Hans Journal of Computational Biology 计算生物学, 2014, 4, 20-31
Published Online June 2014 in Hans. http://www.hanspub.org/journal/hjcb
http://dx.doi.org/10.12677/hjcb.2014.42003

Hans汉斯

The Visual Analysis of Coding and Non-Coding DNA Sequences

Yuqian Liu, Zhijie Zheng

School of Software, Yunnan University, Kunming
Email: 33077511@qq.com

Received: Apr. 22nd, 2014; revised: Apr. 28th, 2014; accepted: May 5th, 2014

Abstract

DNA sequences include complex genetic information; their specific characteristics are contained in both the coding and non-coding sequences. Major gene components in higher levels of organisms are composed of non-coding sequences. In ENCODE project, there are evidences that 98% of the human genomes are non-coding forms and 80% of them with functions, so the research on coding region and non-coding region has become an important research hotspot. This paper provides models and experiment results which using visual representation techniques to distinguish differences between coding and non-coding sequences. This model uses probability measurements on the DNA sequences to coding and non-coding regions respectively to distinguish patterns identified from different sequences.

Keywords

Non-Coding Sequences, Graphic Representation Technique, Probability Measurements

编码和非编码DNA序列的可视化分析

刘玉倩，郑智捷

云南大学软件学院，昆明
Email: 33077511@qq.com

收稿日期：2014年4月22日；修回日期：2014年4月28日；录用日期：2014年5月5日

摘　要

DNA序列作为一种复杂的遗传信息，其具体特性不仅体现在编码序列之中，也包含在非编码序列之中。在高等生物体中主要基因成分为非编码序列，在ENCODE计划中，有证据表明，在人类基因中有98%为非编码形式，其中80％具有功能性，所以对编码区和非编码区的研究已经成为一类重要研究热点。本文提供的模型和实验结果，使用图形表示方法对编码区以及非编码区基因的差异进行区分。该模型采用的是对编码区以及非编码区的DNA序列进行分段概率测量，从而对不同的基因特征分布进行比较。

关键词

非编码序列，图形表示方法，概率测量

1. 引言

人类基因组计划(HGP)[1]测序结果表明，DNA 编码序列占人类基因组的 2%，剩余的 98%的均为非编码基因。非编码区广泛存在于真核生物中，众多的非编码区在生命活动中具有广泛的调控作用。ENCODE 计划[2]提供了详尽的非编码功能单位的功能图谱，在 98%的非编码基因中有 80%是功能性的[3]。数据表明，在非编码功能单位中富含突变体，有时突变体出现在与特定性状相关的特异细胞中，这说明这些区域可能与疾病相关。从生物进化的观点看来，随着生物体功能的完善和复杂化非编码区序列明显增加的趋势表明：这部分序列必定具有重要的生物功能。普遍的认识是，它们与基因在四维时空的表达调控有关。因此寻找这些区域的编码特征以及信息调节与表达规律是生物信息学[4]的重要研究内容。

面对基因组计划所得到的海量生物学数据，如何分析并从中获得生物学信息是后基因组时代[5]的首要任务。DNA 序列作为一种遗传语言，不仅体现在编码区的序列之中，而且隐含在非编码区[6]的序列之中。基因编码区就是能够翻译成为某种蛋白质的 DNA 序列区域(也就是基因)。近年来完整基因组的研究表明，在细菌这样的微生物中非编码区只占整个基因组序列的 10%到 20%。而高等生物和人的基因组中非编码区都占到基因组序列的绝大部分[7]。

DNA 序列分析主要是分析序列中所表达的结构和功能的生物信息。其研究内容非常丰富，如：序列比较[8]、基因识别等，而图形表示是最近发展起来的应用在 DNA 序列分析方面的强有力的可视化工具[9]，它能够揭示蕴藏在 DNA 序列中的结构和功能的生物信息，可视化分析在人类基因组计划中扮演着重要的角色。

面对海量 DNA 数据，用传统方法研究这些数据，已经不能满足我们的需要。图形表示方法可以直观有效的完成 DNA 序列分析，相对于传统方法来说，可以缩短研究的进程。图形表示方法是对 DNA 序列进行分析的一种工具[10]，在人类基因组计划中，已经成功应用图形表示方法，对 DNA 进行分析处理并得到 DNA 图谱。但是，人类基因组的可视化模式并不适用对编码以及非编码区进行分析，本文提出一种新的方法，它主要应用数学统计的原理，对 DNA 数据进行处理，从而得到可视化结果，对编码以及非编码区的基因分布特性进行分析。

2. 系统架构

在本节中，讨论了系统架构及其组成部分的使用图，定义了测量模型中的公式以及相关变量。

2.1. 体系结构

在本文的体系中包含有三个部分，分别是 DNA 概率测量映射，坐标位置映射，以及图形投影。如图 1。

读取 DNA 序列，选择 N 个基因作为选定的 DNA 序列，作为输入数据。经过概率测量映射模块，得到归一化的概率测度。每个碱基的归一化测度对应计数均为 0 至 1 间的一个百分比，将该数据作为坐标位置映射部分的输入，按照一定的规则，对其进行处理，最终得到每个点的横纵坐标。通过横纵坐标值可以确定该点在笛卡尔坐标中的位置，将每个坐标点作为图形投影部分的输入值，集合所有的选定的 DNA 序列，进行图形投影，最终得到编码以及非编码区的 DNA 特征分布图。

