

文章编号:1673-1689(2008)02-0103-03

用非线性预测方法研究蛋白质序列的特性(Ⅱ)

管维红¹, 徐振源^{2*}, 朱平²

(1. 江南大学 信息工程学院, 江苏 无锡 214122; 2. 江南大学 理学院, 江苏 无锡 214122)

摘要: 为了研究蛋白质序列的内在特性, 通过非线性预测方法将蛋白质序列和随机序列以及混沌序列进行比较。前期研究可知: 每条蛋白质序列的每个特征序列的误差比值($E-D$)图具有特异性, 和随机序列的 $E-D$ 图相比具有明显差异。另外每条蛋白质序列的 $E-D$ 图曲线大体都在 1 附近波动, 波动走势没有任何规律可言。通过与给定混沌序列的 $E-D$ 图相比较, 发现两者之间具有相似的特性。由此得出: 蛋白质序列具有非随机性, 可能具有非线性(混沌)性。

关键词: 混沌性; 随机性; 非线性预测方法; 蛋白质序列

中图分类号: TP 301.6; Q 516

文献标识码: A

Nonlinear Prediction Analysis of Properties in Protein Sequences(Ⅱ)

GUAN Wei-hong¹, XU Zhen-yuan^{2*}, ZHU Ping²

(1. School of Information Technology, Jiangnan University, Wuxi 214122, China; 2. School of Science, Jiangnan University, Wuxi 214122, China)

Abstract: In order to study the properties in protein sequences, three of the characteristic sequence $E-D$ graphs of protein sequences in every four classes with random sequences by using the nonlinear prediction method was investigated. From the latest research, it was known that each characteristic sequences $E-D$ graphs of the protein sequences were specific, and different from that of the random sequences. Furthermore, every curves of protein sequences fluctuated around 1, and the walks of waves had not any pattern. The difference of $E-D$ graphs among four classes of protein sequences was very large. By comparing the $E-D$ graphs of the protein sequences with the chaotic sequences, it was also found similar properties. It was concluded that the protein sequences were non-random, and may be with the nonlinear properties such as chaos.

Key words: chaos; randomness; nonlinear prediction method; protein sequence

蛋白质的结构预测问题是当前分子生物学的一大难题^[1]。为了更好地研究蛋白质序列的结构, 越来越多的研究者开始着手研究氨基酸序列的内

在特性^[1-3]。目前, 已有的研究结果给出的结论并非一致, 有的研究者认为蛋白质序列是随机的, 其他则指出蛋白质序列是非随机。如 White, Ja-

收稿日期: 2007-05-10.

基金项目: 国家自然科学基金资助项目(10372054, 60575038).

作者简介: 管维红(1983-), 女, 江苏连云港人, 计算机应用技术硕士研究生。

* 通讯作者: 徐振源(1946-), 男, 上海人, 理学硕士, 教授, 博士生导师, 主要从事混沌控制与同步、生物信息学研究。

Email: xu_zhenyuan@yahoo.com.cn

cobs^[4]通过对蛋白质链上的亲水性氨基酸的统计分析指出:绝大多数蛋白质序列的亲水性分布和随机性分布的序列没有显著区别。而 Pande 等^[5]利用映射将蛋白质序列映射到随机游走中,认为蛋白质序列不具有纯粹的随机性。另外 Huang 等^[6]则利用非线性预测方法计算 3 种模型下 α , β , $\alpha\beta$ 类每条蛋白质序列的总体误差平均值,由所有序列总体误差平均值的均数说明蛋白质序列是随机的。而在其他研究中发现蛋白质可能具有非线性特性,如 Jura 等认为从 DNA 到蛋白质并不是线性关系,可能用非线性(混沌)描述更为合理^[7]。混沌是一种由确定性非线性系统产生的对于初值极为敏感而具有内禀随机性和长期预测不可能性的往复非周期运动^[8]。它是非线性系统普遍具有的一种复杂的动力学行为,它把表现的无序性与内在的规律性巧妙地融为一体,反映了非线性动力学系统具有的内在随机性,是一种貌似无规则的在确定性系统中出现的类随机的过程。混沌这一复杂现象,很多方面和目前对于蛋白质序列的了解有相似之处。

作者采用非线性预测方法,分别得到随机序列和混沌序列的 E - D 图,进而和 4 类蛋白质序列各自特征序列的 E - D 图相比较。从中发现,蛋白质序列的 E - D 图大体都在 1 附近无规律的波动,表现出非随机特性。而每类蛋白质序列之间也具有差异性,通过与混沌序列的比较,得出蛋白质序列可能具有混沌性的猜测。

