文章编号:1673-1689(2008)01-0071-05

用非线性预测方法研究蛋白质序列的特性(Ⅰ)

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:通过详细 HP 模型将一条蛋白质序列构造为 3 个特征序列。其中蛋白质序列取自蛋白质分类数据库 CATH,该库将蛋白质序列根据不同结构域分为 4 类(α 主类,β 主类,αβ 类,低二级结构类)。然后采用非线性预测方法,研究每类蛋白质序列的特征序列,得到特征序列的误差比值(*E-D*)图。从图形上发现,每类蛋白质序列的每条特征序列的曲线起伏波动不断,具有特异性,且 4 类蛋白质序列对应特征序列的 *E-D* 图之间也有很大不同。

关键词:非线性预测方法;蛋白质序列;蛋白质分类
中图分类号:TP 301.6;Q 516 **文献标识码**:A

Nonlinear Prediction Analysis of Properties in Protein Sequences(Ⅰ)

GUAN Wei-hong[1], XU Zhen-yuan*[2], ZHU Ping[2]
(1. School of Information Technology, Jiangnan University, Wuxi 214122, China; 2. School of Science, Jiangnan University, Wuxi 214122, China)

Abstract: In this manuscript, three characteristic sequences for a protein sequence were constructed according to the detailed HP model. Protein sequences were selected from the protein structure classification database CATH. The database classified protein into four classes (mainly alpha, mainly beta, mainly alpha－beta and few secondary structures) based on the different domain structures. By using the nonlinear prediction methods, three characteristic sequences E-D graphs of protein sequences in every four classes were obtained. It was found that each characteristic sequences E-D graphs of protein sequences fluctuated all along and specific. Moreover, the difference of E-D graphs for corresponding characteristic sequences among four classes of protein sequences was very large.

Key words: nonlinear prediction method; protein sequence; protein classification

蛋白质中最基本成分是氨基酸序列,氨基酸序列的特性从根本决定了蛋白质的结构和功能[1-2]。蛋白质的结构预测问题是当前分子生物学的一大难题。为了更好的研究蛋白质序列的结构和功能,越来越多的研究者开始着手研究氨基酸序列的内在特性[3]。

当前用于研究序列特性的方法一般是基于统计学理论的,如相关系数,随机游走,信息熵等,这

收稿日期:2007-01-15.
基金项目:国家自然科学基金项目(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. Email: gwh0229@yahoo.com.cn
*通讯作者:徐振源(1946-),男,上海人,理学硕士,教授,博士生导师,主要从事生物信息学研究. Email: xu_zhenyuan@yahoo.com.cn

些方法都取得了一定的成绩[4]。另外由 Barral[5]，Huang[4]等提出用非线性预测方法来研究 DNA 序列特性，该方法曾成功地用于区分时间序列中混沌和噪声，可以归纳特定的信息，并且对于短序列也可以给出很好的结论[6]。考虑到蛋白质序列的无规律性，且一般蛋白质序列尤其是分类后的都很短，作者采用非线性预测方法研究蛋白质序列的特性。用非线性预测方法研究 DNA 序列的特性，已经有不少研究[5,7-9]，如文献[5]中利用非线性模型研究 DNA 链得出 DNA 链具有非线性确定结构；文献[9]则将非线性预测方法加以改进研究人类的4737个启动子序列发现其中也具有非线性确定结构。作者首先将每条蛋白质序列通过详细 HP 模型，转化为3个特征序列，进而采用非线性预测方法分别研究每类蛋白质序列的特征序列的 *E-D* 图。

1 非线性预测方法

采用的非线性预测方法[4,7-9]。对于任意符号序列 $x_1,x_2,\cdots,x_N$，构建其 d 维向量集(d 为嵌入维数)：

$$\begin{aligned}&\boldsymbol{X}_1 \equiv (x_1,x_2,\cdots,x_d),\\&\boldsymbol{X}_2 \equiv (x_2,x_3,\cdots,x_{d+1}),\\&\quad\vdots\qquad\qquad\vdots\\&\boldsymbol{X}_{N-d} \equiv (x_{N-d},x_{N-d+1},\cdots,x_{N-1}),\\&\boldsymbol{X}_{N-d+1} \equiv (x_{N-d+1},x_{N-d+2},\cdots,x_N),\end{aligned}$$