2.2. 核心模块

2.2.1. 概率测量映射

如图 2 所示的概率映射部分由三个模块组成：柱形图，归一化柱形图和归一化测度构成。这一部分是为了得到归一化测度，以便为后面的坐标位置映射部分提供输入值。

对概率测量映射模块，详细处理流程解释如下：

中间组：

四个碱基的概率测量，表示某一个分段含有某一个碱基的数量。

相关概率测量：四个碱基得到四柱形图的相关概率测量。

输出组：

归一化测度：四归一化柱状图的相关概率测量，表示在一个分段中某一碱基在总分组中所占的比例。

数据经过预处理，形成分段模型，对于选定的 DNA 序列，以各分段中所含某碱基的数量为水平坐标，概率测量相同的则进行叠加，可以得到柱形图。对柱形图进行归一化处理，形成归一化的柱形图，每个碱基的归一化测度对应计数都是 0 到 1 之间的一个数(百分比)，且每一个碱基在同一分组之中的计数之和为 1。

2.2.2. 坐标位置映射

如图 3 所示的坐标位置映射部分，该模块的详细流程描述如下：

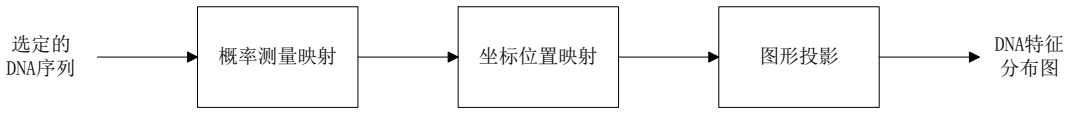


Figure 1. Architecture
图 1. 体系结构

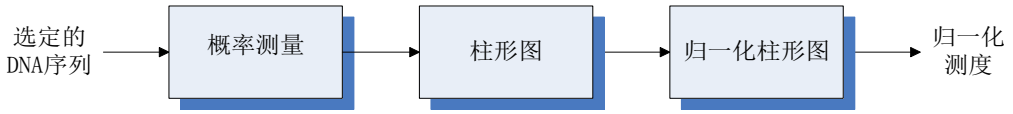


Figure 2. Probability measurement
图 2. 概率测量

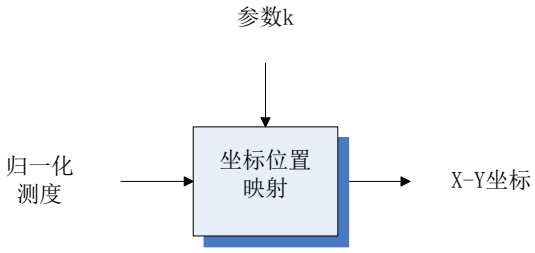


Figure 3. Coordinate position map
图 3. 坐标位置映射图

输入组：

归一化测度：四归一化柱状图的相关概率测量，表示在一个分段中某一碱基在总分组中所占的比例。

可变参数 k：一个可变的控制参数，$n > 0$ n∈Z。

输出组

四个成对的 X-Y 坐标值。

这一部分是为了得到点的坐标，从而进行投影映射，得到可视化结果。概率测量映射得到的归一化测度以及可变参数 k 作为输入信号,在给定的规则下完成坐标映射。最终生成的是不同参数条件下的 X-Y 坐标。

2.2.3. 图形投影

如图 4 所示的图形投影部分，该模块的详细流程描述如下：

输入组：

选择所有的 DNA 序列，即将所有的分组一起映射到 DNA 特征分布图。

四个成对的 X-Y 坐标值。

输出组：

四个二维的 DNA 特征分布图。

这一部分是为了得到最终的 DNA 特征分布图。在坐标位置映射部分产生的坐标值作为基本输入，但由于一个分组只能产生一对坐标值，因此收集所有选定的 DNA，经过图形投影，得到 DNA 特征分布图。

3. 详细描述

3.1. 参数说明

n 一个分段中的元素数量，$n > 0$；

V 象征四个 DNA 符号中的一个，$\{A,G,T,C\}=D$，$V \in D$；

k 表示一个控制参数，$k > 0$；

m_t 表示第 t 个分段；

X^{N_j} 表示具有N_j长的第 j 个 DNA 序列

$$X^{N_j} = \left(X_0, X_1, X_2, X_3, X_4, \cdots, X_k, \cdots, X_{N_j} - 1\right) \quad X_k \in \{A, G, T, C\} \quad 0 \le k \le N_j, 0 < j < M ;$$

M 表示 M 串 DNA 序列；

$\left\{P_H\left(T_i^V\right)\right\}$ 表示四个概率测量值 $0 \le i \le t,\ V \in D\{A, G, T, C\} = D$；

$\left\{\left(X_V^k, Y_V^k\right)\right\}$ 表示一组成对的坐标值，$k > 0$，$V \in D\{A,G,T,C\} = D$；

T_i^V 四个碱基的概率测量，表示某一个分段含有某一个碱基的数量；

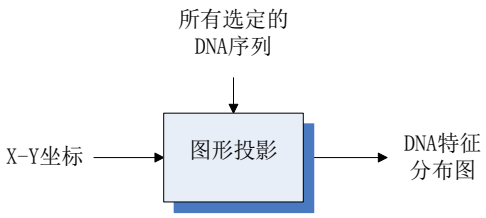


Figure 4. Graphical projection
图 4. 图形投影

$H\left(T_i^V\right)$四柱状图相关概率测量， $V \in D\{A,G,T,C\} = D$；

$\left\{P_H\left(T_i^V\right)\right\}$四归一化柱状图的概率测度，表示在一个分段中某一碱基在总分组中所占的比例，即

$$P_H\left(T_i^V\right) = T_i^V \Big/ \left(\sum_0^t T_i^V\right) V \in D\{A,G,T,C\} = D.$$

