文章编号：2095-2163(2019)06-0052-03 中图分类号：TP391 文献标志码：A

基于互模式熵的 DNA 序列相似性分析

安相静，周小安，张 静，沈冲冲
（深圳大学 信息工程学院，广东 深圳 518000）

摘 要： 随着基因组计划的开展，DNA 序列相似性分析成了现代生物学研究不可缺的一部分。本研究以 H5N1、H1N1 和 H2N2 等 7 种病毒的 DNA 序列作为研究对象，使用整数法将 DNA 序列编码成时间序列信息，计算时间序列之间的互模式熵（Mutual Mode Entropy，MME）。分析不同 DNA 序列的 MME 对序列相似性的表达准确性。实验表明通过整数表示方法的 DNA 序列的 MME 能够定性地解释 7 种 DNA 序列之间的相似性关系。

关键词： 相似性分析；DNA 序列；DNA 表示方法；互模式熵

DNA sequence similarity analysis based on mutual pattern entropy

AN Xiangjing，ZHOU Xiaoan，ZHANG Jing，SHEN Chongchong
（College of Information Engineering，Shenzhen University，Shenzhen Guangdong 518000，China）

【Abstract】 With the development of the genome project，DNA sequence similarity analysis has become an indispensable part of modern biological research. In this study，DNA sequences of seven viruses，such as H5N1，H1N1 and H2N2，were used as research objects，and DNA sequences were encoded into time series information using an integer method，Calculate Mutual Mode Entropy（MME）between time series. The accuracy of expression of sequence similarity by MME of different DNA sequences was analyzed. Experiments have shown that the MME of the DNA sequence by the integer representation method can qualitatively explain the similarity relationship between the seven DNA sequences.

【Key words】 similarity analysis；DNA sequence；DNA representation；mutual mode entropy

0 引 言

在人类基因组计划（Human Genome Project，HGP）实施之前，生物技术比较陈旧，无法获得长的、连续的基因序列，对 DNA 分子序列的研究只局限在分析相邻碱基对之间的相关性，以及 DNA 片段中碱基密度的不均匀性讨论[1]。随着 HGP 顺利实施，生物信息学应运而生，围绕 DNA 序列和蛋白质序列开展了一系列研究分析。DNA 序列的相似性研究就是其中的一个分支。

迄今为止，DNA 序列相似性分析方法层出不穷，Bemdt 与 Clifford 提出的动态时间弯曲（Dynamic Time Warping，DTW）是把时间序列规划和距离测度相结合的非线性规划技术，用于计算两个时间序列的最大相似性[2-3]，李梅等将 DTW 应用到 DNA 序列相似性研究当中，取得了较好的效果[4]。但 DTW 距离在本研究中计算时间复杂度较高。样本熵的方法精度较高，能够分析出更为微小复杂的序列之间的变化，但评估的时间尺度比较单一[5]。近似熵算法虽有一定的抗噪性，但因其在度量序列复杂度时引入了自身数据的比较，会造成统计数据不稳定[6]。本研究采用互模式熵（MME）实现对 DNA 序列相似性估计，MME 算法修改了判定矢量相似的准则，不再考虑被比较的 2 个矢量纵坐标位置是否相同，而通过 2 个矢量对应的波形片段作为 2 个矢量相似的判定依据。此判定准则能有效减少判断矢量相似过程中对容限阈值 r 的依赖，不会因为信号大幅度波动或者信号长度忽然改变影响相似矢量的个数，有效解决了近似熵存在的统计稳定问题。

1 基于 MME 的序列相似性分析算法原理

模式熵（Mode Entropy，ModEn）算法的概念是宁新宝等人在 2005 年首次提出的[7]，有效解决了度量短时大幅度缓慢变化的信号其复杂度的问题。互模式熵（Mutual Mode Entropy，MME）是 ModEn 算法的延伸[8-9]。用来度量不同序列之间是否存在高度

基金项目： 中央财政支持地方高校专项资金（8060000260205）。

作者简介： 安相静（1995-），男，硕士研究生，主要研究方向：机器学习、图像处理；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；张 静（1994-），女，硕士研究生，主要研究方向：非线性序列数据分析；沈冲冲（1995-），男，硕士研究生，主要研究方向：计算机视觉、机器学习、深度学习算法。

收稿日期： 2019-09-12

耦合的问题。计算步骤如下：

首先，对于 2 组包含 N 个数据的时间序列：$\{u(i):0 \leqslant i \leqslant N-1\}$，$\{v(j):0 \leqslant j \leqslant N-1\}$ 从每组数据中连续取 m 个数据点，分别组成其对应的 m 维矢量：

$$X(i) = [u(i), u(i+1), \ldots, u(i+m-1)];$$
$$Y(j) = [v(j), v(j+1), \ldots, v(j+m-1)], \quad (1)$$