得到所有可能的 d 长的符号子序列，然后对其中每个向量 $\boldsymbol{X}_p=(x_p,x_{p+1},\cdots x_{p+d-1}(1\leqslant p\leqslant N-d)$ 寻找其最近邻 $\boldsymbol{X}_{H(p)}\equiv(x_{H(p)},x_{H(p)+1},\cdots x_{H(p)+d-1})$。其中 $\boldsymbol{X}_p$ 和 $\boldsymbol{X}_{H(p)}$ 的相似度通过以下方法定义：一对符号 x_i 和 x_j 的相似度定义为海明(Hanmming)距离：$h(x_i,x_j)=\begin{cases}0 & x_i=x_j\\1 & x_i\neq x_j\end{cases}$；则向量 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 的相似度定义为：$H(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sum_{k=0}^{d-1}h(x_{i+k},x_{j+k})$。给定 $\boldsymbol{X}_p$ 的最近邻 $\boldsymbol{X}_{H(p)}$ 即为使 $H(X_p,X_j)$ 值最小的 X_j，其中 $j\neq p$。一旦最近邻 $X_{H(p)}$ 被确定，就可以计算局部误差平均值，对于多个最近邻：$\varepsilon_p=<h(x_{p+d},x_{H(p)+d}>$，$<\cdot>$ 表示 $\boldsymbol{X}_p$ 与所有最近邻的 $h(x_{p+d},x_{H(p)+d})$ 值的平均值。为此，总体误差平均值为：$<E_d>=\frac{1}{N-d}\sum_{p=1}^{N-d}\varepsilon_p$。

对于周期序列 $<E_d>=0$；对一般不相关的随机序列，即符号 x_{p+d} 和向量 $\boldsymbol{X}_p$ 之间没有关系，其 $<E_d>$ 近似为 $\sum_\alpha p(\alpha)[1-p(\alpha)]$，这里的{$\alpha$}是 x_i 的字符集，$p(\alpha)$ 是符号 α 出现的概率。因此，对这样序列，其总误差平均值 $<E_d>$ 不依赖于所在嵌入维数 d，是一个常量。可以通过 $E_{\text{random}}=\sum_\alpha p(\alpha)[1-p(\alpha)]$ 计算出假定序列随机情况下的 E_{random} 值，再通过 $<E_d>$ 和 E_{random} 的比值来衡量序列的性质，即 $E=<E_d>/E_{\text{random}}$。若 E 的值为1，即 $<E_d>$ 和 E_{random} 相等，说明序列是随机的；若 E 的值不为1，则说明序列非随机，具有一定确定性。以嵌入维数 d 为横坐标，对应 E 值为纵坐标形成二维图形更直观显示序列特性，以下称为 *E-D* 图。

2 数据来源及处理

蛋白质序列，基于氨基酸的物理化学特性，有不同的分类方法。这里考虑用详细 HP 模型[10]将20种氨基酸分为4大类：无极性的，无负荷极性的，负极性的，正极性的，分别用 **A**,**P**,**NC**,**PC**[11]。即：

