Software Engineering and Applications 软件工程与应用, 2014, 3, 41-49
Published Online June 2014 in Hans. http://www.hanspub.org/journal/sea
http://dx.doi.org/10.12677/sea.2014.33006

Hans汉斯

DNA Clustering Distribution Measured with Probability under Nonlinear Function

Lei Du[1], Jeffrey Zheng[2]

[1]Department of Information Security, School of Software, Yunnan University, Kunming.
[2]Key Lab of Yunnan Software Engineering, Kunming
Email: handsome9501@qq.com

Received: Apr. 4th, 2014; revised: May 7th, 2014; accepted: May 15th, 2014

Abstract

Typical clustering analysis can make similarity data together and show the use of the same or different spatial distribution of fragments presented in the sequence. This paper deals with DAN sequences from different sources using statistical calculations and projection characteristics grouped in three different nonlinear functions of the probability value measurements, getting a visual on the genetic characteristics of the formation of the cluster distribution. Comparison showed that similar stratification had the same trend and complementary characteristics at a higher level, but there are obvious differences between the distributions of different types of gene sequences.

Keywords

Cluster of DNA, Probability Value, Nonlinear Function, Genome Sequence

在非线性函数下的DNA概率测量聚类分布

杜　磊[1]，郑智捷[2]

[1]云南大学软件学院信息安全系，昆明
[2]云南省软件工程重点实验室，昆明
Email: handsome9501@qq.com

收稿日期：2014年4月4日；修回日期：2014年5月7日；录用日期：2014年5月15日

摘　要

典型的聚类分析方法可将相似度较高的数据片段依据测量的数值特征聚集在一起，利用空间分布展示序列中存在相同或者不同的片段。本文针对不同来源的DAN序列，利用分组概率值的统计特征进行计算，采用三种非线性函数获得测量的投影测度，得到对应的基因测量特征形成可视化的聚类分布。比较结果显示，同类基因处理结果分层趋势相同，基因子序列分布图示在更高层次呈现出互补结构，而不同种类基因序列之间存在明显的分布差异。

关键词

DNA聚类，概率值，非线性函数，基因组序列

1. 引言

DNA 序列由四元符号{A、T、G、C}组成，从生物学角度理解，它是由四种碱基线性组合而成，并呈现出一定的互补原则；就计算机学科而言，核苷酸表示的译码信息与计算机中“0”“1”所代表的信息一样。因此，DNA 序列从计算机的层次分析，可当作一串随机自然密码，其中隐藏着生物学的规律。针对 DNA 序列分析的模型和方法在现代基因组的各类应用中扮演着重要角色[1]，利用可视化工具展现序列的已知与未知联系，对 DNA 计算[2]和密码学领域的研究和应用提供辅助参考价值。

聚类是一种在数据挖掘中常用的方法，能对无监督数据根据其相似性进行划分并归为可区分的子类。利用聚类分析，通过对相似基因组表达模式的挖掘可以推测出未知基因组的结构和功能。目前用于基因聚类分析的常见方法有把属于同一类的个体间距离尽可能小，把不同类个体间距离尽可能大的分层聚类[3]如 K-means 算法[4]、还有对样本的概率密度分布进行估计的基于混合高斯模型的聚类算法[5]等。通过利用相关的测量模型和方法，形成聚类分布能对后续使用聚类方法进行处理提供支持。

本文所描述的处理模型是对特定分组的基因片段，通过统计相同概率值在非线性函数作用下形成聚类分布的处理过程。该类方法基于概率值[6]，将适用于随机序列的统计分析[7]方法推广到基因序列中，对其进行整理归类，是一类具有应用价值的探索模式。

2. 系统体系构架

概率统计模型[8]是解决复杂问题的有效方法。对于繁杂无序的 DNA 序列，可以认为，反映序列特征最重要的有两方面，其一是碱基的排列顺序，其二是碱基的含量。因此，考虑碱基在序列中所占比率也具有一定意义，根据分组概率值能进行可视化探索。