基准线分别是每一个 m 维矢量的平均值，其计算公式为：

$$B_u(i) = \frac{\sum_{i=0}^{m-1} u(i+l)}{m}, \quad B(j) = \frac{\sum_{i=0}^{m-1} v(j+l)}{m}. \quad (2)$$

根据 $B_u(i)$ 和 $B_v(i)$ 的值，重新构造向量：

$$\Psi_u(i) = [u(i) - B_u(i), u(i+1) - B_u(i), \ldots, u(i+m-1) - B_u(i)] = [\varphi_u(i), \varphi_u(i+1), \ldots, \varphi_u(i+m-1)], \quad (3)$$

$\Psi_v(i)$ 同上，L_{ij} 是矢量 $\Psi_u(i)$ 和 $\Psi_v(i)$ 之间对应元素中相差最大的值：

$$L_{ij} = L[\Psi_u(i), \Psi_v(j)] = \max_{k=0\to m-1} [|\varphi_u(i+k) - \varphi_v(j+k)|], \quad (4)$$

之后定义这两个矢量相似的概率：

$$C_i^m(r) = \frac{1}{N-m+1} \sum_{j=0}^{N-m} \theta(r - L_{ij}), \quad (5)$$

对 $C_i^m(r)$ 取对数，得到编码长度 m 时的概率。计算 $m+1$ 时的概率，如式(7)得到 MME：

$$\phi^m = \frac{1}{N-m+1} \sum_{i=0}^{N-m} \ln C_i^m(r), \quad (6)$$

$$MME(m, r, N) = \varphi^m - \varphi^{m+1}, m \geqslant 1. \quad (7)$$

MME 算法是不同序列之间的相似程度的量化。不仅可以用于计算不同 DNA 序列之间的差异，还可以用于了解同一个序列不同区间段之间的差异，对于 DNA 序列相似性的分析研究具有十分重要的意义[10-11]。

2　实验结果及分析

2.1　实验数据

本文实验中采用 7 种 DNA 片段序列数据，是由 NCBI 数据库中下载（详细信息见：http://www.ncbi.nlm.nih.gov），接下来运用 ModEn 算法及 MME 法来分析研究这些 DNA 片段序列。

2.2　DNA 序列的整数表示方法

由于 7 种病毒 DNA 序列的片段信息都是字符串形式，不利于实验分析研究，因此需要将其转化为时间序列。采用整数表示方法，其映射关系为：

$$D_n = \begin{cases} 0 & X_n = A \\ 1 & X_n = G \\ 2 & X_n = C \\ 3 & X_n = T. \end{cases} \quad (8)$$

2.3　实验结果

从公式(7)可知，DNA 序列之间的 MME 值是由编码长度 m、容限阈值 r、序列长度 N，3 个参数共同决定。依次计算为 m= 1, 2,…, 7 时，H5N1(1)与其它 6 种病毒之间的 MME 值。令 $R = |0.2 * cov(u, v)|$，$N = 900$。实验结果见表 1。

表 1　整数表示方法下不同 m 值时 H5N1(1)与其它 6 种病毒 DNA 之间的 MME

Tab. 1　Integer representation of the MME between H5N1(1) and the DNA of six other viruses at different m

编码长度 m	$m=1$	$m=2$	$m=3$	$m=4$	$M=5$	$m=6$	$m=7$
H5N1(2)	0.007 4	**0.003 0**	0.029 5	0.038 5	0.031 7	0.017 2	0.013 6
H1N1	0.023 4	0.149 0	0.486 0	1.000 6	1.233 0	1.136 4	1.271 3
H2N2	0.015 1	0.142 3	0.452 9	0.810 0	1.094 0	1.068 9	1.195 7
H3N2	0.057 1	0.175 1	0.478 4	0.822 5	1.048 7	1.134 6	0.734 1
H7N9	0.031 4	0.144 1	0.465 9	0.822 6	0.884 4	1.082 2	1.175 9
SARS	0.064 9	0.157 2	0.419 2	0.788 1	0.928 5	1.154 8	1.070 6

由表 1 可知，$m=2$ 时，MME 最小；$m=1,3$ 时，H5N1(1)与 H5N1(2)之间的 MME 值和 H5N1(2)与其它 5 个序列之间的 MME 值相差 1 个数量级，$m=4, 5, 7$ 时，MME 相差 1~2 个数量级，$m=2, 6$ 时，MME 相差 2 个数量级，差异最大。综上所述，当编码长度 $m=2$ 时最能保证实验性能和准确性。

令编码长度 $m=2$，$r = |0.2 * \mathrm{cov}(u, v)|$ 分别对 7 种病毒的 DNA 序列进行 MATLAB 仿真，得到 7 种病毒的 DNA 序列之间的 MME 值见表 2。

由表 2 实验结果可知：H5N1(1)与 H5N1(2)之间的 MME 最小，说明其相似程度最高。H5N1(1)与 SARS 的 MME 值在表 2 的第一行中是最大的，也就是说 H5N1(1)与 SARS 之间的相似程度最小。