A={A,I,L,M,F,P,W,V}；

NC={D,E}；

P={N,C,Q,G,S,T,Y}；

PC={R,H,K}；

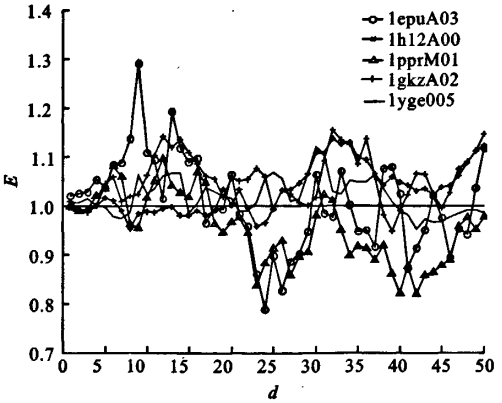
通过详细 HP 模型，蛋白质序列转化为4个集合组成的有限长的字符串。接下来构造蛋白质序列的特征序列[11]。首先，将4个集合无极性的 A，无负荷极性的 P，负极性的 NC，正极性的 PC，分为两组：{A,P}和{NC,PC}。按照这个分组，用1表示属于{A,P}里的氨基酸，而用0表示属于{NC,PC}里的氨基酸。这样就把蛋白质序列简化成一个(0,1)序列，作者称这个为这个蛋白质序列为蛋白质序列的第一特征序列。同样还可以将这4个集合 A,P,NC,PC，分为{A,PC}和{P,NC}，并利用上面的方法将蛋白质序列写成另外一个(0,1)序列，序列为第二特征序列。将这4个集合 A,P,NC,PC，分为{A,NC}和{P, PC}，并简化这个蛋白质序列为一个(0,1)序列时，称为第三特征序列，得到3个(0,1)序列，别称为蛋白质序列的第一、二、三特征序列。这里，每一个(0,1)序列实际上都是原来蛋白质序列的粗粒化描述，在这个过程中，原来的蛋白质序列中的有些信息被忽略了，有些信息却被强调了[11]。

对于蛋白质序列，有不同的结构分类方法。目前，最为常用的蛋白质结构分类数据库[1]为 SCOP (http://scop. berkeley. edu/)和 CATH (http://www. cathdb. info/latest/ index. html)。CATH 数据库由英国伦敦大学开发和维护，将使用计算机程序和人工检查结合起来，把 PDB 数据库中的蛋白质分为4类，即 α 主类，β 主类，αβ 类和低二级结构

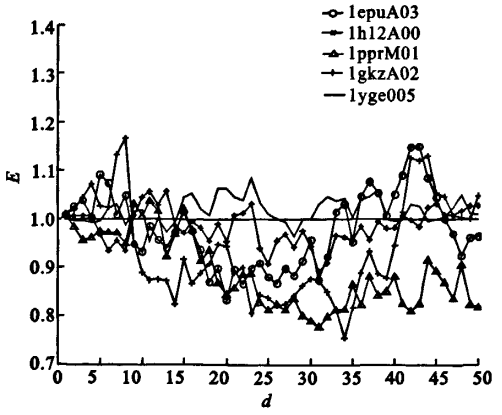
类[1]。α 主类指蛋白质二级结构主要是 α 螺旋型的,β 主类主要是指 β 折叠型,αβ 类则是指 α 螺旋和 β 折叠交替或者连续出现,低二级结构类是指二级结构成分含量很低的蛋白质分子。本文中数据都从 CATH 数据库 v3.0.0 中选取。选取的原则是按折叠类型以分层抽样的方法取出每类蛋白质序列 50～100 条以上,然后计算出每条序列当嵌入维数 $d\in[1,50]$ 时的 *E-D* 图,最后从每类最具代表性的序列中随机选取 5 条。

3　结果及讨论

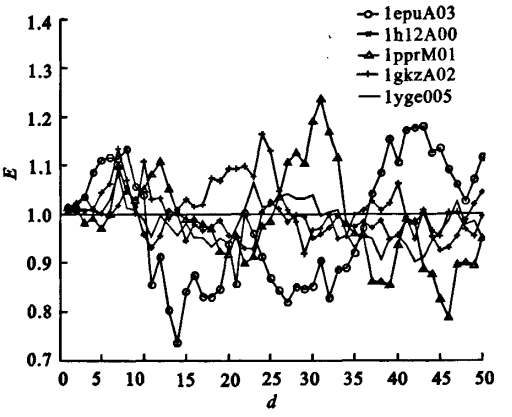
在图 1 中,给出了 α 主类的 5 条蛋白质序列的第一、二、三特征序列的 *E-D* 图。总体来说,随着嵌入维数 *d* 的变化,每条蛋白质序列的每个特征序列的 *E* 值都在 1 附近波动,就曲线而言具有特异性。而对于每条蛋白质序列,3 个特征序列的曲线波动幅度大体相似,如图中 1epuA03 及 1pprM01 两条蛋白质序列的第一二三特征序列分别对应的 3 条曲线波动起伏都很大,而 1h12A00 的和 1yge005 两条蛋白质序列分别对应的 3 条曲线走势相对较为平稳。



