

文章编号: 2095-2163(2019)01-0019-05

中图分类号: Q811.4

文献标志码: A

基于多尺度熵的 DNA 序列相似性分析

张 静, 周小安, 赵 宇

(深圳大学 信息工程学院, 广东 深圳 518000)

摘 要: DNA 序列的相似性分析已成为当前生物信息学科中的研究热点, 对分析算法的需求也逐步增加, 基于样本熵的 DNA 序列相似性分析方法存在一定的效率问题。本文提出了一种基于多尺度熵的分析方法, 以 7 种病毒 DNA 序列作为实验研究的对象, 采用整数法将其分别表示为时间序列, 而后通过对比多个时间尺度下序列之间样本熵互值大小来显示序列之间的相关性, 并与原有的样本熵算法的分析结果进行比较。实验表明, 本文提出多尺度熵分析方法是切实可行的。

关键词: 相似性分析; DNA 序列; 样本熵; 多尺度熵

The similarity analysis of DNA sequences based on multiscale entropy

ZHANG Jing, ZHOU Xiao'an, ZHAO Yu

(College of Information Engineering, Shenzhen University, Shenzhen Guangdong 518000, China)

【Abstract】 The similarity analysis of DNA sequences has become a research hotspot in the bioinformatics discipline, and the demand for analysis algorithms has gradually increased. There are certain efficiency issues based on analyzing the similarity of DNA sequences with sample entropy. This paper studies the application of multiscale entropy for similarity analysis of DNA sequences. The DNA sequences of seven viruses are used as experimental objects, which are converted into digital sequences by the numerical representation of DNA sequences. Then, by comparing the mutual values of sample entropy between DNA sequences at multiple time scales, the similarity of DNA sequences is analyzed. And compared with the results of sample entropy method, the experiments are designed. Experiments results strengthens the conclusion that it is feasible to analyze the DNA sequences similarity by multiscale entropy.

【Key words】 similarity analysis; DNA sequence; sample entropy; multiscale entropy

0 引 言

人类基因组计划完成, 基因测序技术得到飞速的发展, 生物序列也已吸引到多方关注与瞩目, 如何分析和解读生物序列中所包含的有用信息已然成为目前生物学研究的关键。对于获得的一个新物种, 如果能证实其与某些已知序列存在的一定的联系, 那么就能分析出该新物种的结构和功能与相似序列之间存在的共同性, 如此将会大大减轻基因检测与新序列测定的工程量。面对数量级巨大的生物序列, DNA 序列的相似性分析即已显得尤为重要^[1-4]。

分析 DNA 序列相似性的传统方法, 不仅计算量大, 且存在一定的缺陷。诸如, 点阵图分析方法, 实验数据是不能插入空格的^[5]; 而傅里叶变换方法则会丢失序列中的部分信息, 也不能清楚地显示序列之间的异同^[6]; 但将信息理论方法运用在相似性分析上, 取得的实验结果也并不理想^[7-10]; 此外的统计特征方法对数据量却有着极高的要求^[11]。因此

找寻一种不仅能真实反映出序列的复杂度情况, 还能高度准确地从相关性角度全面分析生物序列的相似性的计算方法就已成为本文设计讨论的重点。

近年来, 作为一种数字化表示时间序列复杂度和研究时间序列基本性质的重要计算方法, 样本熵 (sample entropy, SampEn) 在 DNA 序列的相似性分析方法中得到了充分的认可与使用^[12]。在分析比较 DNA 序列相似性的研究中, 样本熵能有效解读序列变化中很小的复杂性变化, 精确度也较高, 但也依然存在一定的局限性。首先, DNA 分子是一种高分子聚合物, 在本次实验研究中, 仅是选取其中的一个片段, 若运用样本熵的测度方法, 分析整个 DNA 序列, 其巨大的工程量, 也将凸显出此方法的短板与不足。其次, 样本熵的测度方法具有时间尺度单一的特点, 而时间尺度对于衡量序列的复杂度是有影响的。为了能对 DNA 序列的复杂度进行全面量化, 本文拟将引入一种基于多尺度熵的测度算法。相对于样本熵算法而言, 多尺度熵能够更加高效、清晰地展示出 DNA 序列的相似程度。对此内容可展开研究论述如下。