1 数据来源及处理

混沌序列取自典型的 logistic 映射系统^[9-10], Logistic 系统在特定的初值和参数取值范围内处于混沌状态,对于不同的初值和参数下所产生的序列有很大差别。这里定义的 logistic 映射^[8]为: $z_{i+1} = l(z_i)$, $l(z) = \mu z(1-z)$, $0 \leq z_i \leq 1$, $0 \leq \mu \leq 4$ (初值 $z = 0.4$), 此处分别在参数 μ 取 3.970, 3.980, 3.997, 3.999, 4.000 时,取出 5 条混沌序列。数字序列长度取 500 (因为一般分类后蛋白质序列长度较短)。序列中每个元素大于等于 0.5 的取为 1, 小于 0.5 的取为 0, 则混沌序列转换为用 (0,1) 表示的符号序列。

随机序列来自 Matlab 软件自动生成的 0,1 随机序列,长度也取 500。

2 结果与讨论

采用非线性预测方法,分别得到随机序列和所选混沌序列关于不同嵌入维数 d 的总体误差比值

E 。以嵌入维数 d 为横坐标,总体误差比值 E 为纵坐标,形成的图称为 E - D 图。

在图 1 中,给出了随机序列的 E - D 图,明显可以看出随机序列的 E 值曲线始终在 1 附近波动,且波动的幅度较小(波动范围: [0.907 59, 1.119 5])。这和已知结论(对一般的随机序列,其总体误差比值 E 不依赖于嵌入维数 d ,可近似为一个常量)相一致。

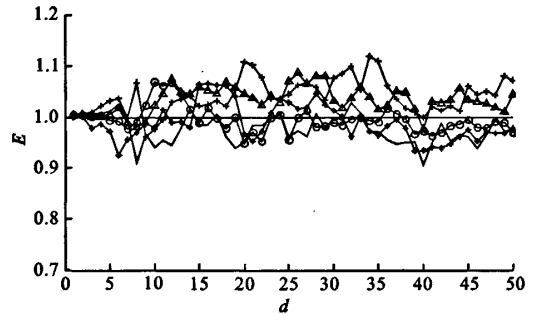


图 1 随机序列图

Fig. 1 The random sequences

从图 2 混沌序列的 E - D 图中发现,随着嵌入维数 d 的变化,混沌序列的 E 值一直在 1 附近波动,对于所选序列来说位于 1 下方的比重较大。图中每条曲线波动的起伏有大有小。已知一般随机序列的 E 值不随嵌入维数 d 的变化而变化,其 E 值可近似为 1。则图中反映的现象和混沌序列内禀随机性,却又含有一定规律性,长期预测不可能性的复杂现象相吻合。那么也可以作这样的推论: E 值在 1 附近波动的序列可能具有一些确定性特性如混沌性。

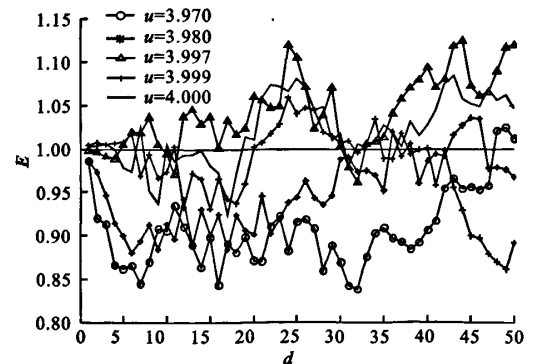


图 2 混沌序列图

Fig. 2 The chaos sequences

在图 3 中,给出了来自 logistic 映射系统的两条混沌序列的 E - D 图以及两条蛋白质序列 (liw1A00, 1mqoA00) 的第一特征序列的 E - D 图的比较。可看出这 4 条序列的 E 值都在 1 附近波动,

且大体走势较为相像,随着 d 的增长,曲线起伏由大变小,但是具体每处的波动却完全不同。

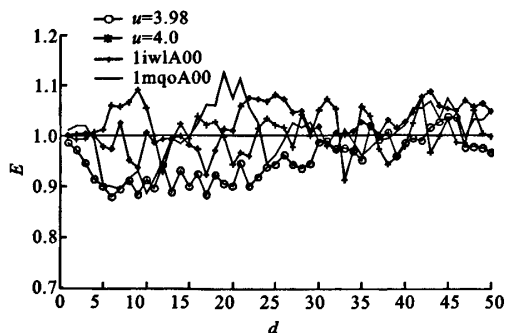


图3 两条混沌序列(参数为 3.98, 4.00)和两条蛋白质序列(1iw1A00, 1mqoA00)的第一特征序列的 E - D 图

Fig.3 E - D graph of two chaos sequences (3.98, 4.00) and the first characteristic sequences of two protein sequences (1iw1A00, 1mqoA00)

