DOI:10.16262/j.cnki.1000-8217.1999.03.007

·成果简介·

用几何学方法分析 DNA 序列

张春霆*

(天津大学生命科学与生物工程研究院,天津 300072)

[关键词] DNA 序列,序列分析,几何学方法,Z 曲线,基因识别,生物信息学

DNA 序列是生物遗传信息的载体,数十亿年来生物进化的历史就记录在这长长的序列之中,隐藏着大自然的奥秘。但是,其意义并不是一目了然的,需要进行分析。传统的分析方法是基于统计学的,基本上属于代数学范畴。然而从笛卡尔时代起,人们就知道代数学与几何学方法是等价的,但各有其特点,相互补充,缺一不可。那么,能不能用几何学方法来分析 DNA 序列呢?经过作者十余年的研究,答案是肯定的。在本项目的早期阶段[1—6],研究工作需要一种工具将 DNA 序列的碱基组成直观地显示出来,为此我们从分析碱基的对称性入手。我们注意到 4 种碱基按不同标准可以分成 3 对:嘌呤-嘧啶;氨基-酮基;强氢键-弱氢键。参照物理学中的一贯做法,我们寻求若干几何体来表示这种对称性。结果正六面体正好满足要求。按上下、左右和前后,正六面体的 6 个面正好分成 3 对。不失一般性,把左、右面分别称为嘌呤、嘧啶面;把前、后面分别称为氨基、酮基面;而把上、下面分别称为强、弱氢键面。由于每一对的每一方均由 2 个碱基组成(例如嘌呤由 A 和 G 组成,嘧啶由 C 和 T 组成等),所以可在正六面体的顶角上适当地标上 A,C,G,T,4 个字母以反映这一事实,最后得到一个正四-六面体体系。这一几何体系准确地反映了 DNA 序列中 4 种碱基(或核苷酸)的对称性,是本项目研究的基础。基于这一体系,利用几何学知识做些数学推导,很快就得到一个显示碱基组成的图形方法。以上想法一提出来就得到国外同行的高度评价,并推荐在诸如《Nuclear Acids Research》这样的核酸研究权威刊物上发表[7]。利用这一方法来分析大量基因的碱基组成,很快发现 4 种碱基出现频率之和恒小于 1/3,很少有例外[7,8]。曾应用这一图形法来分析 10 余种生物的基因组中碱基的频率分布与密码子选用。但限于工作繁忙,大部分工作尚未来得及整理发表,已发表的有 HIV[9],人类[10,11]以及大肠杆菌[12]。研究结果显示,不同物种基因组中的密码子选用和碱基组成都有所不同,而图形法为研究这些差异提供了方便,因而可用于研究分子进化[12]。图形法在蛋白结构类预测[13],反义核酸[14]和氨基酸的亲、疏水性与分类[15]等问题上均获得了具体的应用。图形法与信息熵的结合,产生了诸如信息流矢量、等信息熵面等新概念,并用来研究分子进化[16]。后来,应国外权威学者的邀请,作者写了 1 篇综述文章,总结了图形法及其研究成果[17]。

到 1994 年这项研究又取得了新的突破。在四—六面体体系中建立了一个特殊坐标系,可将任一 DNA 序列一一对应地映射为一三维空间曲线,称之为 Z 曲线[18]。由于 Z 曲线是 DNA 序列的一个等价表示,它携带了 DNA 序列的全部信息,故对 DNA 序列的研究可转化为对 Z 曲线的研究。Z 曲线在 3 个坐标轴上的投影有着明确的生物学意义。其中 x 分量表征着嘌呤/嘧啶碱基沿序列的分布;y 分量表征着氨基/酮基碱基沿序列的分析;而 z 分量则表征着强氢键/弱氢键碱基沿序列的分布。这 3 种独立的分布完整地描述了所研究的 DNA 序列[18]。这种新颖的学术观点为引进更多的数学工具来分析 DNA 序列提供了广阔的前景。在 Z 曲线的框架下,我们发现 DNA 序列的对称性是由 4 阶交代群所描述。利用有限群的表示论我们建立了 DNA 序列的对称性理论[19]。这套理论不仅有学术上的价值,而且在分析限制性内切酶对 DNA 序列的识别等方面

---

* 中国科学院院士.

国家自然科学基金资助项目,批准号 39570187 及 19573002.

本文于 1998 年 12 月 14 日收到.