(a) Mainly Alpha 1



(b) Mainly Alpha 2



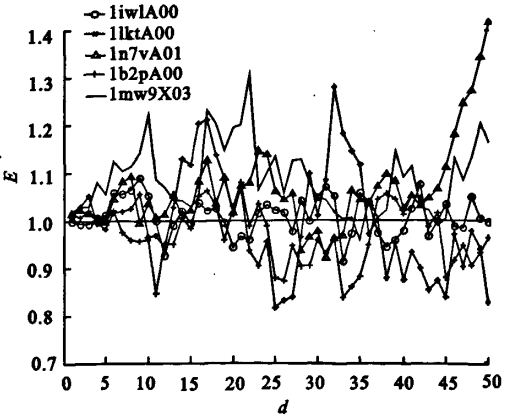
(c) Mainly Alpha 3

图 1　分别为从 α 主类中任意选取的 5 条蛋白质序列的第一、二、三特征序列关于不同嵌入维数 *d* 对应的 *E-D* 图

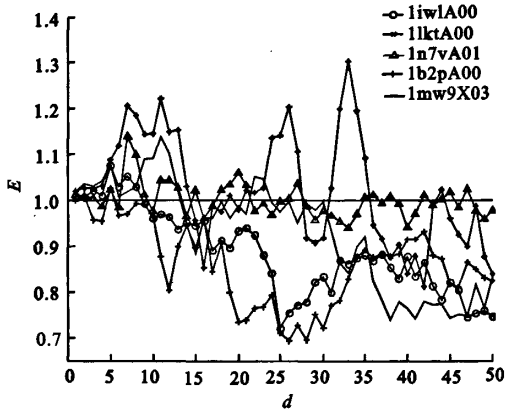
Fig. 1　The values of *E* versus the embedding dimension *d* calculated for three characteristic sequences of Mainly α protein sequences

在图 2 中,给出了 β 主类的 5 条蛋白质序列的第一、二、三特征序列的 *E-D* 图。同样每条曲线的 E 值都在 1 附近波动,但是明显不同于 α 主类,曲线表现出更强的特异性,且每条序列的 3 个特征序列的曲线也不具有相似性。如 1iw1A00,在第一特征序列下,曲线走势较为平稳,而在第二特征序列的曲线中走势起伏相当大。1n7vA01 却正好相反。即 β 主类可能表现出的某种特性比 α 主类来得更强。

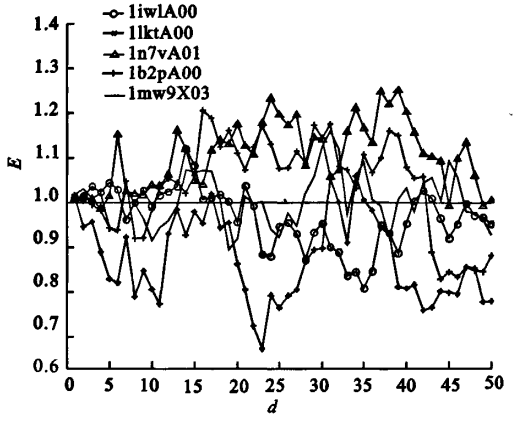
在图 3 中,给出了 αβ 类的 5 条蛋白质序列的第一、二、三特征序列的 *E-D* 图。曲线整体看上去虽然都在 1 附近波动,但是不同于 α 主类和 β 主类,αβ 类中多条曲线整体几乎在 1 以上或以下波动,如第三特征序列中的曲线几乎都在 1 上方。每条蛋白质序列的 3 个特征序列表现得特性差异也很大。