3.2. 概率测量

全部的 DNA 序列，分成长度为 N 的 j 个分组，将 N 个 DNA 序列分为每一段包含固定 n 个元素数量的 t 个分组。

在第一段至第 t 段，让 $0 < i < t$，利用统计原理，每个分段之中某一碱基的含量为 T_i^V，

则有 $T_i^A + T_i^T + T_i^C + T_i^G = n$，每个分段序列的概率为 $T_i^{V\,t-1}_{i=0}$，由此可以得到柱状图分布，记为 $H\left(Ti_v\right)$。它满足条件：

$$H\left(Ti_v\right) = 1,\ 0 \le T_i^V \le n$$

收集全部的，就可以建立柱状分布图

$$H\left(T_v\right) = \sum_{I=0}^{t-1} H\left(T_i^V\right)$$

在这个条件下，四个碱基的概率为 $P_H\left(T_v\right) = \left\{T_l^A, T_l^T, T_l^C, T_l^G\right\}_{l=0}^{n}$，$T_l^v \in [0,1]$。

四个向量进行归一化为

$$\sum_{l=0}^{n} \quad T_l^V = 1。$$

至此，四个碱基形成了完整的测量向量。

3.3. 坐标位置映射

使用上述的测量向量，可以用两个映射函数计算的值来映射到一个二维从而可以分析 DNA 序列。

让 $y_1 = F\left(P,V,k\right)$，$x_1 = F\left(P,V,\frac{1}{k}\right) y_2 = F\left(\left(P,V\right)^{\frac{1}{k}}\right)$，$x_2 = \left(F\left(P,V\right)\right)^{\frac{1}{k}}$

$\left\{\left(x_v^k, y_V^k\right)\right\}$是由以下方程定义的两组值，

$$x_1 = x_v^k = F\left(P,V,1/k\right) = \sqrt[k]{\sum_{l=0}^{n}\left(p_l^v\right)^k}$$

$$y_1 = y_v^k = F\left(P,V,k\right) = \left(\sum_{l=0}^{n}\sqrt[k]{p_l^v}\right)^k$$

$$x_2 = x_v^k = \left(F\left(P,V\right)\right)^{1/k} = \sqrt[k]{\sum_{i=0}^{n} e^{pi}}$$

$$y_2 = y_v^k = F\left(\left(P,V\right)^{1/k}\right) = e^{\left(\sum_{i=0}^{n}\sqrt[k]{pi}\right)}$$

成对的坐标值分别在直角坐标系中的特定位置形成一个点，从而可以形成二维的 DNA 特征分布图。

3.4. 图形投影

确定选定的 DNA 序列(即 X^{N_j})，在直角坐标系中的特定位置只产生一个坐标点，所以有必要应用相对大量的 DNA 序列作为输入来产生可见的分布。这种类型的操作在图形投影之中完成。

对于每个分段均按照上述的规则进行操作，每个分组 X^{N_j} 得到一个坐标点，j 个分组则会得到 j 个坐标点，将这 j 个坐标点映射到同一个坐标系之中，可以得到四个碱基的 DNA 特征分布二维图。同时，对于同一生物的编码区以及非编码区，还要进行数据量以及坐标的统一，将各个变量进行统一，以免对结

果造成影响。

4. 示例结果

4.1. 样品结果

利用 DNA 序列编码以及非编码区的文件，在两个可控的参数下(分段长度 n，k)可以形成二维的图形，采用控制变量的方法，可以得到分段长度 n 以及可变参数 k 分别取值为多少时，可以得到较好的可视化效果。

如图 5(选用样品为线虫编码区碱基 A)。

由上图所示，图 5 表示分组长度 $N = 500$，分段长度固定为 10，选用不同的可变参数 k 所得到的 DNA 特征分布图。图(a)至图(f)分别表示 k 值为 2,3,4,5,6,7 的可视化效果。观察各个图的聚类效果，当 $k = 4$ 时的图形投影效果较好。

