序列分析技术在 DNA 序列挖掘中的应用①

The Application of Sequence Analysis for DNA Sequence

罗春雨 毛国君 邱洪君

(北京市多媒体与智能软件重点实验室、北京工业大学计算机学院 100022)

摘要: 数据挖掘技术在生物科学领域的应用越来越广泛,本文介绍了序列分析技术在 DNA 序列分析上的应用,对相应的算法做了综述研究,讨论了这些算法的优缺点。最后对 DNA 序列分析的前景作了展望。

关键词: 数据挖掘 序列分析 分片 DNA

1 引言

二十一世纪是生物技术的时代,生物技术是当前最热门的研究领域之一。人类基因组工程有"生命登月计划"之称,它的任务是破译人类分布在细胞核中的 23 对染色体上的约 6 万至 10 万个基因,约 30 亿个碱基,而这些碱基构成了人类基因草图。

人类基因草图是由 4 个字符 A(腺嘌呤)、G(鸟嘌呤)、T(胸腺嘧啶)、C(胞嘧啶) 按一定顺序排成的长约 30 亿的序列,其中没有断句,也没有标点符号。目前除了知道这 4 个字符表示 4 种碱基外,人们对它知之甚少。破译基因图是二十一世纪最重要的任务之一。DNA 全序列具有什么结构、由这 4 个字符排成的看似随机的序列中隐藏着什么规律、各个碱基的功能等都有待于进一步研究。人类基因草图绘制后,接下来的任务是寻找各种基因的精确位置。目前,科学家们已发现了 DNA 序列中的一些规律性与结构。例如,由这 4 个字符组成的 64 种不同的三字符串,其中大多数是用于编码构成蛋白质的 20 种氨基酸;而在不用于编码的蛋白质的序列片段中,A 和 T 的含量特别多些,等等。这些发现让人们相信,DNA 序列中存在着局部性与全局性的结构[1]。

人类在生命科学中不断取得突破,积累了大量的生物数据,提供了揭开生命奥秘的数据基础。而这些丰富的生物数据,由于种类多,维数高,本质上具有异质性与网络性,远远超出传统的分析方法的能力和速度,生物数据的分析成为生物研究的瓶颈,其处理、挖掘、分析和理解的要求日益迫切。所以,研究人员把数据挖掘技术引入到生物学领域。

数据挖掘就是从数据库中抽取隐含的、以前未知的、具有潜在应用价值的信息的过程[2]。数据挖掘技术在生物信息学中的应用具有重要的科学和应用价值,主要体现在以下几个方面:

(1) 由于 DNA 序列的局部和全局的结构性,充分挖掘序列的结构对理解和破译 DNA 序列有十分重要的意义。

(2) 分子生物学结合信息技术产生了生物信息学这一崭新领域。生物数据挖掘等决策支持技术因其在大规模数据处理方面的卓越能力而在其中占据越来越重要的地位。

(3) 现代生物学的研究与发展,越来越离不开信息技术的支持,生物学的每次重大发现都离不开信息技术的支持。

(4) 现代生物研究更多地依赖信息技术的分析结果提供进一步研究的线索和依据,强有力的数据处理分析工具成为现代生物科学研究发展的关键。

2 算法综述

DNA 序列分析中,基于 DNA 的可信的基因检测是可计算基因发现的关键。DNA 序列可分为编码区和非编码区,DNA 中编码区常常是基因所在的区域,因此,

① 基金项目:本文得到国家自然科学基金(No.60173014);北京市自然科学基金(No.4022003)资助

对 DNA 序列进行分片对研究基因等有重要意义。随着 DNA 序列研究的不断深入，研究者已提出多种方法用于 DNA 序列分片。下文，给出了比较有代表性的一些算法。