作者简介: 张 静(1994-), 女, 硕士研究生, 主要研究方向: 非线性序列数据分析; 周小安(1968-), 男, 博士, 副教授, 主要研究方向: 混沌系统、保密通信、非线性系统理论; 赵 宇(1991-), 男, 硕士研究生, 主要研究方向: 图像处理、非线性序列数据分析。

收稿日期: 2018-09-30

1 分析方法

1.1 基于样本熵的序列相似性分析算法原理

基于样本熵的 DNA 序列相似性分析方法主要计算和的对数,过程中避免分析自身无意义的值。其物理意义是衡量序列的复杂程度,是一种有关时间复杂度的分析方法。序列的样本熵值越大,其复杂程度越高,序列“同源”的可能性就越小,反之,序列的样本熵值越低,序列相似的可能性就越大^[13-16]。其计算过程可表述为:

(1)将 N 点时间序列 $\{u(i) : 1 \leq i \leq N\}$ 按顺序组成 m 维矢量,对此可表示如下:

$$\mathbf{X}_m(i) = [u(i), u(i+1), \dots, u(i+m-1)] \quad i = 1 \sim N-m \quad (1)$$

(2)对每一个 i 值计算矢量 $\mathbf{X}_m(i)$ 与其余矢量 $\mathbf{X}_m(j)$ 之间的距离,定义 $d[\mathbf{X}_m(i), \mathbf{X}_m(j)]$ 为矢量 $\mathbf{X}_m(i)$ 和 $\mathbf{X}_m(j)$ 中对应元素差值的最大值,即:

$$d[\mathbf{X}_m(i), \mathbf{X}_m(j)] = \max |x(i+k) - u(j+k)| \quad (2)$$

其中, $i, j = 1 \sim N-m+1; i \neq j$ 。

(3)设定匹配过程中正数阈值 $r (r > 0)$,对每一个 i 值统计 $d[\mathbf{X}_m(i), \mathbf{X}_m(j)] < r$ 的个数与总的矢量个数 $(N-m+1)$ 的比值,记为 $B_i^m(r)$,即:

$$B_i^m(r) = \{d[\mathbf{X}_m(i), \mathbf{X}_m(j)] < r \text{ 的数目}\} / (N-m+1) \quad (3)$$

然后对所有 i 求平均值,记作 $B^m(r)$,即:

$$B^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} B_i^m(r) \quad (4)$$

(4)将维数加 1 变为 $m+1$,重复步骤(1)~(3),得到 $B^{m+1}(r)$,即:

$$B^{m+1}(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} B_i^{m+1}(r) \quad (5)$$

(5)运算得出理论上的样本熵值为:

$$\text{SampEn}(m, r) = \lim_{N \rightarrow \infty} \{-\ln [B^{m+1}(r) / B^m(r)]\} \quad (6)$$

此时,令 N 为无限大,依上述步骤计算得出长度为 N 时,序列的样本熵值约为:

$$\text{SampEn}(m, r, N) = -\ln [B^{m+1}(r) / B^m(r)] \quad (7)$$

分析可知, SampEn 的值与 m, r, N 的值相关,所以本文中取 $m = 2, r = 0.1 \sim 0.25SD(x)$ 。

1.2 基于样本熵的序列相似性分析算法原理

在 2002 年, Costa 等人^[17]首次提出一种基于样本熵的、用来分析有关时间序列复杂程度的重要度量方法——多尺度熵 (MultiScale Entropy, MSE)。该方法很好地解决了样本熵算法单尺度的缺陷,多

尺度熵是在样本熵计算方法基础上的改进,研究推得其计算过程如下。

(1) 设 N 点离散时间序列为 $X = \{x(i) : 1 \leq i \leq N\}$, 在给定的嵌入维数 (即窗口长度) m 和相似容限 r 之下, 构建新的粗粒化向量。可得其数学计算公式如下:

$$y_\tau(j) = \frac{1}{\tau} + \sum_{i=j-1}^{j+\tau-1} x(i) \quad 1 \leq j \leq \lfloor N/\tau \rfloor \quad (8)$$