本文所构建的非线性函数 DNA 概率测量系统体系框架如图 1 所示。

系统处理包含概率值计算，非线性函数测量和可视化处理三个模块。概率值计算时需要进行四个操作，为递进关系。基于上述框架，处理流程可简述如图 2 所示。

输入的 DNA 序列依次进行分组、统计、分类、概率计算后，进入非线性函数处理的流程，最终将输出可视化的聚类分布图形。

3. 核心模块方法

3.1. 序列分组统计

选取 DNA 序列 T，按照以下模式处理：(见图 3)

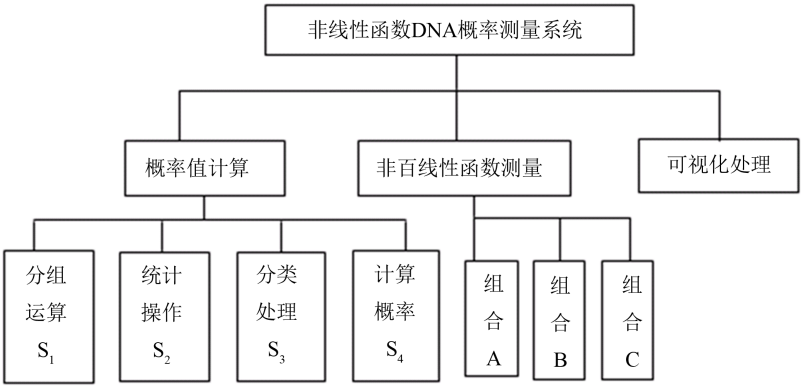


Figure 1. Nonlinear function of DNA probabilistic measurement system
图 1. 非线性函数 DNA 概率测量系统框架图

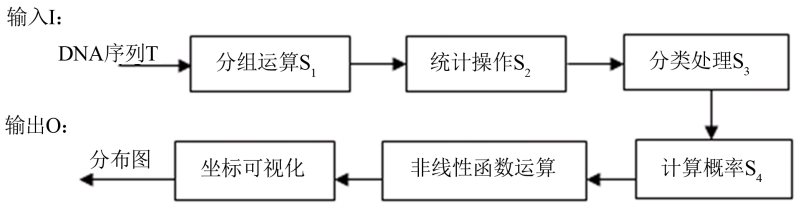


Figure 2. Process flow diagram
图 2. 处理流程简述图

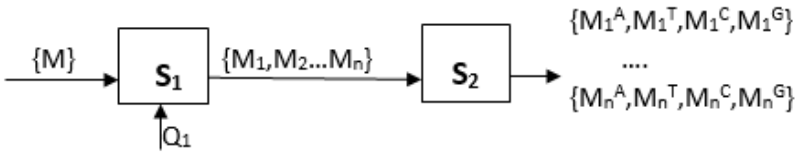


Figure 3. DNA sequence grouping statistical methods
图 3. DNA 序列分组统计方法

输入组：

{M}表示序列 T 的个数集合，M∈N，N 为正整数。

中间组：

S_1 运算盒表示将序列 T 进行一定数量 Q_1 的分组。运算后，序列 T 的个数被分为 n 个序列 T_n，其中 $0 \le M_n \le M$，$n = \lfloor M/Q_1 \rfloor$，$M_n$ 表示分组后一个序列的碱基个数，且 $M_1 = M_2 = \cdots = M_n = Q_1$，多余或不足数据舍弃。

S_2 运算盒将 S_1 得到的 n 个序列逐一统计相应碱基数目。

输出组：

经过 S_2 处理后，对于每一个序列 T_n，都将输出四个值 $M_n^A, M_n^T, M_n^C, M_n^G$，$\{M_n^A, M_n^T, M_n^C, M_n^G\}$ 表示此序列 T_n 中 ATCG 的个数。

至此序列一级分组的工作完成，有 M 个碱基的序列 T 经过上述操作后，被分为 n 组，得到的数据量为 4n。

3.2. 序列分类统计

如图 4 所示。

输入组：

$M_1^A, M_2^A \cdots M_n^A$ 为 S_2 处理后每个序列 T_n 中碱基 A 的个数，总计 n 个，M_1^A 与 M_n^A 可能相同，也可能值不一样。

