinto H. Prefixspan: Mining sequential patterns efficiently by prefix-projected growth. In: Proc. of the 17th Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2001. 215−224.
[9] Han J, Pei J, Mortazavi-Asl B, Chen QM, Dayal U, Hsu MC. Freespan: Frequent pattern-projected sequential pattern mining. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery in Databases. New York: ACM, 2000. 355−359.
[10] Zaki MJ. SPADE: An efficient algorithm for mining frequent sequences. Machine Learning, 2001,42(1-2):31−60.
[11] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Johh Wiley & Sons, 1990.
[12] Day WHE, Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. Journal of Classification, 1984, 1(1):7−24.
[13] Fix E, Hodges JL. Discriminatory analysis—Nonparametric discrimination: Consistency properties. Int'l Statistical Review, 1989, 57(3):238−247.
[14] Quinlan JR. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1993.
[15] Ripley BD. Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press, 1996.
[16] Vapnik VN. The Nature of Statistical Learning Theory. 2nd ed., New York: Springer-Verlag, 2000. 138−167.
[17] Ester M, Zhang X. A top-down method for mining most specific frequent patterns in biological sequence data. In: Berry MW, Dayal U, Kamath C, Skillicorn DB, eds. Proc. of the 4th SIAM Int'l Conf. on Data Mining. 2004. 90−101.
[18] Wang K, Xu Y, Jeffrey XY. Scalable sequential pattern mining for Biological sequences. In: Grossman D, Gravano L, Zhai CX, Herzog O, Evans DA, eds. Proc. of the 13th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2004. 178−187.
[19] Wexler Y, Yakhini Z, Kashi Y, Geiger D. Finding approximate tandem repeats in genomic sequences. Journal of Computational Biology, 2005,12(7):928−942.
[20] Yang J, Wang W. CLUSEQ: Efficient and effective sequence clustering. In: Dayal U, Ramamritham K, Vijayaraman TM, eds. Proc. of the 19th Int'l Conf. on Data Engineering. Bangalore: IEEE Computer Society, 2003. 101−112.
[21] Yang J, Wang W. Towards automatic clustering of protein sequences. In: Proc. of the 1st IEEE Computer Society Bioinformatics Conf. (CSB2002). Washington: IEEE Computer Society, 2002. 175−186.
[22] Lesh NB, Zaki MJ, Ogihara M. Mining features for sequence classification. In: Conf. on Knowledge Discovery in Data, Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 1999. 342−346.
[23] Chen X, Kwong S, Li M. A compression algorithm for DNA sequences and its applications in genome comparison. Genome Inform Ser Workshop Genome Information, 1999,10:51−61.
[24] MountDW. Bioinformatics Sequence and Genome Analysis. New York: Cold Spring Harbor Laboratory Press, 2001.
[25] Rogozin IB, Milanesi L, Kolchanov NA. Gene structure prediction using information on homologous protein sequence. Computer Applications in Biosciences, 1996,12(3):161−170.
[26] Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. National Biomedical Research Foundation, 1978,5(3):345−352.
[27] Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. Proc. of the National Academy Sciences of the United States of America (PNAS), 1992,89(22):10915−10919.
[28] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 1970,48(3):443−453.
[29] Smith T, Waterman MS, Burks C. The statistical distribution of nucleic acid similarities. Nucleic Acids Research, 1985,13(2): 645−656.
[30] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc. of the National Academy Sciences of the United States of America (PNAS), 1988,85(8):2444−2448.
[31] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology, 1990, 215(3):403−410.
[32] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 1994,22(22):4673−4680.
[33] Morgenstern B, Dress A, Werner T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. of the National Academy Sciences of the United States of America (PNAS), 1996,93(22):12098−12103.
[34] Notredame C, Higgins DG. SAGA: Sequence alignment by genetic algorithm. Nucleic Acids Research, 1996,24(8):1515−1524.
[35] Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 2004,32(5): 1792−1797.
[36] Wang GR, Ge J, Xu HY, Zheng RS. A sequence similarity query processing technique based on two-partitioning frequency

transformation. Journal of Software, 2006,17(2):232−241 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/232.htm

[37] Kahveci T, Singh AK. An efficient index structure for string databases. In: Apers P, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass RT, eds. Proc. of the 27th Int'l Conf. on Very Large Data Bases (VLDB). Rome: Morgan Kaufmann Publishers, 2001. 351−360.