其中, $\lfloor N/\tau \rfloor$ 表示向下取整, $\tau = 1, 2, 3, \dots$ 是正整数, 称为尺度因子。

粗粒化过程也就是用长度为 τ 的窗口进行滑动平均的过程。将其描绘出来, 则如图 1 所示。

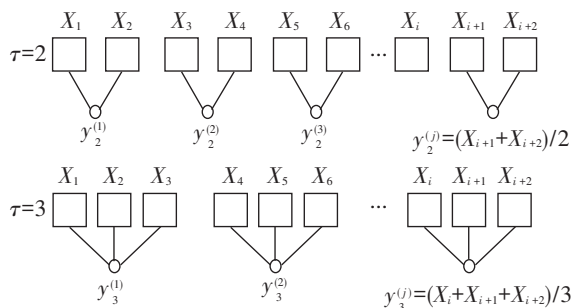


图 1 时间序列的粗粒化处理过程

Fig. 1 Coarse granulation process of time series

(2) 计算在不同尺度因子粗粒化后的序列的样本熵值, 每个值对应一个尺度因子, 也就是计算序列对应的尺度熵值。设最大的尺度因子为 τ_{\max} , 分别计算尺度 $[1, 2, \dots, \tau_{\max}]$ 对应的粗粒化向量 $[y_1(j), y_2(j), \dots, y_{\tau_{\max}}(j)]$ 对应的样本熵值, 从而得到原始序列的多尺度熵值 $MSE = [SE_1, SE_2, \dots, SE_{\tau_{\max}}]$ 。

2 DNA 序列的整数表示方法

DNA 序列是由 4 种不同的碱基组合而成的, 在文献^[18]中, 提供了一种基于数值的表示方法。其本质就是将 4 种碱基字符串依照适当的映射规则一一对应映射为数值序列, 这样编码得到的 DNA 序列就可以对其进行后续的数据处理和数据分析。研究求出其映射关系表示方法如下:

$$D_n = \begin{cases} 0 & X_n = A \\ 1 & X_n = G \\ 2 & X_n = C \\ 3 & X_n = T \end{cases} \quad (9)$$

3 实验数据

本文从 NCBI 数据库 (<http://www.ncbi.nlm.nih>).

表3 $\tau=2$ 时,7种病毒的DNA序列之间的互样本熵矩阵Tab. 3 Multiscale entropy between DNA sequences of seven viruses when $\tau=2$ 10^{-2}

VIRUS	H5N1(1)	H5N1(2)	H1N1	H2N2	H3N2	H7N9	SARS
H5N1(1)	0.573 8	0.645 4	2.574 5	1.943 1	2.162 4	1.811 2	1.947 1
H5N1(2)		0.617 2	2.602 7	2.054 1	2.274 9	1.791 8	1.945 9
H1N1			0.844 2	1.909 5	2.107 4	1.713 8	1.883 0
H2N2				0.806 1	2.639 1	2.021 0	1.921 8
H3N2					0.751 0	1.543 3	1.414 9
H7N9						0.693 1	1.386 3
SARS							0.802 3

当尺度因子 $\tau = 3$ 时,计算 7 种 DNA 序列之间的多尺度熵值,即 $\tau = 3$ 时的样本熵值。综合运算结果见表 4。

$\tau = 1$ 时,实际上计算的就是 DNA 序列之间的样本熵值,从表 2~表 4 中,可知 $\tau = 3$ 与 $\tau = 2$ 时的实验结果与 $\tau = 1$ 时的实验结果相同,都表现出 H5N1(1) 与 H5N1(2) 有着很高的相似性,这也验证、并说明了基于多尺度熵的相似性分析方法是切