中间组：

S_3 运算盒表示将输入数据按照 Q_2 个一组进行分类划分，得到 j 组，j = [n/Q_2]。为使最终可视化分别更具有表达性，$Q_2 > Q_1$。

输出组：

输出的 j 组数据中，每组有 Q_2 个数据。

对于其他碱基，按照相同步骤处理。

3.3. 概率值 P 计算

如图 5 所示。

输入组：

$\left\{M_1^A, M_2^A \cdots M_{Q2}^A\right\}$ 为 S_3 的第一组输出结果。

中间组：

S_4 运算盒的处理过程如下：

1) 将 Q_2 个数据的值相加，得到总和 sum；

2) 检索 Q_2 个值，把相同值相加。例如有 5 个数据分别为 32, 5, 40, 32, 21，处理后所得结果为 64, 5, 40, 21；

3) 用步骤 2 中所得结果分别除以总和 sum，得到其概率值为 $P_1, P_2 \cdots P_{Q2}$，最多 Q_2 个 P 值。

输出组：

输出概率 $P_1, P_2 \cdots P_{Q2}$ 且 $P_1 + P_2 + \cdots + P_{Q2} = 1$，$0 \le P_n \le 1$。

对于其他数据，按照相同步骤处理。

3.4. 非线性函数测量

设对概率值 P 处理的非线性函数为 $f_1(P), f_2(P)$。$f_1(P)$所计算的值作为横坐标 X 的值，$f_2(P)$所计算的值作为相应纵坐标 Y 的值。这样对于每一个 P_n 值，可以在平面坐标系中得出一个点 $W_n(f_1(P_n), f_2(P_n))$：(见图 6)。

表 1 给出了本次所采用的非线性函数，对一个 P_n 值，采用三种不同的非线性函数计算：

sqrt(x)表示求 x 的平方根；pow(x, y)表示求 x 的 y 次方；exp(x)表示 e 的 x 次方。

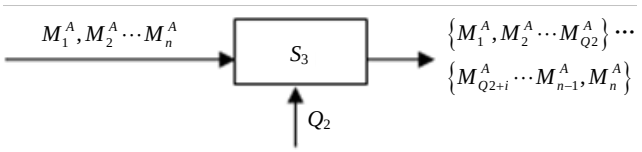


Figure 4. DNA sequence classification statistical methods
图 4. DNA 序列分类统计方法

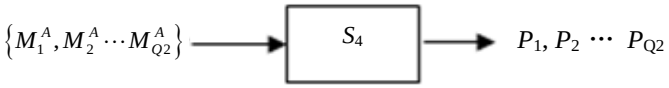


Figure 5. DNA sequence probability P calculated methods
图 5. DNA 序列概率值 P 计算方法

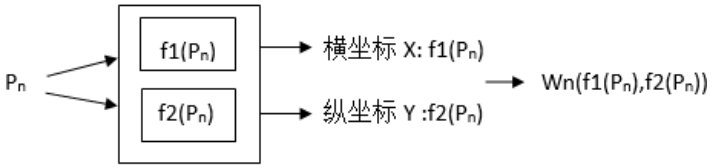


Figure 6. Nonlinear function of the probability value processing methods
图 6. 非线性函数对概率值的处理方法

Table 1. Nonlinear function evaluating expressions
表 1. 非线性函数计算表达式

投影类型	$f_1(P)$	$f_2(P)$
A	sqrt(P)	pow(P,4)
B	sqrt(P)	exp(P)
C	pow(P,4)	exp(P)

4. 测量结果

本次试验中，选取了基于 RC4[9]算法产生的伪随机数对应的 DNA 序列[10]、动物鼠[11]和植物水稻[12]的部分碱基。

方便计算，定义总碱基数目 M = 1,200,000，$Q_1 = 20$，$Q_2 = 30$，则 $n = M/Q_1 = 60{,}000$，$j = n/Q_2 = 2000$。