[38] Colin M, Jignesh MP, Shniti K. OASIS: An online and accurate technique for local alignment searches on biological sequences. In: Freytag JC, Lockemann PC, Abiteboul S, Carey MJ, Selinger PG, Heuer A, eds. Proc. of the 29th Int'l Conf. on Very Large Data Bases (VLDB). Berlin: Morgan Kaufmann Publishers, 2003. 910−921.

[39] Bajcsy P, Han JW, Liu L, Yang J. Survey of biodata analysis from a data mining perspective. In: Wang JTL, Zaki MJ, Toivonen HTT, Shasha DE, eds. Proc. of the Data Mining in Bioinformatics. London: Springer-Verlag, 2005. 9−39.

[40] DMGroup, China. 2007. http://www.dmgroup.org.cn

[41] Zhu JQ. The research on a data mining platform and it's key technologies [Ph.D. Thesis]. Shanghai: Fudan University, 2002 (in Chinese with English abstract).

[42] Li R, Zhang ZP, Cao SL, Zhu YY, Li YX. A scalable data mining architecture for bioinformatics. In: Ebecken NFF, Brebbia CA, Zanasi A, eds. Proc. of the Data Mining IV. Wessex Institute of Technology Press, 2003. 583−592.

[43] Deng XB, Zhu YY. L-tree match: A new data extraction model and algorithm for huge text stream with noises. Journal Computer Science and Technology, 2005,20(6):763−773.

[44] Deng XB, Zhu YY. ReDE: A regular expression-based method for extracting biological data. Journal of Computer Research and Development, 2005,42(12):2184−2191 (in Chinese with English abstract).

[45] Xiong Y, Zhang R, Chen Y, Zhu YY. Research on biological sequence database management system. In: Huang DS, Liu HY, Shi YY, Chen GL, eds. The Study of Intelligent Computing Theory and Methodology in Bioinformatics. Hefei: University of Science and Technology of China Press, 2006. 134−138 (in Chinese with English abstract).

[46] Cao SL, Qin L, He ZW, Zhong Y, Zhu YY, Li YX. Semantic search among heterogeneous biological databases based on gene ontology. Acta Biochimica et Biophysica Sinica, 2004,36(5):365−370.

[47] Li R, Cao SL, Li YY, Tan H, Zhu YY, Zhong Y, Li YX. A measure of semantic similarity between gene ontology terms based on semantic pathway covering. Progress in Natural Science, 2006,16(7):721−726.

[48] Xiong Y, Zhu YY. A multi-supports-based sequential pattern mining algorithm. In: Wei D, Xie Z, Wang H, Shi B, eds. The 5th Int'l Conf. on Computer and Information Technology Proc. Washington: IEEE Computer Society, 2005. 170−174.

[49] State Key Laboratory of Bioelectronics, China. 2007. http://www.lmbe.seu.edu.cn/lmbenew/content.jsp

[50] Centre for Bioinformatics, China. 2007. http://www.cbi.pku.edu.cn/

[51] BioDW, China. 2007. http://www.scbit.org/services/index.html

[52] Brejova B, DiMarco C, Vinar T, Hidalgo SR. Finding patterns in biological sequences. Technical Report, CS-2000-22, University of Waterloo, 2000.

[53] Brazma A, Jonassen I, Eidhammer I, Gilbert D. Approaches to the automatic discovery of patterns in Biosequences. Journal of Computational Biology, 1998,5(2):279−305.