实可行的。 $\tau = 2$ 时实验展示的 DNA 序列之间相似性程度比 $\tau = 1$ 时实验效果要更加明显,且在 $\tau = 3$ 中的数据差异则尤其明显,这也进一步展现出其在研究 H5N1(1) 与 H5N1(2) 具有很高的相似性、而与其它序列相似性较低方面的优越性能。综上分析可知,多尺度熵分析算法可以运用在 DNA 序列相似性分析研究上,不仅能降低计算量、提高实验分析的效率,还能更加突出显示序列整体之间的相似性程度。

表4 $\tau=3$ 时,7种病毒的DNA序列之间的互样本熵矩阵Tab. 4 Multiscale entropy between DNA sequences of seven viruses when $\tau=3$ 10^{-2}

VIRUS	H5N1(1)	H5N1(2)	H1N1	H2N2	H3N2	H7N9	SARS
H5N1(1)	0.387 8	0.446 3	2.639 1	2.351 4	3.135 5	1.722 8	1.945 9
H5N1(2)		0.313 7	2.351 4	2.154 7	2.233 6	1.845 8	1.609 4
H1N1			0.405 5	2.345 1	2.483 7	2.028 1	2.484 9
H2N2				0.479 6	1.897 1	2.944 4	1.504 1
H3N2					0.525 0	1.223 8	1.981 0
H7N9						0.578 1	1.386 3
SARS							0.470 0

5 结束语

本文在样本熵的基础上,运用多尺度熵的分析方法分析 7 种病毒序列的相似性,实验结果表明,该方法能够有效表现序列之间的相似性程度。相对于样本熵分析算法,多尺度熵分析算法效率更高,但是多尺度熵算法中数据分析的精确度却降低了,所以本文的方法适用于分析研究序列长度较大的数据对象。在具体实验中,要根据整体详尽的实验需求,有针对性地选择最合适的 DNA 序列的分析算法。

参考文献

- [1] PAL S K, BANDYOPADHYAY S, RAY S S. Evolutionary computation in bioinformatics: A review[J]. IEEE Transactions on Systems, Man, & Cybernetics, Part C, 2006, 36(5):601-615.
- [2] 鲁卫平,周元国. 生物信息学的现状和展望[J]. 国际检验医学杂志, 2002, 23(5):254-255,274.

- [3] 张春霆. 生物信息学的现状与展望[J]. 世界科技研究与发展, 2000, 22(6):17-20.
- [4] 唐玉荣. 生物信息学中的序列比对算法[J]. 计算机工程与应用, 2003, 39(29):5-7.
- [5] GIBBS A J, MCINTYRE G A. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequence [J]. European Journal of Biochemistry, 1970, 16(1):1-11.
- [6] 张少宏,戴宪华. 基于对齐的生物序列相似性分析[J]. 生物信息学, 2005, 3(2):81-84.
- [7] GATLIN L L. Information theory and the living system[M]. New York:Columbia University Press, 1972.
- [8] HARIRI A, WEBER B, OLMSTED J. On the validity of Shannon-information calculations for molecular biological sequences [J]. Journal of Theoretical Biology, 1990, 147(2):235-254.
- [9] 刘芳. 基于信息离散度的 DNA 序列相似性分析研究[D]. 长沙:湖南大学, 2009.
- [10] FANG Weiwu, ROBERTS F S, MA Zhengrong. A measure of discrepancy of multiple sequences [J]. Information Sciences, 2001, 137(1-4):75-102.
- [11] LIAO Bo, WANG Tianming. New 2D graphical representation of

- DNA sequences[J]. *Journal of Computational Chemistry*, 2004, 259(11):1364-1368.
- [12] ZHANG Xun, ZHOU Xiaolan, Yu Yunhui. Similarity analysis of DNA using improved approximate entropy[C]//2012 International Conference on Biomedical Engineering and Biotechnology (iCBEB). Macau, Macao:IEEE, 2012:511-514.
- [13] LAKE D E, RICHMAN J S, GRIFFIN M P, et al. Sample entropy analysis of neonatal heart rate variability [J]. *Am J. Physiol. Regul. Integr. Comp. Physiol.*, 2002, 283(3):789-797.
- [14] PINCUS S M. Approximate entropy as a measure of system complexity[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 1991, 88(6):2297-2301.
- [15] ALCARAZ R, RIETA J J. A review on sample entropy applications for the non-invasive analysis of atrial fibrillation electrocardiograms[J]. *Biomedical Signal Processing & Control*, 2010, 5(1):1-14.
- [16] LIU Lizhi, QIAN Xiyuan, LU Hengyao. Cross-sample entropy of foreign exchange time series[J]. *Physica A: Statistical Mechanics and its Applications*, 2010, 389(21):4785-4792.
- [17] COSTA M, GOLDBERGER A I, PENG C K. Multiscale entropy analysis of physiologic time series[J]. *Physical Review Letters*, 2002, 89(6):068102-1-4.
- [18] ROSEN G L. Examining coding structure and redundancy in DNA [J]. *IEEE Engineering in Medicine & Biology Magazine*, 2006, 25(1):62-68.