有重要的应用[19]。一段时期以来,国际上对 DNA 序列的长程关联进行了大量的研究。学者们普遍采用所谓随机行走(Random Walk)方法来将 DNA 序列转变为一数字序列。我们研究发现,"行走"的结果只不过是 Z 曲线的一个分量而已。因为 Z 曲线携带了 DNA 序列的全部信息,故远比一维随机行走所包含的信息要完整。我们用 Z 曲线形式研究了 DNA 序列的长程关联,不仅得到了全新的结果,而且还提出了识别一个序列是否包含内含子的新算法,准确率达 90%以上[20]。Z 曲线理论最重要的应用可能在基因识别方面。对 Z 曲线的 3 个分量分别施行傅里叶变换以探查其周期性,以此来识别人类基因组中的外显子,准确率达到 84.9%[21]。这一方法可与现有的最好的单一方法相媲美。而且,Z 曲线方法在提高基因识别准确率方面潜力还很大。例如,Z 曲线的一阶微分在原核生物非编码区与编码区的边界处,真核生物基因中内含子和外显子交界处,均出现反常行为。显然可以利用这一现象来提高基因识别的准确率。有关工作正在抓紧进行之中。

总之,这项工作的意义在于,这是由中国人立足于国内,用中国人的思维方法所建立起的一套 DNA 序列的完整的理论体系。与众不同之处在于它将坐标系、多面体、投影、曲线、曲线微分等几何学概念与抽象的 DNA 序列建立起紧密的联系,创造出新的知识,在国际上独树一帜,自成一派。而现有的这一切成果仅仅意味着工作刚开始,用几何学方法分析 DNA 序列将有着广阔的发展前景,可能发展成为生物信息学中相对独立的一个分支学科。

参 考 文 献

[1] Zhang C T. Upper limit for the variances of some helical parameters in DNA double helix. Int. J. Biol. Macromol., 1989, **11**:9—12.

[2] Zhang C T, Zhou G F. Analysis of sequences of twist angles in DNA double helix. Int. J. Biol. Macromol., 1989, **11**:165—168.

[3] Zhang C T, Cen Y Q. The study of DNA sequences by their sequences of twist angles. J. Theor. Biol., 1989, **138**:457—465.

[4] Zhang C T, Zhou G F. Analysis of pattern of twist angles in DNA double helix. Int. J. Biol. Macromol., 1990, **12**:225—232.

[5] Zhang C T. Equations between frequencies of amino acids in organism. J. Theor. Biol. 1990, **142**:281—284.

[6] Zhang C T, Shang Z X. The study of stacking energy for natural DNA sequences. J. Theor. Biol., 1991, **149**:257—263.

[7] Zhang C T, Zhang R. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. Nucl. Acids Res., 1991, **19**: 6313—6317.

[8] Zhang C T, Zhang R. Diagrammatic representation of distribution of DNA bases and its applications. Int. J. Biol. Macromol., 1991, **13**: 45—49.

[9] K.C. Chou, Zhang C T. Diagrammatization of codon usage in 339 human immunodeficiency virus proteins and its biological implication. AIDS Res. and Human Retroviruses, 1992, **8**:1967—1976.

[10] Zhang C T, Chou K C. Graphic analysis of codon usage strategy in 1490 human proteins. J. Protein Chem., 1993, **12**:329—335.

[11] Zhang C T, Zhan Y. Analysis on the distribution of bases for 1487 human protein coding sequences. J. Theor. Biol., 1994, **167**:161—165.

[12] Zhang C T, Chou K C. A Graphic Approach to Analyzing Codon Usage in 1562 E. Coli Protein Coding Sequences. J. Mol. Biol., 1994, **238**:1—8.

[13] Chou J J, Zhang C T. A joint prediction of the folding types of 1490 human proteins from their genetic codons. J. Theor. Biol., 1993, **161**:251—262.

[14] Zhang C T, Chou K C. An Analysis of Base Frequencies in the Anti-Sense Strands Corresponding to 180 Human Protein Coding sequences. Amino Acids, 1996, **10**:253—262.

[15] Zhang R. Distribution of mapping points of 20 amino acids in the tetrahedral space Amino Acids, 1997, **12**:167—177.

[16] Wang J H, Zhang C T. Studies on the Isentropic Equations of Nucleotide Sequences and Their Application. J. Theor. Biol., 1996, **181**:197—202.

[17] Zhang C T. Diagrammatic representation of base composition in DNA sequences. In: Visualizing Biological Information, Pickover C A ed. Singapore: World Scientific, 1995, 84—95.

[18] Zhang R, Zhang C T. Z Curves, an Intuitive Tool for Visualizing and Analyzing DNA sequences. J. Biomol. Struc. Dynamics., 1994, **11**: 767—782.

[19] Zhang C T. A Symmetrical Theory of DNA Sequences. J. Theor. Biol., 1997, **187**:297—306.

[20] Zhang C T, Lin Z S, Zhang R. A Novel Approach to Distinguish Between Intron-Containing and Intronless Genes Based on the Format of Z Curves. J. Theor. Biol., 1998, **192**:467—473.

[21] Yan M, Lin Z S, Zhang C T. A New Fourier Transform Approach for Protein Coding Measure Based on the Format of the Z Curve. Bioinformatics, 1998, **14**:685—690.

ANALYSIS OF DNA SEQUENCES BY A GEOMETRICAL APPROACH

Zhang Chunting

(*Institute of Life Science and Biotechnology, Tianjin University, Tianjin* 300072)

Key words DNA sequence, sequence analysis, geometrical approach, Z curve, gene identification, Bioinformatics