2.1 最大相似度方法

在生命科学领域，一些生物学家为了科研的需要，把一些简单的 DNA 序列分片思想引入的 DNA 序列分析中。在前人的基础上，[Fu and Curnow, 1990][3] 提出了最大相似度计算的分段方法，其早期目标是想用统计学的方法来找到蛋白质的二级结构。后来，他们发现，这种方法能较好的应用到 DNA 序列的划分上，其做法是：通过计算最大相似度来估计在最小长度上作了限制的经过改造的片段数。用 $Y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})$ 表示第 i 个核苷，$Yi = (1,0,0,0)$ 表示第 i 个核苷为 A，Yi 服从独立多项分布，概率分布向量为 $P = (P_1, P_2, P_3, P_4)$，这时 Yi 的概率为：$P[Y_i = y_i] = p_1^{y_{i1}} p_2^{y_{i2}} p_3^{y_{i3}} p_4^{y_{i4}}$。定义序列 $c_i (i = 1, \cdots, n)$，$c_i = 0$ 表示无变化片段，$c_i = 1$ 表示变化片段。定义似然率 $\ln LR = \ln \frac{\prod_{i=1}^{n} P[Y_i = y_i | c_i]}{\prod_{i=1}^{n} P[Y_i = y_i | c_i = 0]} = \sum_{k=1}^{l} \sum_{i=m_k}^{i=n_k} \ln \left(\frac{\prod_{i=1}^{n} P[Y_i = y_i | c_i = 1]}{P[Y_i = y_i = y_i | c_i = 0]} \right)$，$m_k, n_k, k = 1, \cdots l$，表示 l 个变化片段第 k 个片段的开始和结束。对任意长度的片段 $S = (Y_m, \ldots, Y_n)$，定义函数 $f(S) = \sum_{i=m}^{n} \ln \left(\frac{\prod_{i=1}^{n} P[Y_i = y_i | c_i = 1]}{P[Y_i = y_i | c_i = 0]} \right)$。此时，对于变化片段 $S_k, k = 1, \cdots, l$，log 相似度记为：$\ln LR = \sum_{k=1}^{l} f(S_k)$，这时，可以通过计算最大相似度找出变化片段。

为了计算和查找最大相似度的片段，可能需要把序列划分成需要的结构。这种强制划分的方法需要大量的计算，并且在序列很长或要改造的片段较多的情况下，这个结果是不切实际的。为了更好地适用于 DNA 序列，Fu. 和 Curnow 进一步提出先预定给定变化片段的最小长度，有效的改善了最大相似度算法的效率。其基本步骤为：1) 找出不和 l − 1 个中任意两个片段不重叠的最佳片段；2) 找出最佳分割并且扩展每 l − 1 个片段；3) 从 1)、2) 选出相似度增长达的片段。

2.2 马尔科夫链模型

[Churchill, 1992][4] 提出隐含马尔科夫链来建模 DNA 序列，并且预测在线粒体或抗菌素基因组中分片的位置。隐含马尔科夫模型假定可以把不同的片段划分到一个确定的有限状态中。在每个状态中，核苷数据服从某概率分布，这些状态以低概率从一种状态转换到另一种状态，从而形成了一个隐含马尔科夫链。假定 r 个有穷状态，在观察状态下，S_i 服从马克尔科夫分布，转移矩阵为 $\Lambda = (\lambda_{ij})$，转移等式 $P[S_i = s_i | S_i = s] = \prod_{j=1}^{r} \prod_{k=1}^{r} \lambda_{jk}^{s_i s_{i+1}}$。现在，假定观察 $Yi = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})$ 在状态上服从多项分布，这时：$P[Y_i = y_i | S_i = s] \prod_{j=1}^{4} p_{s,j}^{y_{ij}}$，可以获得系统等式为 $P[Y_i = y_i | Y_{i-1} = y_{i-1}, S_i = s] = \prod_{j=1}^{4} \prod_{k=1}^{4} P_{sk}^{y_{i-1,j} y_{ik}}$，这时可以得到平滑等式 $P[S_i = s | Y_1, \cdots Y_n]$ 并且划分在序列中的同质区。Churchill 指出，未知的状态分布和在某一状态的分布可以通过 EM 算法从数据上估计，并且可以用 BIC (Bayesian Information Criterion) 来决定必要的状态数，参数值可以估计，使算法的执行有较高的自动化和好的灵活性。但是，当额外的数据不明确时，形成的结果仍是光滑的，对用户来是不明确的。