为了将所得到的坐标形象化的显示出来，使用 MATLAB[13]绘制图像。统一坐标后，得到以下可视化的结果：

A. $f_1(P) = sqrt(P)$, $f_2(P) = pow(P, 4)$:

B. $f_1(P) = sqrt(P)$, $f_2(P) = exp(P)$:

C. $f_1(P) = pow(P, 4)$, $f_2(P) = exp(P)$:

5. 比较分析聚类结果

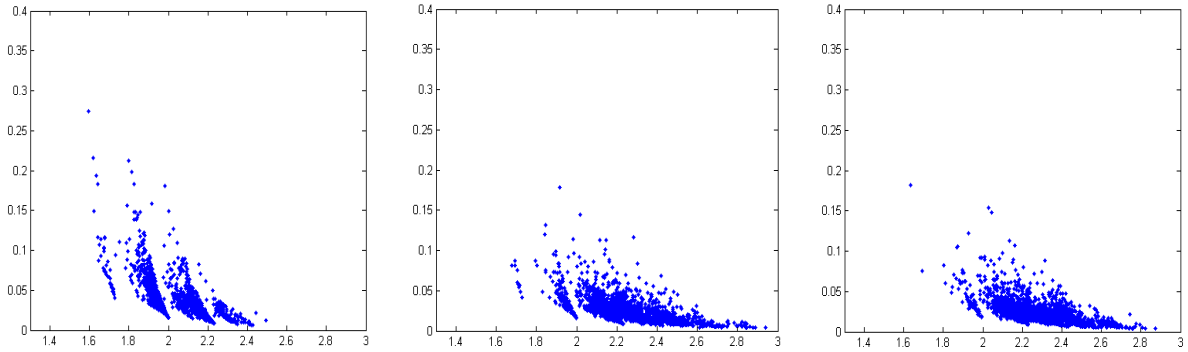
通过数据挖掘可视化的投影方法，绘制出了上文所示的各种聚类分布图，可获得不同的分布信息。

观察比较图 7 和图 8，基于 RC4 算法产生的伪随机数对应的 DNA 序列与动植物所得到的可视化结果存在比较明显的差异。前者主要分布在 1.6~2.4 的区间，而对于鼠和水稻的显示看来，其开合度和聚散程度有着很大的相似度，集中体现在 1.8~2.8。能看出，RC4 碱基 A、T 与鼠和水稻的 C、G 分布相似，而 RC4 碱基 C、G 与鼠和水稻的 A、T 分布形态相同。图 9 并无明显的分层现象，碱基分布均呈现出束状。但同等条件下 RC4 开合度更大，鼠和水稻分布范围更广。

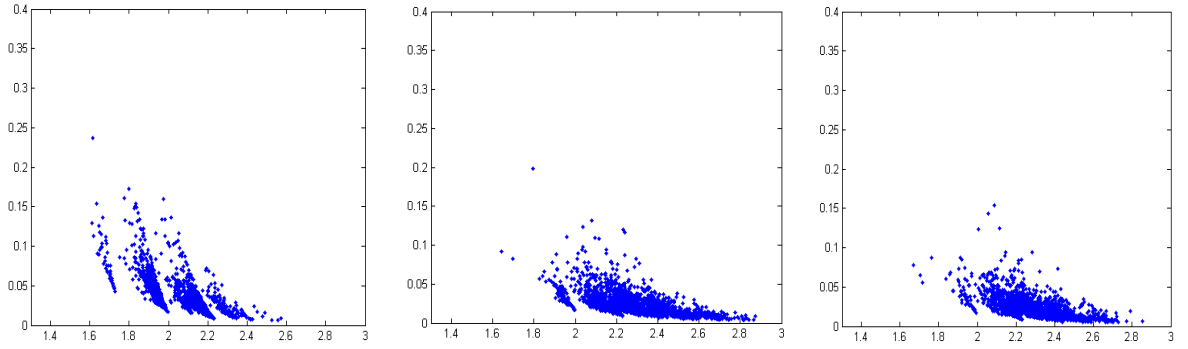
综合着不同算法相比较，碱基序列在同种非线性函数的测量下，所得到的图形走势大体相同。碱基 A 与 T，C 与 G 的延展性也整体一致，说明了它们互补的特性。此外，选取 C 函数测量时没有明显聚类效果，区间范围也不同，而 A、B 两种的分层效果明显，聚集范围一致。但 B 函数影响下的聚类分层可视化结果最为清晰。