[54] Smith HO, Annau TM, Chandrasegaran S. Finding sequence motifs in groups of functionally related proteins. Proc. of the National Academy Sciences of the United States of America, 1990,87(2):826−830.

[55] Jonassen I. Efficient discovery of conserved patterns using a pattern graph. Computer Applications in the Biosciences, 1997,13(5):509−522.

[56] Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics, 1998,14(2):229−229.

[57] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals a gibbs sampling strategy for multiple alignment. Science, 1993,262(5131):208−214.

[58] Li M, Ma B, Wang L. Finding similar regions in many strings. In: Proc. of the 31st Annual ACM Symp. on Theory of Computing (STOC'99). New York: ACM Press, 1999. 473−482.

[59] Liang CZ. COPIA: A new software for finding consensus patterns in unaligned protein sequences [MS. Thesis]. University of Waterloo, 2001. 1−22.

[60] Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins, 1990,7(1):41−51.

[61] Grundy, WN, Bailey, TL, Elkan, CP, Baker ME. Meta-MEME: Motif-Based hidden Markov models of protein families. Computer Computer Applications in the Biosciences, 1997,13(4):397−406.

[62] Nevill MCG, Wu TD, Brutlag DL. Highly specific protein sequence motifs for genome analysis. Proc. of the National Academy Sciences of the United States of America (PNAS), 1998,95(11):5865−5871.

[63] Yan XF, Han JW, Afshar R. CloSpan: Mining closed sequential patterns in large datasets. In: Barbara D, Kamath C, eds. Proc. of the 3rd SIAM Int'l Conf. on Data Mining. San Francisco, 2003. 166−177.

[64] Wang JY, Han JW. BIDE: Efficient mining of frequent closed sequences. In: Proc. of the 20th Int'l Conf. on Data Engineering.