3 结 语

根据以上 3 幅图及对应分析,并结合前期结论:每类蛋白质序列的每条特征序列的 E 值曲线在 1 附近起伏波动不断,具有特异性。另外每条曲线的波动走势是没有任何规律可言的,但是却仍都在 1 附近波动,没有跳出这个范围。即蛋白质序列本身看似杂乱的变化过程中,其实可能含有内在的规律性。而来自 logistic 映射系统的 5 条混沌序列的

E - D 图也表现出了相似的特性,虽然两者的比较并没有找到走势完全一致的 E - D 曲线,但这正和混沌的貌似无规则,类似随机而长期预测不可能的特性相吻合。由此得到结论:蛋白质序列具有非随机性,可能具有非线性确定性质(混沌性)。其中 4 类蛋白质序列表现的非线性特性不同:如 α 主类中每条蛋白质序列的 3 个特征序列的 E - D 图走势具有一定相关性,在 1 附近波动也较为均衡。 β 主类的 E - D 图表现出比 α 主类更强的特异性,每条蛋白质序列的 3 个特征序列也不具有明显的相关性。 $\alpha\beta$ 类的多条特征序列的曲线在 1 的上方或者下方波动。低二级结构类的整体波动最为剧烈,最大起伏达到了 0.903 2,对应特性更为复杂。

这里的结论和文献[4]和[6]中的结论不同,文献[4]中指出蛋白质亲水性分布和随机分布没有显著区别,只看到了蛋白质序列中类似随机性的一面。文献[6]中也是如此,将每类蛋白质序列中所有序列的相应 $\langle E_d \rangle$ 值取平均值,这样的统计方法就可能屏蔽掉了原序列中的非线性。文献[5]中也给出了蛋白质序列非随机的结论,但是其给出非随机结论主要是因为忽略了 Pande 模型下分类的不平衡。总体来说用非线性预测方法来研究蛋白质序列的特性是很合理的,更有利于发现其真实特性。

从生物意义上来说,从混沌的角度来解释目前对蛋白质序列的认识是比较吻合的。蛋白质序列可能正像混沌序列的表现无序性与内在规律性一样,在看似随机的表面具有一定的确定性。

[1] 赵国屏. 生物信息学[M]. 北京: 科学出版社, 2004.

[2] 王翼飞, 史定华. 生物信息学——智能化算法及其应用[M]. 北京: 化学工业出版社, 2006.

[3] 宋江宁, 李炜娜. 用氨基酸组分预测含顺式肽键蛋白质[J]. 无锡轻工大学学报(食品与生物技术学报), 2003, 22(3): 7-11.

SONG Jiang-ning, LI Wei-jiang. Predicting proteins containing cis peptide bonds using amino acid compositions[J]. Journal of Wuxi University of Light Industry (Journal of Food of Science and Biotechnology), 2003, 22(3): 7-11. (in Chinese)

[4] WHITE S H, JACOBS R E. Statistical distribution of hydrophobic residues along the length of protein chains[J]. Biophysical Journal, 1990, 57(4): 911-921.

[5] PANDE V S, GROSBERG A Y. Nonrandomness in protein sequences: Evidence for a physically driven stage of evolution? [J]. Biophysics, 1994, 91(12): 12972-12975.

[6] HUANG Y Z, XIAO Y. Nonlinear deterministic structures and the randomness of protein sequences[J]. Chaos, Solitons and Fractals, 2003, 17: 895-900.

[7] JURA J, WEGRZYŃ P. Regulatory mechanisms of gene expression: complexity with elements of deterministic chaos[J]. Acta Biochimica Polonica, 2006, 53(1): 1-9.

[8] 刘延柱, 陈立群. 非线性动力学[M]. 上海: 上海交通大学出版社, 2000.

[9] 林卫强, 黄元石. Logistic 混沌序列的随机性分析[J]. 福州大学学报: 自然科学版, 2004, 32(3): 0270-0275.

LIN Wei-qiang, HUANG Yuan-shi. Analysis for randomness of the series generated by chaos Logistic system[J]. Journal of Fuzhou University: Natural Sciences Edition, 2004, 32(3): 0270-0275. (in Chinese)

[10] 胡文立, 王玖. Logistic 数字混沌序列的性能分析[J]. 桂林电子工业学院学报, 2001, 21(1): 26-30.

HU Wen-li, WANG Jiu. The correlation property of digital chaotic sequence by Logistic maps[J]. Journal of Guilin Institute of Electronic Technology, 2001, 21(1): 26-30. (in Chinese)

(责任编辑: 朱 明)