(a) Mainly Beta 1



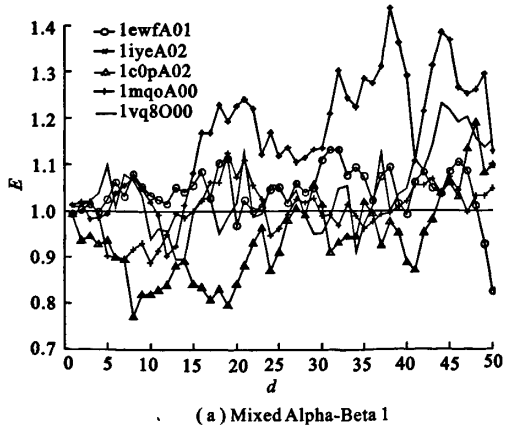
(b) Mainly Beta 2



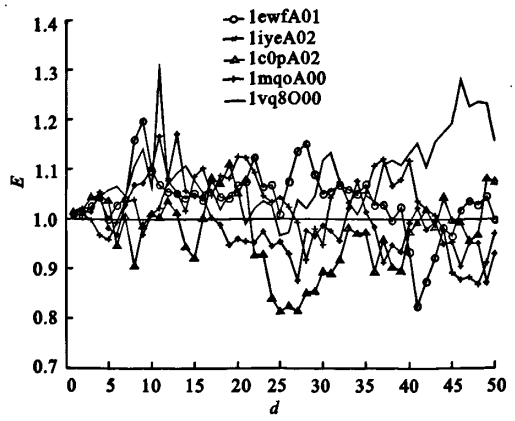
(c) Mainly Beta 3

图 2 分别为从 β 主类中任意选取的 5 条蛋白质序列的第一、二、三特征序列关于不同嵌入维数 d 对应的 E-D 图

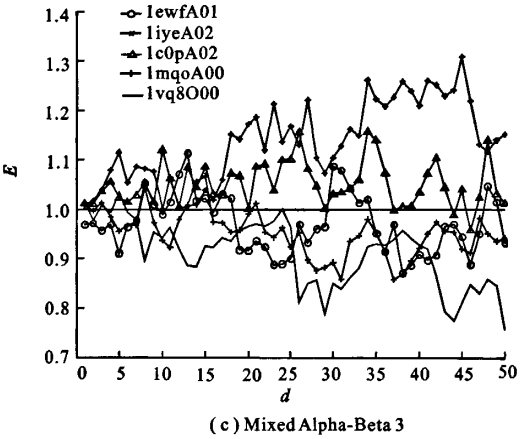
Fig. 2 The values of E versus the embedding dimension d calculated for three characteristic sequences of Mainly β protein sequences



(a) Mixed Alpha-Beta 1



(b) Mixed Alpha-Beta 2



(c) Mixed Alpha-Beta 3

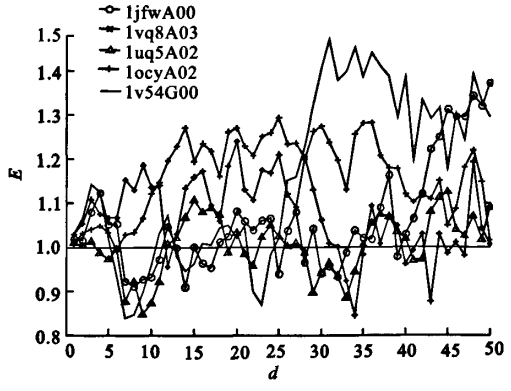
图 3 分别为从 α-β 主类中任意选取的 5 条蛋白质序列的第一、二、三特征序列关于不同嵌入维数 d 对应的 E-D 图

Fig. 3 The values of E versus the embedding dimension d calculated for three characteristic sequences of Mixed α-β protein sequences