(上接第 18 页)

- [2] 曹立杰, 郭戈, 靳玉峰, 等. 基于传感器网络的海洋水质监测及赤潮预报系统的设计[J]. *大连海洋大学学报*, 2014, 29(6):664-668.
- [3] 田野, 郭子祺, 乔彦超, 等. 基于遥感的官厅水库水质监测研究[J]. *生态学报*, 2015, 35(7):2217-2226.
- [4] 李家良. 水面无人艇发展与应用[J]. *火力与指挥控制*, 2012, 37(6):203-207.
- [5] 张树凯, 刘正江, 张显库, 等. 无人船艇的发展及展望[J]. *世界海运*, 2015, 38(9):29-36.
- [6] NAEEM W, IRWIN G W, YANG A. COLREGs-based collision avoidance strategies for unmanned surface vehicles [J]. *Mechatronics*, 2012, 22(6):669-678.
- [7] CAMPBELL S, NAEEM W, IRWIN G W. A review on improving the autonomy of unmanned surface vehicles through intelligent collision avoidance manoeuvres [J]. *Annual Reviews in Control*, 2012, 36(2):267-283.
- [8] PRACZYK T. Neural anti-collision system for autonomous surface vehicle [J]. *Neurocomputing*, 2015, 149:559-572.
- [9] 刘钰. 基于导航误差约束的水面无人艇路径规划方法研究[D]. 南京:东南大学, 2017.
- [10] 郑佳春, 吴建华, 马勇, 等. 混合模拟退火与粒子群优化算法的水面无人艇路径规划[J]. *中国海洋大学学报(自然科学版)*, 2016, 46(9):116-122.
- [11] 马文耀, 吴兆麟, 杨家轩, 等. 人工鱼群算法的避碰路径规划决策支持[J]. *中国航海*, 2014, 37(3):63-67.
- [12] 范云生, 赵永生, 石林龙, 等. 基于电子海图栅格化的无人水面艇全局路径规划[J]. *中国航海*, 2017, 40(1):47-52.
- [13] 霍凤财, 任伟建, 刘东辉. 基于改进的人工势场法的路径规划方法研究[J]. *自动化技术与应用*, 2016, 35(3):63-67.
- [14] 邢海洋, 张军, 王楠. 自动导引车云导引平台的研究与设计[J]. *计算机工程*, 2017, 43(7):64-69.
- [15] 梁献霞, 刘朝英, 宋雪玲, 等. 改进人工势场法的移动机器人路径规划研究[J]. *计算机仿真*, 2018, 35(4):291-294, 361.
- [16] 郭娜. 基于模拟退火-Q学习的移动机器人路径规划技术研究[D]. 南京:南京理工大学, 2009.
- [17] 陈自立, 徐娅萍, 顾立彬. 基于模糊 Q 学习算法的 AGV 路径规划研究[J]. *制造业自动化*, 2012, 34(11):4-6, 16.
- [18] 董培方, 张志安, 梅新虎, 等. 引入势场及陷阱搜索的强化学习路径规划算法[J]. *计算机工程与应用*, 2018, 54(16):129-134.
- [19] 于乃功, 王琛, 默凡凡, 等. 基于 Q 学习算法和遗传算法的动态环境路径规划[J]. *北京工业大学学报*, 2017, 43(7):1009-1016.