6. 结束语

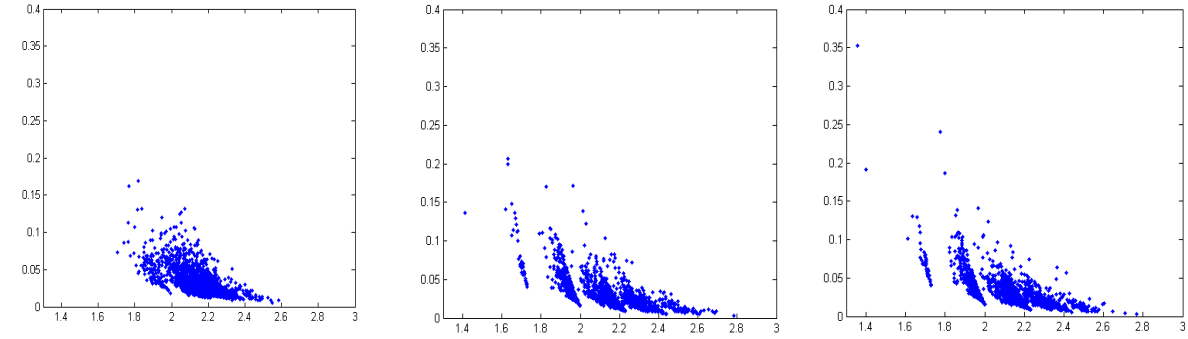
DNA 序列的分析是现代分子生物学中最重要的部分[14]。从生物数据库中的基本序出发，结合当前新兴的数据挖掘技术，分析基因数据，比较 DNA 序列相似性，可以为生物信息、计算机安全学等方面提供一定的研究基础。



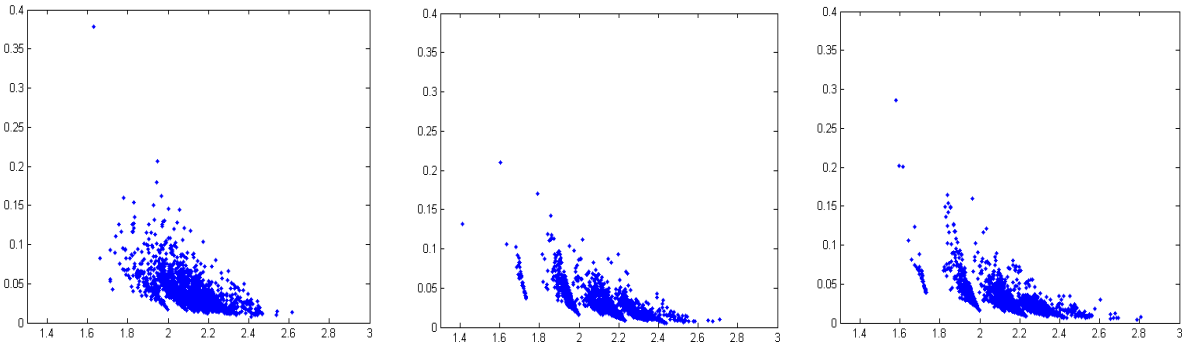
RC4(A)鼠(A) 水稻(A)



RC4(T)鼠(T) 水稻(T)

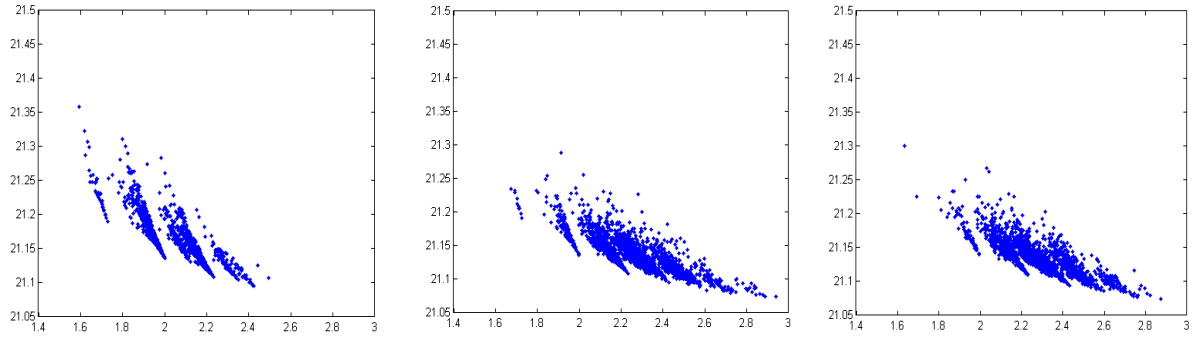


RC4(C)鼠(C) 水稻(C)

