Hans Journal of Computational Biology 计算生物学, 2019, 9(2), 14-21
Published Online June 2019 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2019.92003

Hans汉斯

The Visual Analysis of Packet Measurement about Tobacco DNA Sequence

Hengyi Qu, Xin Yuan, Linwei Ma, Zhijie Zheng

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

Received: May 24th, 2019; accepted: Jun. 7th, 2019; published: Jun. 14th, 2019

Abstract

With the development of genetic engineering technology, the gene database is expanding, and the methods of DNA data processing are improving. In this paper, the complex tobacco DNA data are projected into a visualized graph by using the variable map technique. And the data are projected separately using different segments. These results provide a reference for further tobacco genome research.

Keywords

Tobacco DNA, Visualization, Variable Map

烟草DNA序列分组测量可视化分析

曲恒熠，袁　鑫，马林威，郑智捷

云南大学软件学院，云南　昆明
Email: hengyiqu@qq.com

收稿日期：2019年5月24日；录用日期：2019年6月7日；发布日期：2019年6月14日

摘　要

随着基因工程技术的发展，基因数据库也在不断扩大，DNA数据的处理研究的方法也在不断改进。本文则是利用变值图技术，将复杂的烟草DNA数据投射成可视化的图，并且对数据使用不同的分段分别进行投射。这些分析和展示的结果为深入烟草基因组研究提供了参考。

关键词

烟草DNA，可视化，变值图

文章引用: 曲恒熠, 袁鑫, 马林威, 郑智捷. 烟草 DNA 序列分组测量可视化分析[J]. 计算生物学, 2019, 9(2): 14-21.
DOI: 10.12677/hjcb.2019.92003

1. 引言

在现代生物学研究中，DNA 序列正在以大数据流的形式对广泛的物种(从单细胞到人类)进行测序。而从整个基因序列中对不同种类的 DNA 序列进行分类和鉴定是一项困难的工作。目前基因组研究的方法是通过对 DNA 序列进行多角度、多层次的处理和分析，获取更多的生物信息。近年来，生物基因数据的处理和利用以多种处理方式进行[1] [2] [3] [4]，如基因特征提取、基因序列定位等。

变值图是一种新型处理方式，以四种符号为元结构来处理密码序列、DNA 序列到心电信号的随机序列的技术。该类方法基于概率值，适用于将随机序列的统计分析方法应用到基因序列，利用类推法从所选序列中生成多个统计概率分布，形成二维－三维的可视化图，可视化效果良好，映射结果可用于探索全基因组的非线性复杂行为[5] [6] [7] [8] [9]。

烟草是最早应用于基因工程研究的植物之一[10]，但在早期，其育种的速度慢于其他植物，遗传图谱的构建也远落后于其他植物，原因之一是其基因序列具有高度的复杂性和特殊性，导致基于基因数据的研究困难重重。

在本文中，将使用变值图把从烟草基因序列中得到的一系列基因序列数据[11] [12]经过处理、投射成映射结果，以可视化的结果展示烟草基因组的复杂行为。

2. 系统架构

2.1. 架构

过程模块的架构示意如图 1 所示。处理模型由五部分构成：输入，处理，测量，投影和输出，其中包含三个模块：处理，测量和投影。

输入：烟草 DNA 序列。

输出：二维图像。

模块：在处理模块中，我们将从 Ensembl Plants 基因库下载的烟草 DNA 序列[11] [12]以固定长度 m 连续划分为多个 DNA 段。在测量模块中，计算每个段{A, C, G, T}四种符号的数量，并将测量段转为四个测量序列。在投影模块中，X:{AT}和 Y:{AG}的特定结合决定了在投影位置的四种测量序列，整个测量序列最终投影为二维图像。

2.2. 处理模块

输入的 DNA 序列中，多个段能够以固定长度 m 划分来生成一个段的序列。

输入：一条 DNA 序列。

输出：一条段序列。

2.3. 投影模块

投影模块图示如图 2 所示，其包含了位置与投影两个部分。对于每一种测量，X{AT}与 Y{AG}独自决定了两条轴心线。当所有测量完成后，二维图像的统计分布建立了二维柱形图。