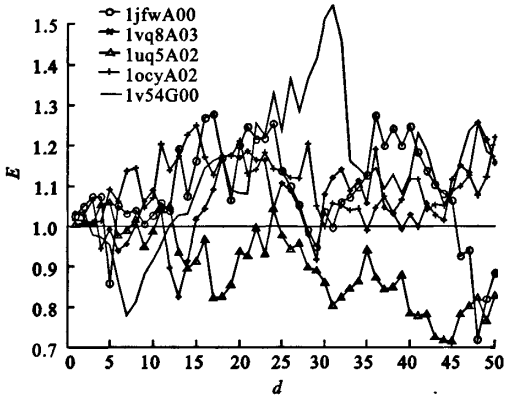
在图 4 中，给出了低二级结构类的 5 条蛋白质序列的第一、二、三特征序列的 E-D 图。和图 2，3，4 的 α 主类，β 主类及 α、β 类相比，低二级结构类的曲线波动最为剧烈，最大起伏达到 0.903 2 (1v54G00 对应的第三特征序列)。无论是每条蛋白质序列之间还是一条蛋白质的 3 个特征序列之间都具有很大的差异。

4 结 语

根据以上 4 组图及对应分析，得出每类蛋白质序列的每条特征序列的 E^2 值曲线在 1 附近起伏波动不断。其中 4 类蛋白质序列之间其对应特征序

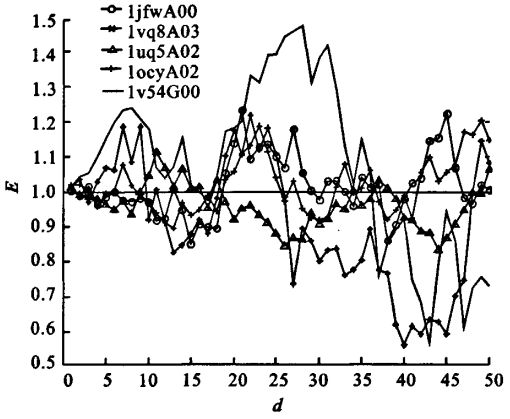


(a) Few Secondary Structures 1



(b) Few Secondary Structures 2

列具有一定得差异。每条蛋白质序列的 3 个特征序列之间也有很多不同。即蛋白质序列的 E 值曲线具有特异性，和前人得出的关于蛋白质序列具有随机性的结论不同。该结论为进一步研究蛋白质序列的特性提供了一定的依据。



(c) Few Secondary Structures 3

图 4　分别为从低二级结构类中任意选取的 5 条蛋白质序列的第一、二、三特征序列关于不同嵌入维数 d 对应的 E-D 图

Fig. 4　The values of E versus the embedding dimension d calculated for three characteristic sequences of Few Secondary Structure protein sequences

参考文献(References)：

[1] 赵国屏. 生物信息学[M]. 北京：科学出版社，2004. 149－165.

[2] 宋江宁，李炜疆. 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，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: Food Science and Biotechnology**, 2003, 22(3): 7－11. (in Chinese)

[3] 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.

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

[5] BARRAL J P, HASMY A. Nonlinear modeling technique for the analysis of DNA chains[J]. **Physical Reviewe**, 2000, 61 (2): 1812－1815.

[6] GARCIA P, JIMENEZ J. Local Optimal metrics and nonlinear modeling of chaotic time series[J]. **Physical Review Letters**, 1996, 76(9): 1449－1452.

[7] XIAO Y, HUANG Y Z, LI M F, et al. Nonlinear analysis of correlations in Alu repeat sequences in DNA[J]. **Physical Reviewe**, 2003, 68: (06)1913－1918.

[8] EUGENE S H. Nonlinear aspects of coding and noncoding DNA sequences: Annual March Meeting[C]. American Physical Society, 2001.

[9] YANG Hui-jie, ZHAO Fang-cui. Nonlinear modeling approach to human promoter sequences[J]. **Journal of Theoretical Biology**, 2006, 241: 765－773.

[10] YU Z G, ANH V O, LAU KA-SING. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses[J]. **Journal of Theoretical Biology**, 2004, 226: 341－348.

[11] 何平安. DNA 序列及蛋白质序列的分析与比较[D]. 大连：大连理工大学，2004.

(责任编辑：朱 明)