实验结果及分析证明 MME 算法在 DNA 序列

之间的相似性研究中,能有效判断出不同 DNA 序列 之间的相似程度。

表 2 基于整数值的 DNA 表示方法下 7 种病毒 DNA 序列之间的 MME 值

Tab. 2 MME values between seven viral DNA sequences based on integer value-based DNA representation m

VIRUS	H5N1(1)	H5N1(2)	H1N1	H2N2	H3N2	H7N9	SARS
H5N1(1)	0	**0.003 0**	0.149 0	0.142 3	0.154 1	0.144 1	0.157 2
H5N1(2)		0	0.155 7	0.151 7	0.180 6	0.150 8	0.176 0
H1N1			0	0.104 5	0.178 0	0.137 6	0.201 5
H2N2				0	0.197 9	0.143 7	0.255 5
H3N2					0	0.103 1	0.162 2
H7N9						0	0.205 0
SARS							0

3 结束语

本研究通过 MME 算法对 7 种病毒做 DNA 序列相似性研究,实验结果的定性分析更加稳定、有效。本研究是 MME 算法应用的冰山一角还有更多非线性研究领域可以对其展开更深入的研究,对 DNA 序列相似性进行定量分析是未来的研究方向。

参考文献

[1] WANG X, XIA Z, CHEN C, et al. The international Human Genome Project (HGP) and China′s contribution[J]. 蛋白质与细胞, 2018, 9(4):317-321.

[2] 张艳萍. 生物序列相似性向量及其识别 DNA 结合蛋白的效果研究[D]. 天津:南开大学, 2014:1-4.

[3] KEOGH E, CHOTIRAT ANN Ratanamahatana. Exact indexing of dynamic time warping[J]. Knowledge & Information Systems, 2005, 7(3):358-386.

[4] 李梅, 白凤兰. 基于 DTW 距离的 DNA 序列相似性分析[J]. 生物数学学报, 2009(2):374-378.

[5] 李立, 曹锐, 相洁. 脑电数据近似熵与样本熵特征对比研究[J]. 计算机工程与设计, 2014, 35(3):1021-1026.

[6] Ning XB, Xu YL, Wang J, et al. Approximate entropy analysis of short-term HFECG based on wave mode [J]. Physica A, 2005, 346(3): 475-483.

[7] 张洁, 景晓军, 刘馨靖,等. 一种基于模式熵的残缺指纹识别算法[J]. 电子与信息学报, 2012, 34(12):3040-3045.

[8] ZHANG J. An Incomplete Fingerprint Recognition Algorithm Based on Pattern Entropy[J]. Journal of Electronics & Information Technology, 2012, 34(12):3040-3045.

[9] 蔺博宇, 郭志刚, 李弼程,等. 基于随机化映射和模式熵的近似重复图像检测[J]. 数据采集与处理, 2012, 27(3):363-367.

[10] 徐寅林, 宁新宝, 陈颖. 模式熵与高频心电图信号不规则性的动态分析[J]. 科学通报, 2004, 49(13):1317-1321.

[11] 吴俊, 王俊. 基于模式熵的室性早搏与房性早搏识别[J]. 生物医学工程学杂志, 2010(3):516-518.

(上接第 51 页)

[13] GEHLER P V, ROTHER C, BLAKE A, et al. Bayesian color constancy revisited[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.

[14] CARDEI V C, FUNT B, BARNARD K. Estimating the scene illumination chromaticity by using a neural network[J]. Journal of the Optical Society of America A, 2002, 19(12): 2374-2386.

[15] LOU Z, GEVERS T, HU N, et al. Color Constancy by Deep Learning[C]// British Machine Vision Conference. 2015: 76.1-76.12.

[16] OH S W, KIM S J. Approaching the computational color constancy as a classification problem through deep learning[J]. Pattern Recognition, 2017, 61: 405-416.

[17] BIANCO S, CUSANO C, SCHETTINI R. Color constancy using CNNs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 81-89.

[18] BIANCO S, CUSANO C, SCHETTINI R. Single and multiple illuminant estimation using convolutional neural networks[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4347-4362.

[19] SHI W, LOY C C, TANG X. Deep specialized network for illuminant estimation[C]//European Conference on Computer Vision. Springer, Cham, 2016: 371-387.

[20] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[21] SHI L. Re-processed version of the gehler color constancy dataset of 568 images ([DB/OL]). 2010, http://www.cs.sfu.ca/~color/data/.

[22] CHENG D, PRASAD D K, BROWN M S. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution[J]. Journal of the Optical Society of America A, 2014, 31(5): 1049-1058.

[23] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional architecture for fast feature embedding[C]// Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[24] CHAKRABARTI A, HIRAKAWA K, ZICKLER T. Color constancy with spatio-spectral statistics[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(8): 1509-1519.

[25] BARRON J T. Convolutional color constancy[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 379-387.