[Fickett, Torney and Wolf, 1992][6] 在隐含马尔科夫的基础上，提出了滑动的马尔科夫模型方法，改善了隐含马尔科夫模型的性能。

2.3 贝叶斯方法

[Liu, Lawrence, 1999][7] 对贝叶斯在 DNA 序列分割的应用做了进一步的工作。用硬币游戏来说明其基本思想：投掷两种硬币的组合模拟四个碱基 A, G, C, T，用两种硬币的 n 个投掷来描述 DNA，开始的 A 个硬币组成的片段中，θ_1 为正面朝上的概率，剩下的 n − A 正面朝上的概率为 θ_2，把硬币 A 作为变异点，A 的先验概率为 1/(n+1)，当中经过蒙特卡罗法近似之后，得到 A 的后验概率。某点的后验概率最大，则作为真正的变异点，对序列进行划分。

2.4 分治法 (divide − and − conquer approach)

经验熵分类方法是一种被广泛应用的一种分类方法，主要用于对有限的字符资源分类。在 DNA 序列分析上，是一种把 DNA 序列中不同性质部分处理成同类的子序列。

尽管经验分类中的所有的字母和匹配规则能方便的用于描述 DNA 序列，仍不能识别在已经获得的范围内的所有生物学功能。[Pedro Bernaola − Galva n, Ramo n Roma n − Rolda n 等, 1995][8] 将分治法用于 DNA 序列分片中。分治法是基于 Jensen − Shannon 经验熵的方法。

分治法以全局的观点描述长序列，比以前基于滑

动窗口的分片法能更好的描述长序列的全局特征。基本思想:序列的每个位置把长序列分为左右两个子序列,通过计算每个位置的 Jensen – Shannon 经验熵,选取最大值点作为一个边界点,递归调用此方法,完成长序列的分片。当 Jensen – Shannon 经验熵低于所建立最小意义的 Jensen – Shannon 熵时,不在进行分片,算法终止。在此算法中,基于 2 个或 4 个特征字符,即:{R(A or G), Y(C or T) or S(G or C), W(A or T)}或{A,G,C,T}。

[Pedro Bernaola – Galván, Ivo Grosse 等,1999]进一步改进基于经验熵的分片方法来检测编码区和非编码区的边界。这是基于 12 个特征字符的。这种方法查找编码区和非编码区的边界时有很高的精度,并且不需要在已知的数据集上进行训练。

在划分自然 DNA 序列时,一般来说有多个编码区和非编码区组成,当这些区域很多时,复杂性非常高,采用了启发式的算法来解决这个问题。通过滑动指针在每个位置上移动,把队列分为两个部分,当 C 值达到某一个给定的极限 S 时,认为所在点为一个边界。在没有超过极限 S 的有意义的值时,算法终止。

[Wentian Li,2000]改善了 BIC,AIC 用于 DNA 分片的停止条件。[Daniel Nicorici, Jaakko Astola,2003]把在分治法和 Wentian Li 工作的基础上,对分治法做了进一步的改善,该算法扩展成 18 个字符的算法,进一步改善了算法的性能。

分治法中,由于在序列的每个位置都要进行计算,以全局的观点描述序列,因此有很高的精度;分治法运算采用递归调用的处理方法,在满足停止条件是停止递归调用,这样方法处理简单,但由于递归调用要占用大量的系统资源,当序列长度巨大时,递归调用的开销是非常巨大的,要采用合适的处理方法来解决分治法处理大序列的问题。