图 6(a)~图 6(f)为 k 值确定，分段长度不固定，$n = 5,10,15,20,25,30$ 时的投影图。

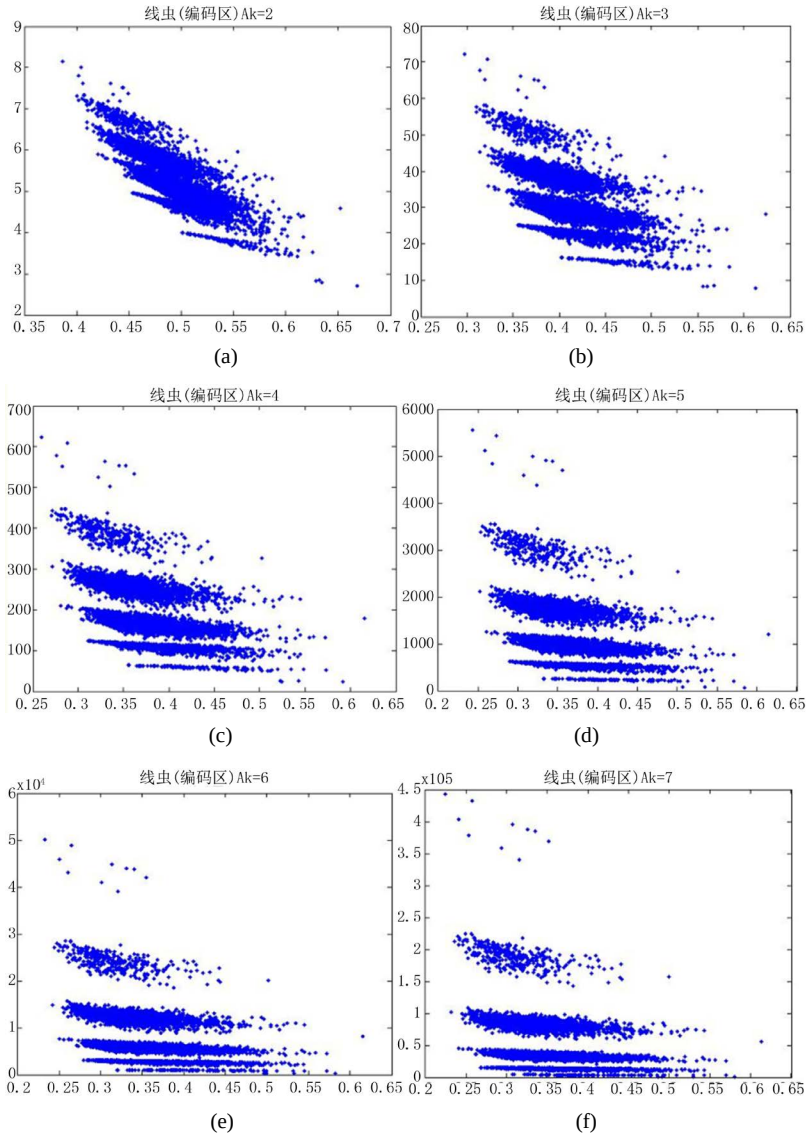


Figure 5. The variable results map of k value
图 5. k 值可变结果图

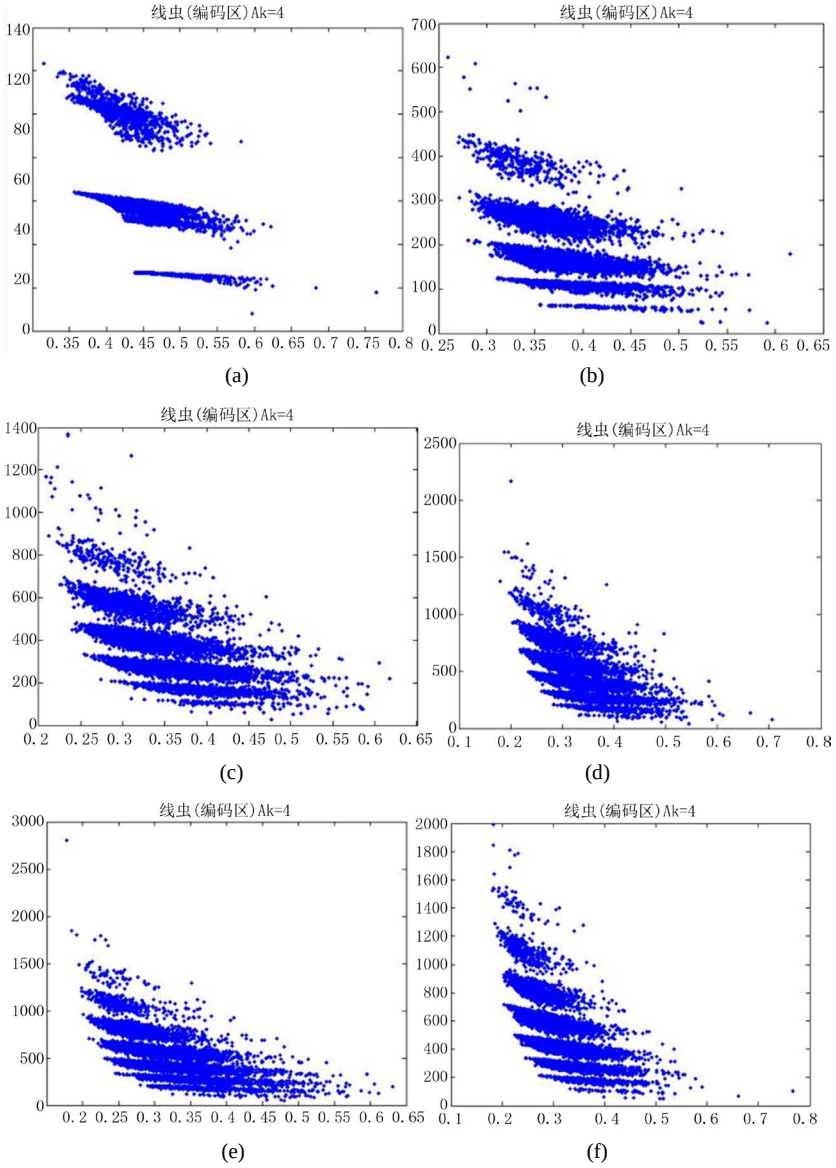


Figure 6. The variable results map of *n* value
图 6. *n* 值可变结果图

由上图所示，图 6 表示分组长度 $N = 500$，可变参数固定为 4，选用不同的分段长度 n 所得到的 DNA 特征分布图。图(a)至图(f)分别表示 n 值为 $5,10,15,20,25,30$ 的可视化效果。观察各个图的聚类效果，当 $n = 10$ 时的图形投影效果较好。

当 $n = 5$ 时，改变 K 的值，效果图 7：

在 $n = 5$ 时，由于分段长度较小，投影效果随着参数的改变并没有明显的差异，不适宜用于分析编码区与非编码区可视化的分析，故文中选用分段长度(即 n)为 10 进行可视化分析。

4.2. 结果分析

本文对拟南芥，玻璃海鞘，水稻，线虫，大肠杆菌，沙门氏菌，幽门螺杆菌以及黑猩猩的编码以及非编码进行可视化。结果图示本文以进化程度较高的线虫，黑猩猩以及比较低等的沙门氏菌为例。