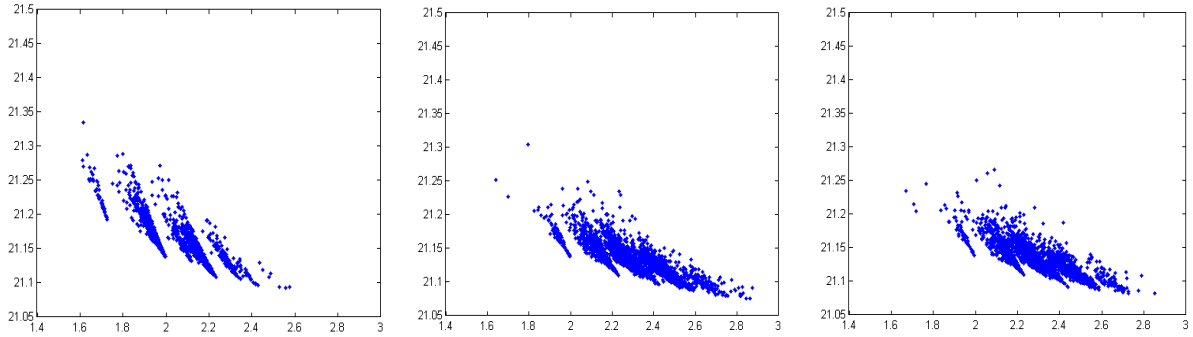


RC4(G)鼠(G) 水稻(G)

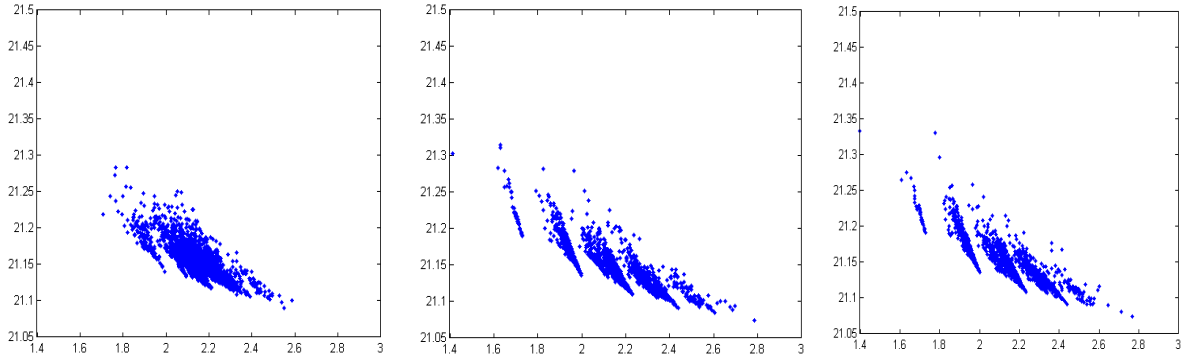
Figure 7. Class A projected visualization results
图 7. A 类投影可视化结果



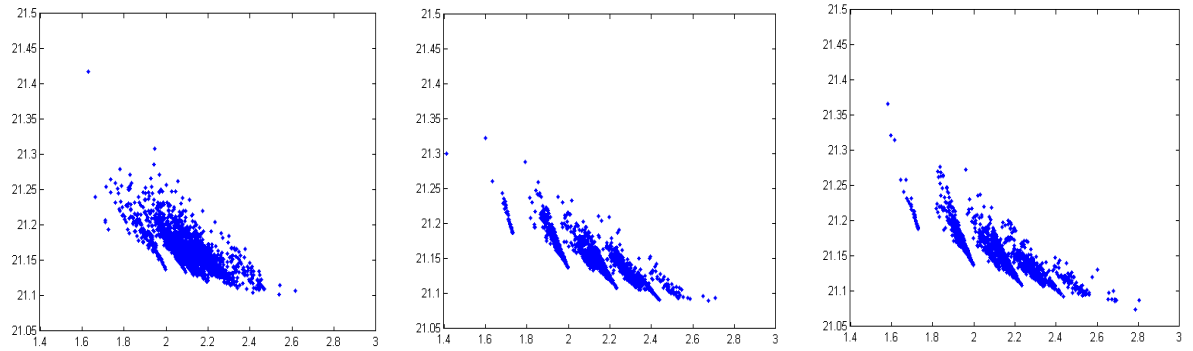
RC4(A)鼠(A) 水稻(A)