Washington: IEEE Computer Society, 2004. 79−90.
[65] Tzvetkov P, Yan XF, Han JW. TSP: Mining top-*k* closed sequential patterns. Knowledge and Information Systems, 2005,7(4): 438−457.
[66] Cong S, Han JW, Padua DA. Parallel mining of closed sequential patterns. In: Proc. of the 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining. New York: ACM Press, 2005. 562−567.
[67] Shapiro JA, Sternberg RV. Why repetitive DNA is essential to genome function. Biological Reviews, 2005,80(2):227−250.
[68] Makalowski W. Not junk after all. Science, 2003,300(5623):1246−1247.
[69] Int'l Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, 431(7011): 931−945.
[70] Benson G. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Research, 1999,27(2):573−580.
[71] Apostolico A, Prefarata F. Optimal off-line detection of repetitions in a string. Theoretical Computer Science, 1983,22(3): 297−315.
[72] Kolpakov R, Kucherov G. Finding maximal repetitions in a word in linear time. In: Proc. of the 1999 Symp. on Foundations of Computer Science. Washington: IEEE Computer Society, 1999. 596−604.
[73] Wang D, Wang G, Wu QQ, Chen BC. Finding LPRs in DNA sequence based on a new index SUA. In: Proc. of the IEEE 5th Symp. on Bioinformatics and Bioenginerering (BIBE 2005). Washington: IEEE Computer Science, 2005. 281−284.
[74] Delgrange O, Rivals E. STAR: An algorithm to search for tandem approximate repeats. Bioinformatics, 2004,20(16):2812−2820.
[75] Kolpakov R, Kucherov G. Finding repeats with fixed gap. In: Proc. of the 7th Int'l Symp. on String Processing and Information Retrieval (SPIRE). Washington: IEEE Computer Society, 2000. 162−168.
[76] Kolpakov R, Kucherov G. Finding approximate repetitions under hamming distance. Theoretical Computer Science, 2003,303(1): 135−156.
[77] Krishnan A, Tang F. Exhaustive whole-genome tandem repeats search. Bioinformatics, 2004,20(16):2702−2710.
[78] Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: The manifold applications of repeat analysis on a genomic scale. Nucleic Acids Reseach, 2001,29(22):4633−4642.
[79] Landau GM, Schmidt JP, Sokol D. An algorithm for approximate tandem repeats. Journal of Computational Biology, 2001,8(1): 1−18.
[80] Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 1999,15(7):563−577.
[81] Pevzner PA, Sze SH. Combinatorial approaches to finding subtle signals in DNA sequences. In: Proc. of the 8th Int'l Conf. Intelligent System for Molecular Biology. San Diego: AAAI Press, 2000. 269−278.
[82] Li DD, Wang ZZ, Ni QS. An effective algorithm for repeat sequence finding. Bioinformatics, 2005,3(4):163−166 (in Chinese with English abstract).
[83] Wang D, Zhao Y, Chen BC, Wang GR. SUA-Based algorithm for finding SATRs in DNA sequence. Journal of Northeastern University (Natural Science), 2007,28(2):209−212 (in Chinese with English abstract).
[84] Wang D, Wang GR, Chen BC, Wu QQ, Wang B, Han DH. A new light weight index SUA for biological sequence analysis. Journal of Huazhong University of Science and Technology (Nature Science Edition), 2005,33(z1):184−188 (in Chinese with English abstract).
[85] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Buneman P, Jajodia S, eds. Proc. of the 1993 ACM SIGMOD Int'l Conf. on Management of Data. Washington: ACM Press, 1993. 207−216.
[86] Margaret HD. Data Mining Introductory and Advanced Topics. Beijing: Tsinghua University Press, 2003.
[87] Rodriguez A, Carazo JM, Trelles O. Mining association rules from biological databases. Journal of the American Society for Information Science and Technology, 2005,56(5):493−504.
[88] Kiem H, Phuc D. Discovering motiv based association rules in a set of DNA sequences. In: Proc. of the 2nd Int'l Conf. on Rough Sets and Current Trends in Computing. London: Springer-Verlag, 2000. 386−390.
[89] Guan JW, Liu DY, Bell DA. Discovering motifs in DNA sequences. Fundamental Informaticae Archive, 2004,59(2-3):119−134.
[90] Koh JLY, Lee ML, Khan AM, Tan PTJ, Brusic V. Duplicate detection in biological data using association rule mining. In: Boulicaut JF Esposito F, Giannotti F, Pedreschi D, eds. Proc. of the ECML/PKDD Workshop on Data Mining and Text Mining for Bioinformatics. LNCS 3202, Pisa: Springer-Verlag, 2004. 35−41.
[91] Gusfield D. Algorithms on Strings, Trees, and Sequences. New York: Cambridge University Press, 1997. 525−532.
[92] Lopresti D, Tomkins A. Block edit models for approximate string matching. Theoretical Computer Science, 1997,181(1):159−179.
[93] Muthukrishnan S, Sahinalp S. Approximate nearest neighbors and sequence comparison with block operations. In: Proc. of the 32nd Annual ACM Symp. on Theory of Computing. New York: ACM Press, 2000. 416−422.
[94] Guralnik V, Karypis G. A scalable algorithm for clustering sequential data. In: Proc. of the IEEE Int'l Conf. on Date Mining. Washington: IEEE Computer Society, 2001. 179−186.
[95] Morzy T, Wojciechowski M, Zakrzewicz M. Scalable hierarchical clustering method for sequences of categorical values. In: Cheung DW, Williams GJ, Li Q, eds. Proc. of the 5th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD).