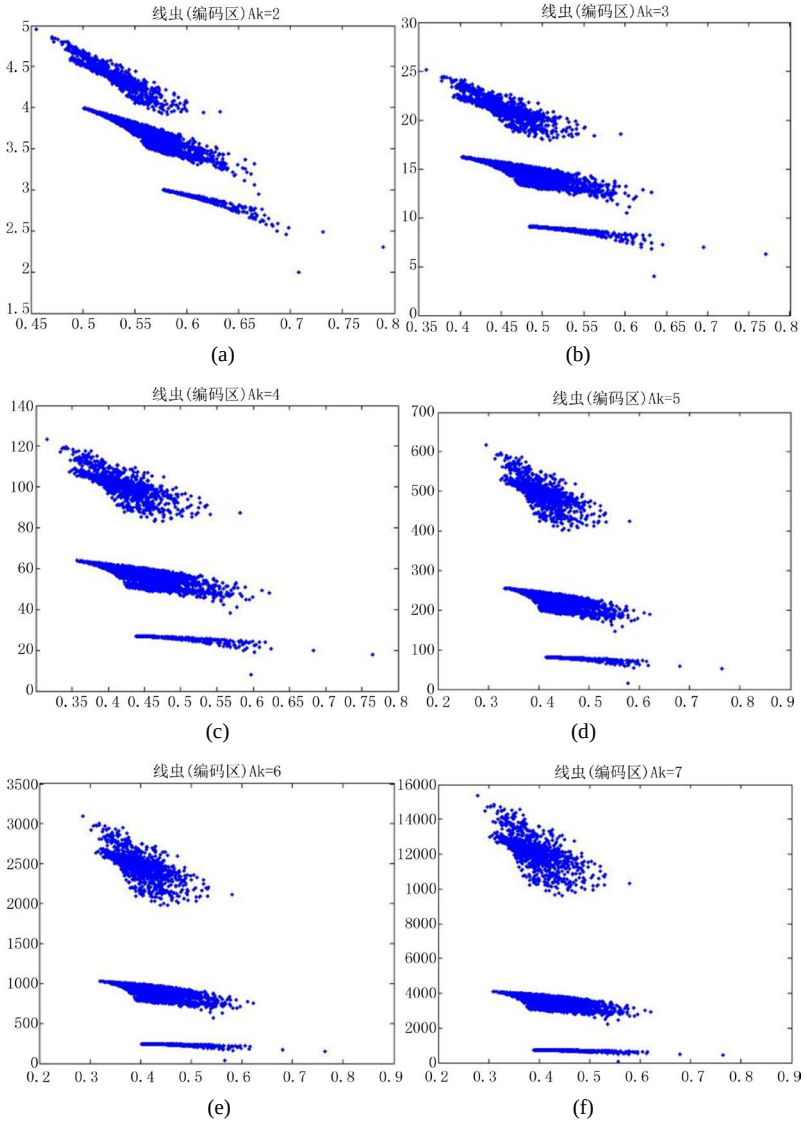


Figure 7. The variable results map of k value at n = 5
图 7. n = 5 k 可变结果图

图 8、图 9 为沙门氏菌编码非编码区在 $k=4, n=10, N=500$ 条件下的可视化效果。
图 10、图 11 为线虫编码非编码区在 $k=4, n=10, N=500$ 条件下的可视化效果。
图 12、图 13 为黑猩猩编码非编码区在 $k=4, n=10, N=500$ 条件下的可视化效果。

$$x_1 = \sqrt[k]{\sum_{l=0}^{n}\left(p_l^v\right)^k} \quad y_1 = \left(\sum_{l=0}^{n}\sqrt[k]{p_l^v}\right)^k$$

$$x_2 = \sqrt[k]{\sum_{i=0}^{n} e^{pi}} \quad y_2 = e^{\left(\sum_{i=0}^{n}\sqrt[k]{pi}\right)}$$

从 DNA 基因分布图可以看出，采用控制变量的方法，只有一个参数可变的情况下，DNA 特征分布图的可视化效果有显著的差别。分段长度为 10 时，k 值从 2 取到 7，图形的分层效果很明显，各个坐标点明显是从集中到分散再到集中，观察得出，k 值取 4 时，可视化的效果是最好的(如图 5 所示)。k 值固定为 4，分段长度分别为 5、10、15、20、25、30，可以观察到随着分段长度的增加，可视化图形的分层