RC4(T)鼠(T) 水稻(T)

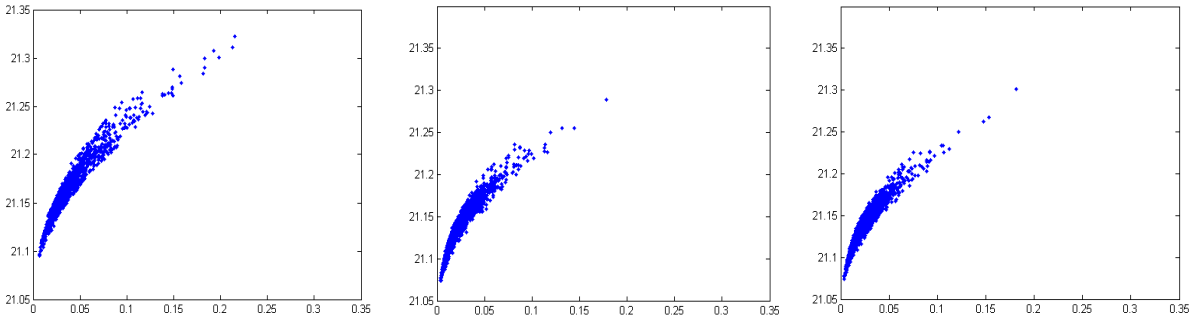


RC4(C)鼠(C) 水稻(C)

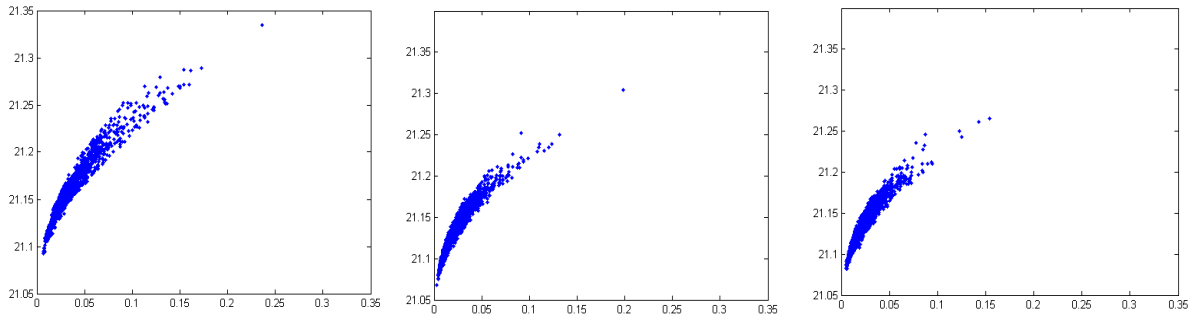


RC4(G)鼠(G) 水稻(G)

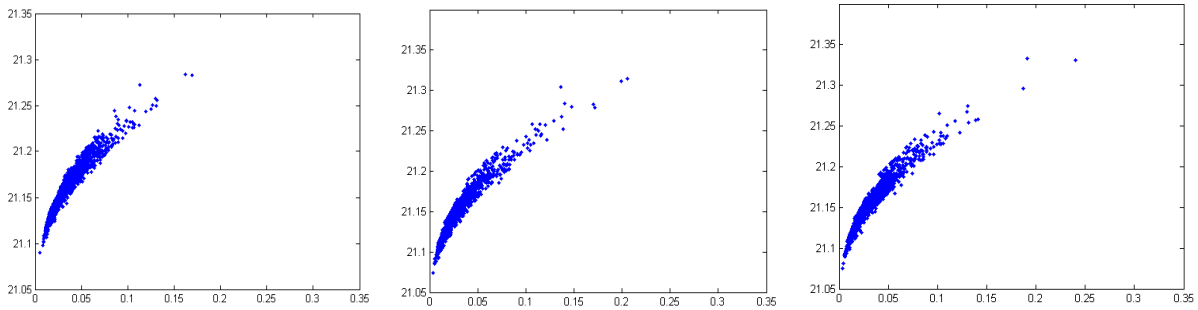
Figure 8. Class B projected visualization results
图 8. B 类投影可视化结果