LNCS 2035, Hong Kong: Springer-Verlag, 2001. 282−293.
[96] Cheng Y, Church G. Biclustering of expression data. In: Proc. of the 8th Int'l Conf. on Intelligent System for Molecular Biology. AAAI Press, 2000. 93−103.
[97] Liu JZ, Yang J, Wang W. BiClustering in gene expression data by tendency. In: Proc. of the Computational Systems Bioinformatics Conf. Washington: IEEE Computer Society, 2004. 182−193. http://www.cs.unc.edu/~weiwang/paper/CSB04_1.pdf
[98] Wang H, Wang W, Yang J, Yu PS. Clustering by pattern similarity in large data sets. In: Franklin MJ, Moon B, Ailamaki A, eds. Proc. of the 2002 ACM SIGMOD Int'l Conf. on Management of Data. Madison: ACM, 2002. 394−405.
[99] Gelfand MS, Mironov AA, Pevzner PA. Gene recognition via spliced alignment. Proc. of the National Academy Sciences of the United States of America (PNAS), 1996,93:9061−9066.
[100] Sze SH, Roytberg MA, Gelfand MS, Mironov AA, Astakhova TV, Pevzner PA. Algorithms and software for support of gene identification experiments. Bioinformatics, 1998,14(1):14−19.
[101] Wang JTL, Rozen S, Shapiro BA, Shasha D, Wang ZY, Yin M. New techniques for DNA sequence classification. Journal of Computational Biology, 1999,6(2):209−218.
[102] Deshpande M, Karypis G. Evaluation of techniques for classifying biological sequences. In: Cheng MS, Yu PS, Liu B, eds. Proc. of the 6th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining. LNCS 2336, London: Springer-Verlag, 2002. 417−431.
[103] Maddouri M, Elloumi M. A data mining approach based on machine learning techniques to classify biological sequences. Knowledge-Based System, 2002,15(4):217−223.
[104] Duda R, Hart P. Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973. 12−26.
[105] Hawkins D. Identification of Outliers. London: Chapman and Hall, 1980. 2−26.
[106] Sun P, Chawla S, Arunasalam B. Mining for outliers in sequential databases. In: Ghosh J, Lambert D, Skillicorn DB, eds. Proc. of the 6th SIAM Int'l Conf. on Data Mining. Bethesda, 2006. 94−105. http://www.siam.org/meetings/sdm06/proceedings/009sunp.pdf
[107] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. Ultraconserved elements in the human genome. Science, 2004,304(5675):1321−1325.
[108] Davidson B, Borchert G. Scientists explore function of 'Junk DNA'. 2006. http://www.sciencedaily.com/releases/2006/11/061113180029.htm
[109] Suurkula J. Over 95 percent of DNA has largely unknown function. 2004. http://www.psrast.org/junkdna.htm

附中文参考文献:

[36] 王国仁,葛健,徐恒宇,郑若石.基于二分频率变换的序列相似性查询处理技术.软件学报,17(2):232−241. http://www.jos.org.cn/1000-9825/17/232.htm
[41] 朱建秋.数据挖掘应用平台及其关键技术研究[博士学位论文].上海:复旦大学,2002.
[44] 邓绪斌,朱扬勇.ReDE:一个基于正则表达式的生物数据抽取方法.计算机研究与发展,2005,42(12):2184−2191.
[45] 熊赟,张锐,陈越,朱扬勇.生物序列数据库管理系统研究.见:黄德双,刘海燕,施蕴渝,陈国良,编.生物信息学中的智能计算理论与方法研究.合肥:中国科学技术大学出版社,2006.134−138.
[82] 李冬冬,王正志,倪青山.一种有效的重复序列识别算法.生物信息学,2005,3(4):163−166.
[83] 王镝,赵毅,陈白尘,王国仁.DNA 序列中基于后缀数组索引的 SATR 查找算法.东北大学学报(自然科学版),2007,28(2):209−212.
[84] 王鏑,王国仁,陈白尘,吴青泉,王斌,韩冬红.一种可用于生物序列分析的轻量级索引结构.华中科技大学学报(自然科学版),2005,33(z1):184−188.

朱扬勇(1963－),男,浙江金华人,博士,教授,博士生导师,主要研究领域为数据挖掘,生物信息学,数据库.

熊赟(1980－),女,博士生,主要研究领域为数据挖掘,生物信息学,数据库.