2.5 MDL(Minimum Description Length)方法

[Wojciech Szpankowski, Wenhui Ren and Lukasz Szpankowski,2003]把 MDL 规则引入到 DNA 分片中。$X_N^1 = X_1 X_2 \cdots\cdots X_N$ 该序列由有限的字符集 A 组成。算法主要思想:一个给定的 DNA 序列{"TAGCATGCTG AGGGATCTAG CAGGTTGAXX ……"},首先对 DNA 序列进行分组替换,把 A 和 G 分为嘌呤组 R = {A,G},把 C 和 T 分为嘧啶组 S = {C,T},替换后,将 DNA 序列划分为等长的片段 $b_1\ b_2\ b_3\ b_4 \cdots\cdots b_K$,其中 $|b_1| + |b_2| + |b_3| + \cdots\cdots |b_K| = N$,$|b_i|$ 表示第 i 块的长度 b($1 \leq i \leq K$, $b * K = N$, N 为 DNA 序列的长度)。然后,根据 Stein—Ziv 定理通过构造最优的判别式函数计算每个片段的经验熵,在所有片断的经验熵中,选择最优值作为一个变化点,即编码区和非编码区的边界。在第一次选择过程中,可能由于片段长度较长,无法准确知道该边界的确切位置,需要采用同样的方法在某一片段内进行二级分片,以确定变化点的确切位置。

MDL 算法采用最优的分片长度,因而有较高的效率,当选取片段长度为 $|bi| = \log(N)$ 时,算法的时间复杂度为 O(N)。MDL 算法实现简单,效率高,但也存在着一些问题:不同分片的交集为空,简化了数据处理,但也存在着一个问题,就是在划分片段时,可能把同质的编码区或非编码区划分到了两个片段,这时可能会造成实际存在的变化点(编码区和非编码区的边界)丢失或产生多余的变化点。

除以上方法外,还有其他的 DNA 序列分析方法,比如间隔序列——模式过滤的方法,神经网络方法,聚类方法也被引入 DNA 序列挖掘中。

在我国,生物信息的数据挖掘研究还刚刚起步,中科院等少数研究机构也开始了这一领域的研究。

3 算法优缺点比较研究

从理论基础上说,MDL 方法和分治法都是基于 Jensen – Shannon 熵的计算的。在传统的 MDL 算法中,采用等长的分片方法,这种方法数据处理简单,分片的效率高,但存在着精度较低的缺点。只能找到边界点在某一片断内,不能找到具体的位置。在传统的分治法中,由于在每个位置都要进行计算 Jensen – Shannon 熵,找到一个最大值点作为边界点,因此,能够找到比较精确的位置,但此方法的运算量大,效率较低。

我们将两种算法的结合起来:首先,用 MDL 算法将要分析的序列分成较小的序列,初步查找变化点的位置,其次,用分治法根据需要进行进一步查找变化点的位置。在查找过程中,如果序列长度很大,可以用 MDL 方法进行多级分片,再用分治法进行计算。

在 MDL 算法中,采用如下的分片规则:$A^N = X_1 X_2 \cdots\cdots X_N$,该序列由有限的字符集 A 组成。我们把长度为 N 的序列分为 K 块,即 $b_1\ b_2\ b_3\ b_4 \cdots\cdots b_K$,其中 $|b_1|$

$+|b_2|+|b_3|+\cdots\cdots|b_K|=N$，$|b_i|$表示第 i 块的长度 $b(1\leq i\leq K, b*K=N)$。可以看出，不同分片的交集为空，简化了数据处理，但也存在着一个问题，就是在划分片段时，可能把同质的编码区或非编码区划分到了两个片段，这时可能会造成实际存在的变化点（编码区和非编码区的边界）丢失或产生多余的变化点。可能发生的情况有：(1) 在实际 DNA 序列中，相邻片段边界附近不存在变化点，在算法运算过程中，却找到了变化点，这时就产生了多余的变化点。(2) 在实际 DNA 序列中，相邻片段存在变化点，由于片段的划分，使算法无法找到实际存在的变化点，造成变化点的丢失。这些错误极其可能是由于在等长划分的情况下，由于相邻片段的交集为空，这时，恰好把性质相同的片段分割开来，划分到了不同的片段，从而导致了以上情况的发生。在 MDL 方法中，如果采取相邻片段之间的交集不为空，可能部分解决变化点丢失和增加的问题。当然，处理的复杂程度和方法需要改进。