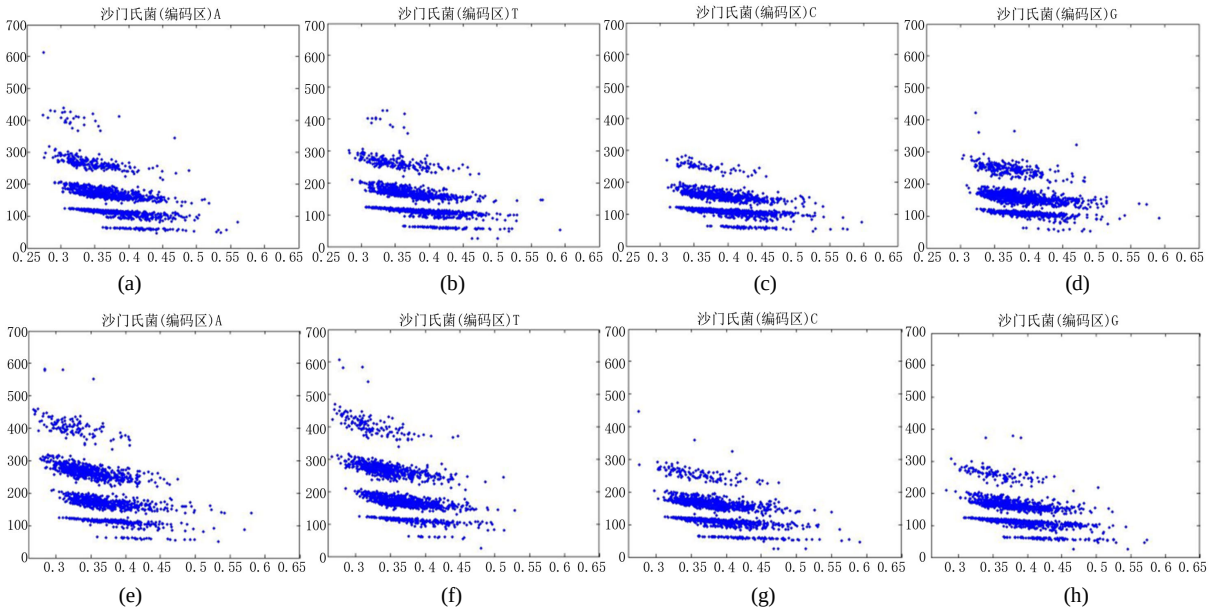


Figure 8. The visualization results of Salmonella 1
图 8. 沙门氏菌编码非编码区的可视化 1

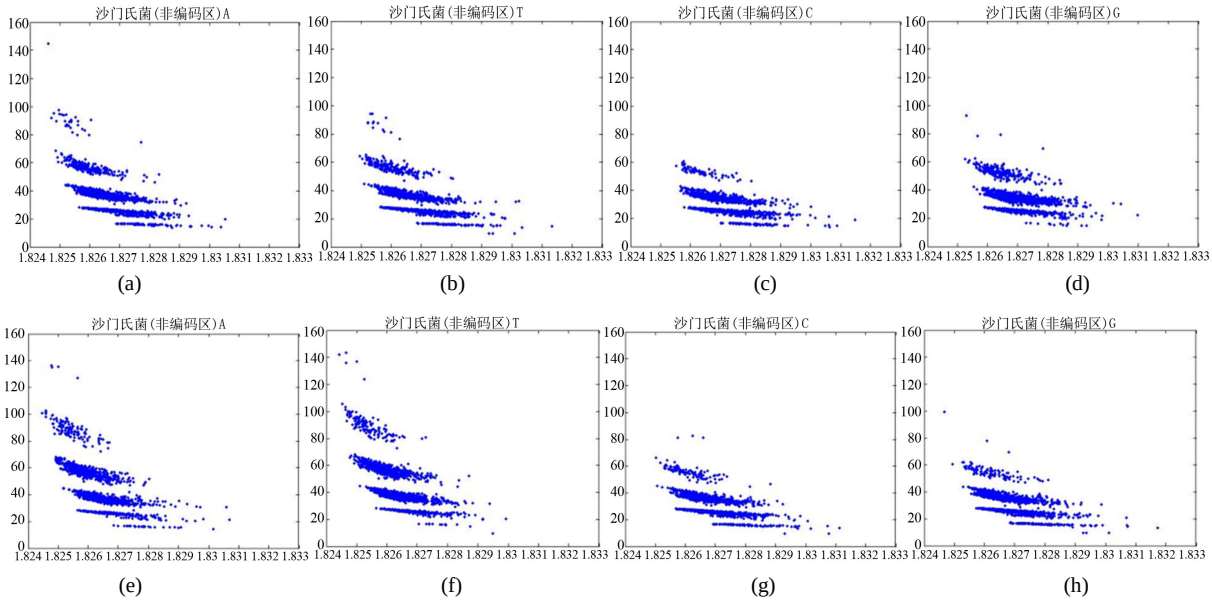
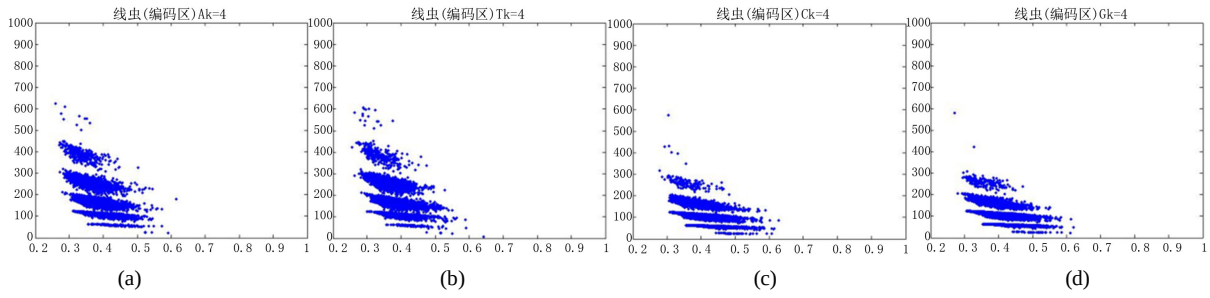


Figure 9. The visualization results of Salmonella 2
图 9. 沙门氏菌编码非编码区的可视化 2



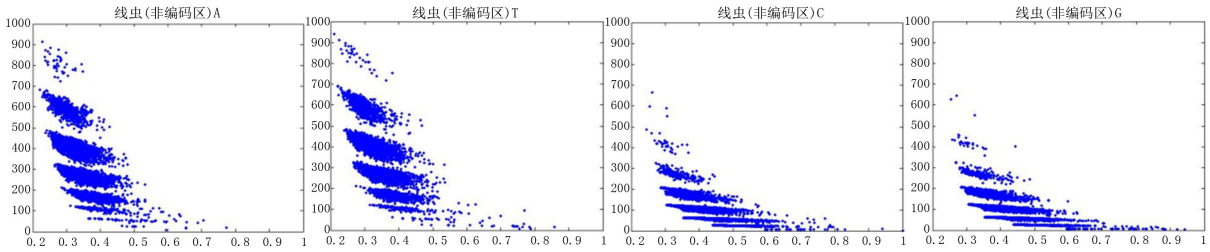


Figure 10. The visualization results of *Caenorhabditis elegans* 1
图 10. 线虫编码非编码区的可视化 1

