Software Engineering and Applications 软件工程与应用, 2013, 2, 86-89
http://dx.doi.org/10.12677/sea.2013.24015 Published Online August 2013 (http://www.hanspub.org/journal/sea.html)

Visualization Model of Packet Measurement about DNA Sequences

Qinxian Bu, Zhijie Zheng

School of Software, Yunnan University, Kunming
Email: bqxian@126.com, conjugatesys@gmail.com

Received: Jun. 20th, 2013; revised: Jul. 4th, 2013; accepted: Jul. 16th, 2013

Abstract: The generation and detection of random sequences play an important role in the application of Cryptography. With the successful implementation of human genome project, using the random sequence detection tool to process DNA sequences has a special significance. Based on the natural randomness of DNA sequences, this paper proposes a visualization model to display spatial distribution of measurement features of DNA sequences by using conjugate maps. The model can provide a reference for the in-depth visualization study of DNA sequences on random measurement features.

Keywords: DNA Sequences; Packet Measurement; Conjugate Maps

DNA 密码序列分组测量可视化模型

卜琴仙，郑智捷

云南大学软件学院，昆明
Email: bqxian@126.com, conjugatesys@gmail.com

收稿日期：2013 年 6 月 20 日；修回日期：2013 年 7 月 4 日；录用日期：2013 年 7 月 16 日

摘　要：随机序列生成和检测在密码学应用中发挥着重要作用，随着人类基因组计划的成功实施，应用随机序列检测工具处理 DNA 序列具有特殊意义。利用 DNA 序列本身具有的天然随机性，本文用共轭测量图示方法，提出了一种展现 DNA 序列测量特征空间分布的可视化模型，该模型能为深入解析 DNA 序列随机性测量特征的可视化研究提供参考。

关键词：DNA 序列；分组测量；共轭图

1. 引言

由于随机序列可产生不可预知的数字串[1]，因此被广泛应用于移动通信[2]、密码安全[3]、环境仿真[4]等多种与密码生成与检测相关的领域。而现代 DNA 序列的研究表明：同源染色体上等位基因间的相互分离与非同源染色体上非等位基因间的自由组合，互不干扰，各自独立分配到配子中去，形成这种线状的排列分布是随机的，因此 DNA 序列具有天然的随机特性。

DNA 序列是遗传信息的载体，可以用 A、T、C、G 四个字符组成的字符串来表示。由于 DNA 序列数据庞大，研究人员很难直接得到 DNA 序列的信息，因此需要使用辅助工具来进行研究[5]。借鉴李清平[6]等人针对元胞自动机序列提出的密码技术中检测随机序列的可视化模型，建立了一套展现 DNA 序列分组测量特征分布的可视化模型，为 DNA 序列测量及随机序列的可视化研究提供参考。

2. 模型和方法

本文建立的模型包括 2 个核心模块：测量模块和可视化模块。处理过程是把一段长的 DNA 序列分成等长的若干段，分别计算每段中各个碱基的测度得到测度序列，最后用二维共轭图方法来展示测度序列的分布特征。该模型的工作流程如图 1 所示。

2.1. 测量模块

在这个模块中，用概率统计方法将 DNA 序列转换成测度序列。

模块的输入：长度为 M 的 DNA 序列及每组的长度 m；

模块的输出：M/m 组测度序列；

模块的处理：以 m 个碱基为一组，用归一化的概率统计方法分别计算每组测量参数的测度值，直到将整条 DNA 序列划分统计完为止，处理过程如图 2 所示。

令 $i \in \{1,2,\cdots,i,\cdots M/m\}$ 为 DNA 序列不同组的编号，本文使用的测量参数如表 1 所示。令 $j \in \{\mathrm{A,C,T,G}\}$，则 $N_j(i)$表示在序列的第 i 段中，碱基 j 的总数目；同理 $N_{\mathrm{A+T}}(i)$与 $N_{\mathrm{C+G}}(i)$分别表示在序列的第 i 段中，碱基 A 与碱基 T 的数目之和及碱基 C 与碱基 G 的数目之和。碱基 j 对应的测度值用 $\bar{P}_i(j)$ 来表示，同理碱基 A + T、C + G 对应的测度值分别用 $\bar{P}_i(\mathrm{A+T})$、$\bar{P}_i(\mathrm{C+G})$来表示，各个测度值的计算方法如表 2 所示。