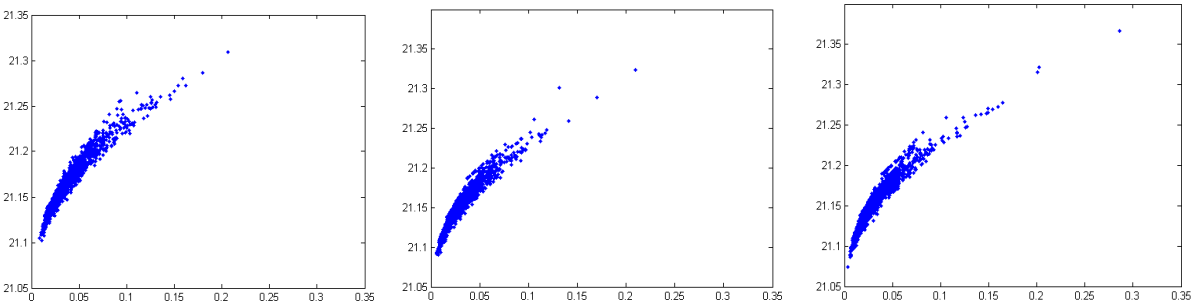
RC4(A)鼠(A) 水稻(A)



RC4(T)鼠(T) 水稻(T)



RC4(C)鼠(C) 水稻(C)



RC4(G)鼠(G) 水稻(G)

Figure 9. Class C projected visualization results
图 9. C 类投影可视化结果

本文通过分析研究不同来源的 DNA 序列，在数据挖掘技术和可视化的基础上，设计了一种基于概率值和非线性函数操作的测量方法。运用该模型，在相同函数的影响下，可以得到趋势相同的分层效果图。由于文中仅针对部分的片段测量结果进行初步研究，存在不可避免的局限及不足。对应的不足之处希望

在下一步的研究上继续完善。

致　谢

感谢国家自然科学基金、云南大学软件学院以及云南省软件工程重点实验室对信息安全研究项目的基金支持。

基金项目

国家自然科学基金资助项目(61362014)；云南大学软件学院 2013 年第四届教育创新基金资助项目(学生专项）

参考文献 (References)

[1] Lieberman-Aiden, E., et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.

[2] Zheng, J., Zhang, W.Q., Luo, J., et al. (2013) Variant map system to simulate complex properties of DNA interactions using binary sequences. *Advances in Pure Mathematics*, **3**, 5-24.

[3] Eisen, M., Spellman, P., Brown, P., et al. (1998) Parallel human genome analysis: Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**, 14863-14868.

[4] Tavazoie, S., Hughes, J.D., Campbell, M.J., et al. (1999) System-attic determination of genetic network architecture. *Nature Genetics*, **22**, 281-85.

[5] Yeung, K.Y., Raley, C., Murua, A., et al. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977-987.

[6] Beyer, O., Hackel, H., Pieper, V. and Tiedge, J. (1980) 概率计算和数学统计. Harri Deutsch 出版社.

[7] Chance, B.L. and Rossman, A.J. (2005) Preface. In: *Investigating Statistical Concepts, Applications, and Methods*, Duxbury Press, New York.

[8] 吴赣昌 (2008) 概率论与数理统计. 中国人民大学出版社, 北京.

[9] Schneier, B. (1995) Chapter 17—Other Stream Ciphers and Real Random-Sequence Generators. In: *Applied Cryptography*: *Protocols, Algorithms, and Source Code in C*, 2nd Edition, Wiley, New York.

[10] 张巍琼, 郑智捷 (2012) 基于不同产生机制的伪随机序列和DNA序列的随机性测量. *成都信息工程学院学报*, **6**, 文章编号: 1671.

[11] http://asia.ensembl.org/Mus_musculus/Info/Index

[12] ftp://ftp.ncbi.nih.gov/genomes/

[13] Chapman, S.J. (2008) MATLAB Programming for Engineers. 2nd Edition, 清华大学出版社, 北京.

[14] Bu, Q.X. and Zheng, J.Z.J. (2013) 2D Conjugate Maps of DNA Sequences. *Journal of Information Security*, **4**, 193-196.