在以上的算法中，实验都是基于一段给定的 DNA 序列中，但在现实的科研中，由于种种原因，许多 DNA 序列并没有被完全解密，也就是说，许多 DNA 序列是不完整的，比如一段 DNA 信息如下："TAGCATGCTG AGGGATCTAG CAGGTTGAXX GATCTAGCAT"（chromosome 9 9q34, human），其中 A、G、C、T 表示组成 DNA 的四种基本核苷，X 表示未知的信息（在此 X 为假定未知），即不知道具体是哪种核苷，造成了 DNA 序列的缺失。尽管缺失的信息只有很少一部分，但对于携带遗传密码的 DNA 来说，是至关重要的，因此，在研究的过程中，不得不考虑这些缺失的信息，以上算法中均未提到如何处理这些缺失的信息。在 DNA 信息缺失的问题上，根据相邻片段或其它的统计信息能够以较大的概率来估计这些缺损信息，减少分片出错的概率。

在目前的大多数 DNA 数据挖掘的算法中，实验基本都是基于一定长度的 DNA 序列的，就是说事先给定了要分析的 DNA 序列的长度。实际上，DNA 序列长度巨大，比如人类的染色体有 30 亿个碱基，不同的物种之间有很大的不同，自然界中可能存在比人类 DNA 序列更长的物种。我们可能无法一次处理如此长度的 DNA 序列，我们需要探索巨大序列如何调入内存处理的问题。随着 DNA 序列挖掘的发展，DNA 序列分片变得越来越重要，提高算法的效率和精度是必然要求。

4　结束语

数据挖掘技术已经成为 DNA 序列的强有力的分析工具。在异构、分布式基因数据库的语义集成，DNA 序列间相似搜索和比较、关联分析、路经分析等方面都取得了一定的成果。此外，DNA 序列是构成一切生命特征的基础，通过分析 DNA 序列，能发现 DNA 序列中的一些特殊信息。同样，构成基因、氨基酸的基本成分都是 DNA 序列中的四种基本核苷，通过对 DNA 序列进行分析，也能促进对基因和氨基酸等的研究。生物信息的数据挖掘还处于起始阶段，随着生物技术领域的不断突破，数据挖掘技术会越来越多的应用到生物技术领域；同时，数据挖掘技术的广泛应用也会不断的促进生物信息学的发展，推动生物科技的进步。

参考文献

1 吴家睿，'生命之书'的解读，http://202.116.43.3/life/teacher/zhourc/life2004/ebook/lifesource/ .

2 Jiawei Han, Micheline Kamber. Data Mining—Concepts and Techniques[M]，北京 高等教育出版社，2001。

3 Fu, Y. -X、Curnow, R. N. , Maximum likelihood estimation of multiple change points, Biometrika 77 , 563 -573, 1990.

4 Churchill, G. A. , Hidden Markov chains and the analysis of genome structure, Computer in Chemistry 16, 107 -115, 1992.

5 Jerome V. Braun and Hans -Georg Muller, Statistical Methods for DNA Sequence Segmentation, Statistical Science, Vol. 13, No. 2, 142 -162, 1998.

6 Fickett, J. W. , Torney, D. C. and Wolf, Base compositional structure of genomes, 1056 - 1064, Genomics 13, 1992.

7 Tun S. Liu, Charles E. Lawrence, Bayesian inference on biopolymer models, Bioinformatics 15, 38 - 52 1999.

8 Pedro Bernaola -Galvan, R. Roman -Rold and J. L. Oliver, Compositional Segmentation and Long -Range Fractal Correlations in DNA Sequences, Phys. Review E, 53, 5181 -5189, 1996.