2.2. 可视化模块

不同于 Poincare 方法仅基于一组测度序列构造基

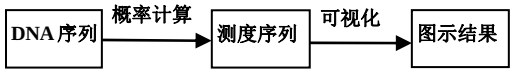

Figure 1. The flow diagram of DNA sequence visualization
图 1. DNA 序列可视化流程图

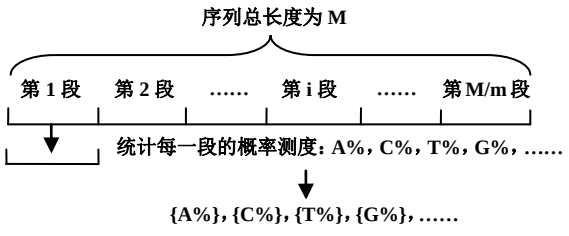

Figure 2. The packet processing diagram of DNA sequences
图 2. DNA 序列分组处理示意图

Table 1. The types of measurement parameters
表 1. 测量参数类型

碱基类型	参数类型	总数目
A	$N_{\mathrm{A}}(i)$	
T	$N_{\mathrm{T}}(i)$	
C	$N_{\mathrm{C}}(i)$	$N = N_{\mathrm{A}}(i) + N_{\mathrm{T}}(i) + N_{\mathrm{C}}(i) + N_{\mathrm{G}}(i)$ $N_0(i) = N_{\mathrm{A+T}}(i) = N_{\mathrm{A}}(i) + N_{\mathrm{T}}(i)$ $N_1(i) = N_{\mathrm{C+G}}(i) = N_{\mathrm{C}}(i) + N_{\mathrm{G}}(i)$ $N = N_0 + N_1$
G	$N_{\mathrm{G}}(i)$	
A + T	$N_{\mathrm{A+T}}(i)$	
C + G	$N_{\mathrm{C+G}}(i)$	

Table 2. The value of probability measurements
表 2. 测量参数的测度值

概率测度	测度值
$\bar{P}_i(\mathrm{A})$	$N_{\mathrm{A}}(i)/N_0(i)$
$\bar{P}_i(\mathrm{T})$	$N_{\mathrm{T}}(i)/N_0(i)$
$\bar{P}_i(\mathrm{C})$	$N_{\mathrm{C}}(i)/N_1(i)$
$\bar{P}_i(\mathrm{G})$	$N_{\mathrm{G}}(i)/N_1(i)$
$\bar{P}_i(\mathrm{A+T})$	$N_{\mathrm{A+T}}(i)/m$
$\bar{P}_i(\mathrm{C+G})$	$N_{\mathrm{C+G}}(i)/m$

于不同距离参数的图示[6]，本文选择两组测度序列来构造二维共轭图示，由于实施的测量方法具有同时给出多组测度序列的能力，因此，本文提出的图示方法能展示更为精细的测量特征分布[6]。表 2 列出该模型的 6 种测量参数，在进行二维可视化时，共有 $C_6^2 = 15$ 种选择，在本文中选择其中的几种组合进行展示。

模块的输入：表 2 中的两个概率测度序列；

模块的输出：相应的图示化结果；

模块的处理：选定将要展示的测度组合，逐一展示这组测度组合在各个段中的分布，直至将所有段即 M/m 段处理完为止，每段中的一组测度组合对应图示中的一个点。这样就得到了整条序列的测量特征分布。

通过可视化模块的工作，最终形成了一系列图例组合。这些结果能直观地显示 DNA 序列测量特征的空间分布信息，为深入研究 DNA 序列以及随机数的特征空间分布提供方便。

3. 可视化结果及分析

3.1. 可视化结果

本文选用两组 DNA 序列作为实例：一组是水稻

属种基因组中从 224150~224326 共 176 个碱基组成的 DNA 序列片段[7]；另一组是玉米基因组中一段具有转位特性的长度为 376 bp 的独立 DNA 序列[8]。为了更好地分析分组长度对 DNA 序列可视化结果的影响，选用 $\{\bar{P}_i(\mathrm{A}),\bar{P}_i(\mathrm{G})\}$、$\{\bar{P}_i(\mathrm{A}),\bar{P}_i(\mathrm{A}+\mathrm{T})\}$ 两个测度组合的结果图示进行比较。同时考虑到图示结果的易读性，每个图均有图名，命名规则是：“序列分组长度 + 序列简写名称 + 基础序列概率统计计算方法 + 可视化的维度模型”。DNA 序列简写名称与原始名称的对应关系如表 3 所示，实例的可视化结果如图 3 和图 4 所示。