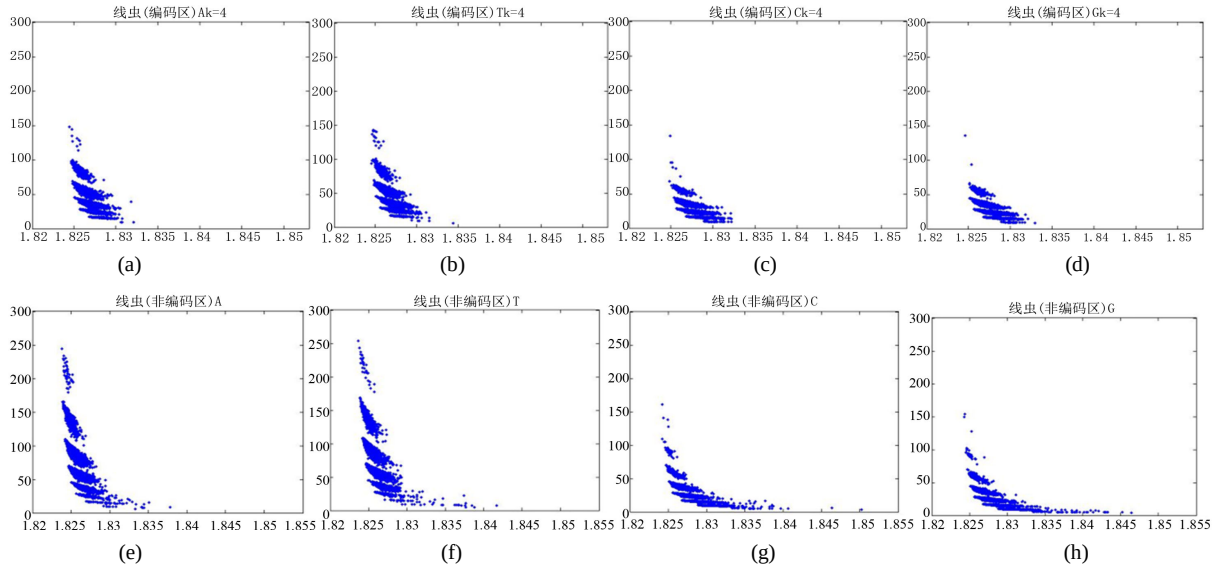


Figure 11. The visualization results of *Caenorhabditis elegans* 2
图 11. 线虫编码非编码区的可视化 2