3.2. 结果分析

可视化结果显示：1) 随着分组长度的增大，相对应的散点数会减少，并且点与点之间越呈聚集的趋势；2) 随着分组长度的增大，不同序列的空间分布越具有差异性。

以图 3 为例，当分组长度分别为 9、15、30 时：水稻序列在该模型下得到的空间点数分别为 17、11、5，并且横坐标的值域从 0.20~1.00 变为 0.35~0.64，纵坐标的值域从 0.00~1.00 变为 0.44~0.82；玉米序列在该模型下得到的空间点数分别为 32、25、12，横坐标的值域从 0.00~1.00 变为 0.12~0.70，纵坐标的值域从 0.00~1.00 变为 0.15~0.83。

随着分组长度的增大，横纵坐标的值域范围缩小，两个子序列完全相同的可能性减少，导致坐标中点与点之间完全重合的可能性减小，这时空间点的分布是 DNA 序列信息的完全表达，此时易于进行序列的差异性分析。

当分组长度为 30 时，比较图 3 中水稻序列和玉米序列的点分布，可以得到：水稻序列中碱基 G 的含量高于玉米序列中的含量，这为进一步分析水稻序列和玉米序列的性状差异原因提供依据。

通过以上分析可知：概率计算方法的特点以及不同DNA序列本身的属性差异形成了图示的分布特征。从测量可视化的角度，在设定分组长度的时候，需要综合考虑信息丢失以及分组的实用性两个因素来得到适合的描述参数。

Table 3. The relationship between abbreviated name and original name of DNA sequences
表 3. DNA 序列的简写名称与原始序列的对应关系

简写名称	所代表的序列	GenBank 索引号	物种名称
OR2402	ORSiTEMT02400002	6979318	水稻
ZR5911	ZRSiTEMT05900011	16225215	玉米

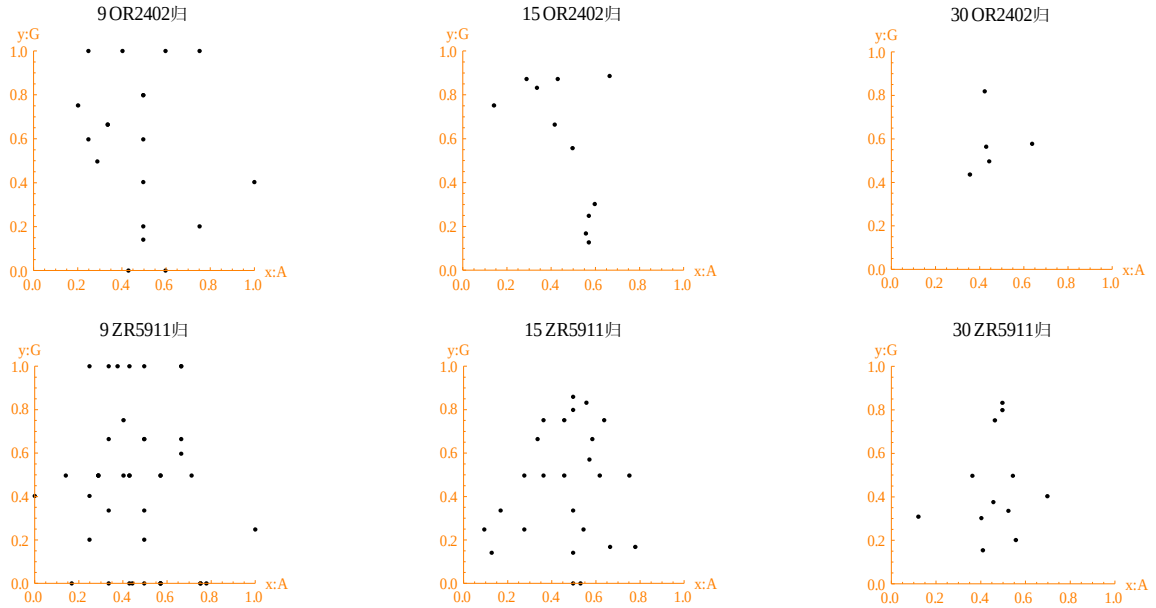

Figure 3. The results of different length by using the measurement combination of $\{\bar{P}_i(A),\bar{P}_i(G)\}$: the first represents oryza sequence, the second line represents maize sequence
图 3. $\{\bar{P}_i(A),\bar{P}_i(G)\}$ 测度组合不同分组长度的可视化结果：其中第一行表征水稻序列，第二行表征玉米序列