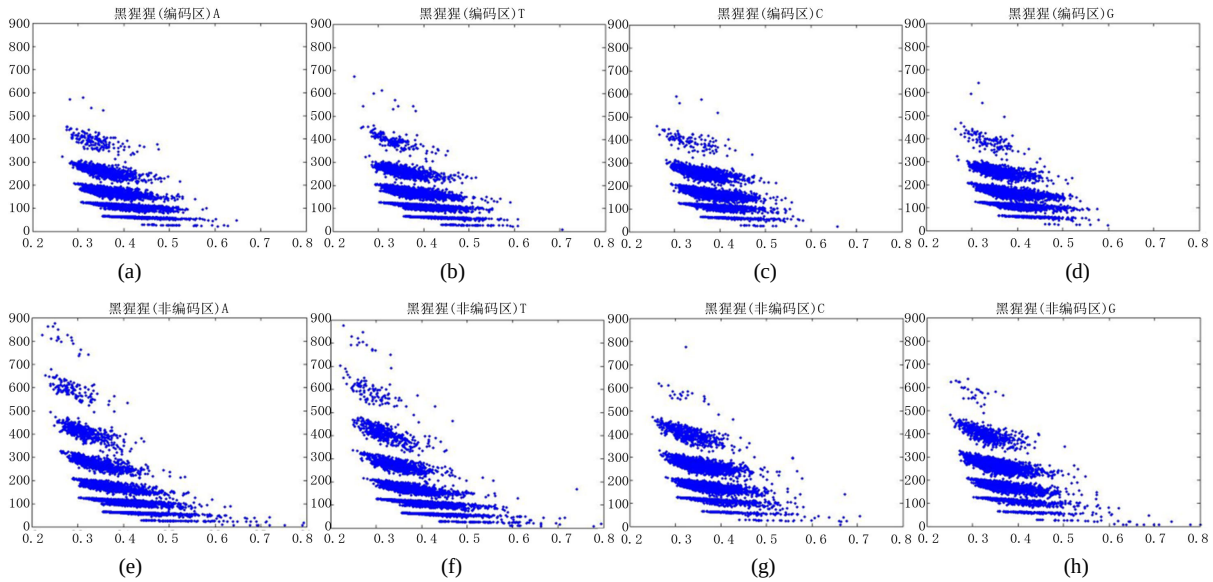


Figure 12. The visualization results of Pan_troglodytes 1
图 12. 黑猩猩编码非编码区的可视化 1

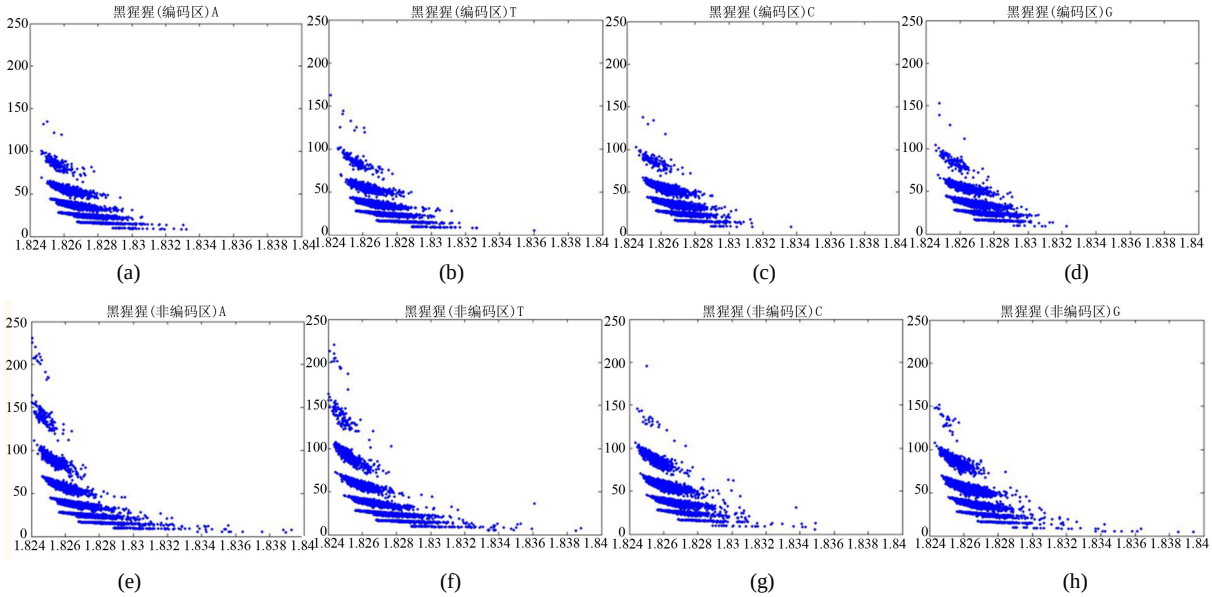


Figure 13. The visualization results of Pan_troglodytes 2
图 13. 黑猩猩编码非编码区的可视化 2

数量越来越多，同时也越来越紧凑，在长度为 5 以及 10 时，分层效果是最明显的(如图 6 所示)。

分段程度 n 为 5 时，对 DNA 数据进行可视化，DNA 特征分布图并不随着参数的改变而有明显的改变(如图 7 所示)，反复试验多组数据，仍得到这个结论，原因可能是因为分段长度过小，从而隐藏 DNA 特征分布。

观察沙门氏菌编码以及非编码的可视化效果(如图 8，图 9 所示)，两类图均存在 A-T 对称，C-G 对称，但是总体看起来 A-T 比 G-C 在非编码投影上要明显的多。但是两类图之间，并没有明显的区别。

观察线虫编码以及非编码的可视化效果(如图 10，图 11 所示)，同样存在两类图在 A-T 对称，C-G 对称，且 A-T 比 G-C 在非编码投影上要明显的多。与沙门氏菌不同的是，线虫的非编码区与编码区之间存在着明显的差别，它的非编码区比编码区的分布范围要大的多。

观察黑猩猩编码以及非编码的可视化效果(如图 12，图 13 所示)，黑猩猩的非编码区与编码区之间的差异与线虫相比更显明显，它的非编码区比编码区的分布范围要大的多。另外，虽然两种方法均能显示出两类图之间的区别，但是方法 2(图 11)的效果比方法 1(图 10)更明显。

从基因而言，黑猩猩是高等动物，它在一定程度上和人类的基因存在很大的相似性，线虫是较为高等的生物，它进化层次比较高，而沙门氏菌是很低等的生物，沙门氏菌比线虫和黑猩猩要低等得多。通过对编码区与非编码区进行 DNA 可视化分析，可以看出线虫的非编码区与编码区有明显的区别，非编码区的分布范围更广，黑猩猩与线虫相比差异更明显些，沙门氏菌则看不出什么区别。由此可以估计，越是低等的生物其基因非编码和编码越分不清，越高等的生物其基因非编码区的可视化分布图分布范围越大。

5. 结论

本文通过对编码区以及非编码区的 DNA 序列分组分段处理，对每个分段进行概率测量，经过归一化处理得到归一化测度，通过改变可变参数 k 来进行坐标映射，将所有的选定的 DNA 序列进行图形投影，从而得到了编码以及非编码区的 DNA 特征分布图。通过比较低等生物以及高等生物编码以及非编码区的基因特征分布，对二者之间的关系，从非生物学角度提供了一定的研究价值。DNA 序列的图形表示方法

为研究 DNA 序列提供了重要手段，利用图形表示方法描述基因序列具有直观性和计算简单等优点，这样相对于传统的研究方法，大大缩短研究的进程。

致　谢

感谢云南大学软件学院、云南省软件工程重点实验室信息安全基金及郑智捷博士国家自然科学基金的支持。

参考文献(References)

[1] Kumar, S.S. (2005) Responsibilities in the post genome era: Are we prepared? *Issues in Medical Ethics*, **10**, 150-151.

[2] Bernstein, B.E., Birney, E., Dunham, I., *et al*. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

[3] Pennisi, E. (2012) Genomics. ENCODE project writes eulogy for junk DNA. *Science*, **337**, 1159-1161.

[4] Ecker, J.R., Bickmore, W.A. and Barroso, I. (2012) Genomics: ENCODE explained. *Nature*, **489**, 52-55. http://www.nature.com/nature/journal/v489/n7414/full/489052a.html

[5] Randić, M., Novič, M. and Plavšić, D. (2013) Milestones in graphical bioinformatics. *International Journal of Quantum Chemistry*, **113**, 2413-2446.

[6] Staden, R. and Mclachlan, A.D. (1981) Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research*, **10**, 141-156.

[7] Michel, C.J. (1986) New statistics approach to discriminate between protein coding and non-coding. *Journal of Theoretical Biology*, **120**, 223-236.

[8] 张春霆 (1999) 用几何学方法分析 DNA 序列. 中国科学基金, **3**.

[9] Li, Q.P. and Zheng, Z.J. (2010) Spatial distributions for measures of random sequences using 2D conjugate maps. *Proceedings of Asia-Pacific Youth Conference on Communication* (*APYCC*) (*ISTP*), Kunming, 64-69.

[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**, 548-555.