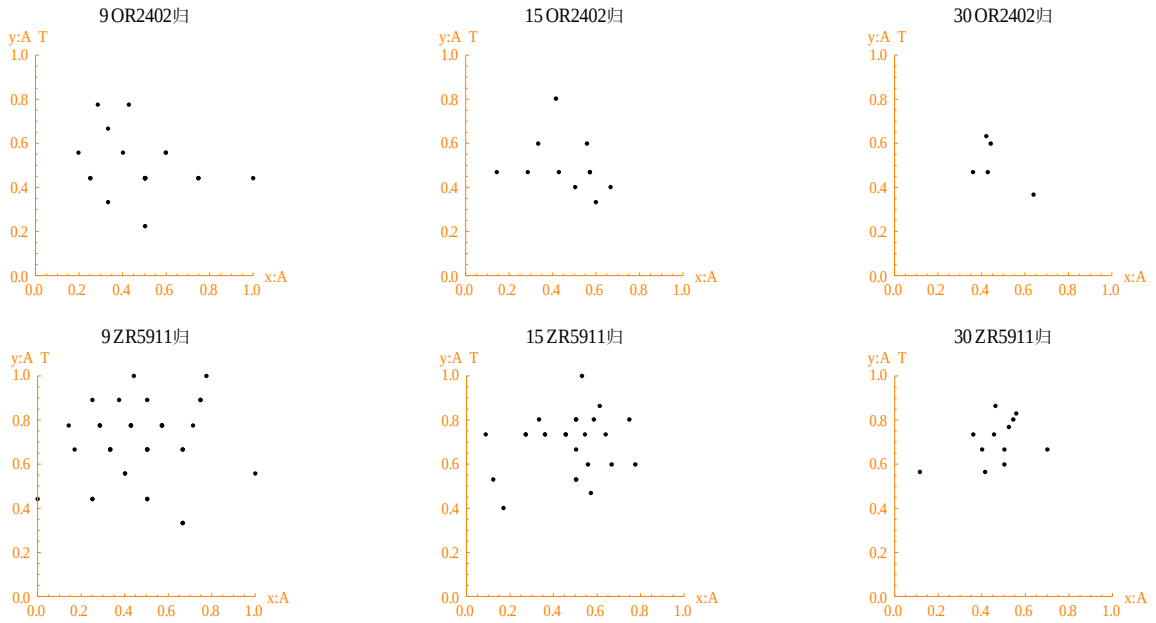

Figure 4. The results of different length by using the measurement combination of $\{\bar{P}_i(A), \bar{P}_i(A+T)\}$: the first represents oryza sequences, the second line represents maize sequence

图 4. $\{\bar{P}_i(A), \bar{P}_i(A+T)\}$ 测度组合不同分组长度的可视化结果：其中第一行表征水稻序列，第二行表征玉米序列

4. 总结

本文建立了用于展现 DNA 序列分组测量特征分布的可视化模型，一方面能为 DNA 序列的分组研究提供参考，达到了简化不同 DNA 序列差异性分析的目的，另一方面能为随机密码序列的可视化研究提供借鉴。进一步的工作将集中在以下两个方面：1) 优化算法，使可视化结果尽可能地展示 DNA 序列的精细特性；2) 以具体应用为切入点，运用建立的模型解决实际问题。

5. 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云南省海外高层次人才项目对本课题的支持。

参考文献 (References)

[1] 梁帆, 张秀龙, 郑智捷. 利用随机性测试方法对特征分布图示的分类和评判[A]. 2010 年亚太青年通信与技术学术会议论文集[C]. 昆明, 2010: P78-P82.

[2] 陈顺林, 杨万全, 董庆蓉. m 序列在移动通信扰码中的应用与仿真[J]. 现代电子技术, 2002, 3: 27-29.

[3] B. Schneier. Secrets & lies: Digital security in networked world. John Wiley & Sons, Hoboken, 2000: 85-101.

[4] 杨睿. 论伪随机序列及其应用[J]. 沈阳工程学院学报(自然科学报), 2009, 5(2): 166-168.

[5] 石龙. 一种 DNA 序列的 2D 图形表示[J]. 科技信息, 2009, 1: 480-483.

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

[7] R. Tarchini, P. Biddle, R. Wineland, S. Tingey and A. Rafalski. The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. Plant Cell, 2000, 12(3): 381-391.

[8] X. Zhang, C. Feschotte, Q. Zhang, N. Jiang, W. B. Eggleston and S. R. Wessler. P instability factor: An active maize transposon system associated with the amplification of tourist-like MITEs and a new super family of transposases. Proceedings of the National Academy of Sciences of the United States of America, 2011, 98(22): 12572-12577.