输入：四个测量序列。

输出：二维图像。

Figure 1. Process module
图 1. 过程模块

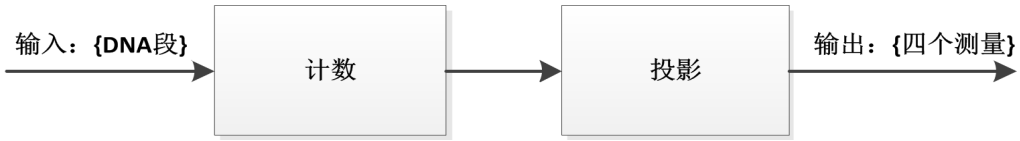


Figure 2. Projection module
图 2. 投影模块

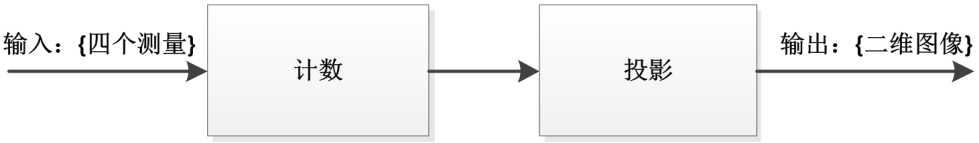


Figure 3. Measuring module
图 3. 测量模块

2.4. 测量模块

测量模块图示如图 3 所示。单独地测量每段 DNA 四种符号{A, G, C, T}的比例情况。最终每种符号的数量是 0 到 m 之间的整数，将一个段序列转换为四个测量序列。

输入：一条段序列。

输出：四个测量序列。

3. 可视化结果及分析

3.1. 参数解释

m：每组处理 DNA 序列的长度(这里我们分别选取了 m = 80，100，120，140，160，180，200，220，240，260 段长度做可视化处理)；

V 为 AT，AG 两个碱基组中的一个，V∈{AT, AG}；

P_V 为碱基组的比例；

Result1、Result2：分别保存 DNA 序列中 AT 数量、AG 数量的数组；

Result1 = NUM(A) + NUM(T)；

Result2 = NUM(A) + NUM(G)；

Result1(x)：表示第 x 组中 A 的个数与 T 的个数之和；

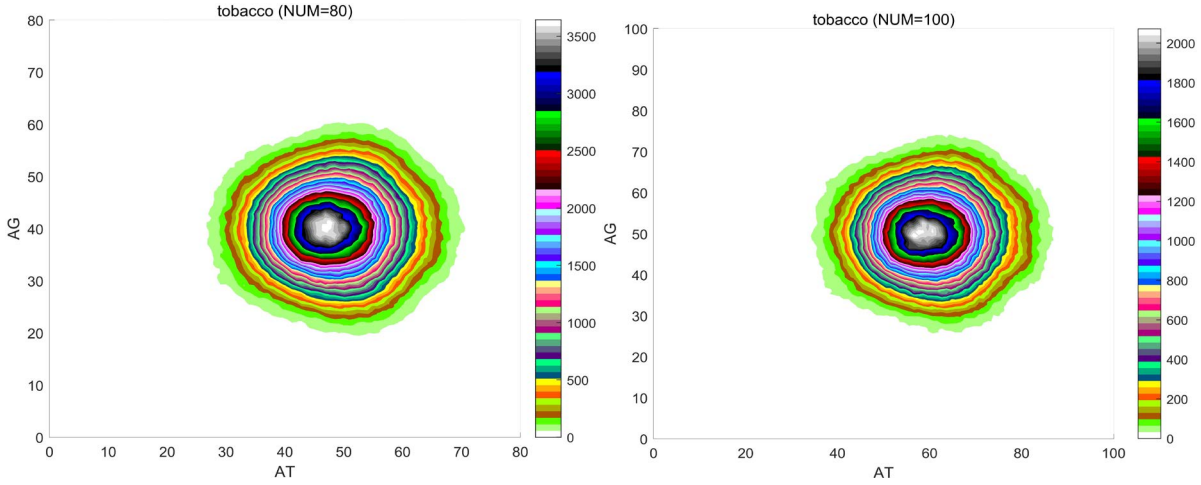
Result2(y)：表示第 y 组中 A 的个数与 G 的个数之和；

(P_{AT}, P_{AG}) = (Result1(x), Result2(y))映射生成图像上的点。

3.2. 不同分段的投影图像

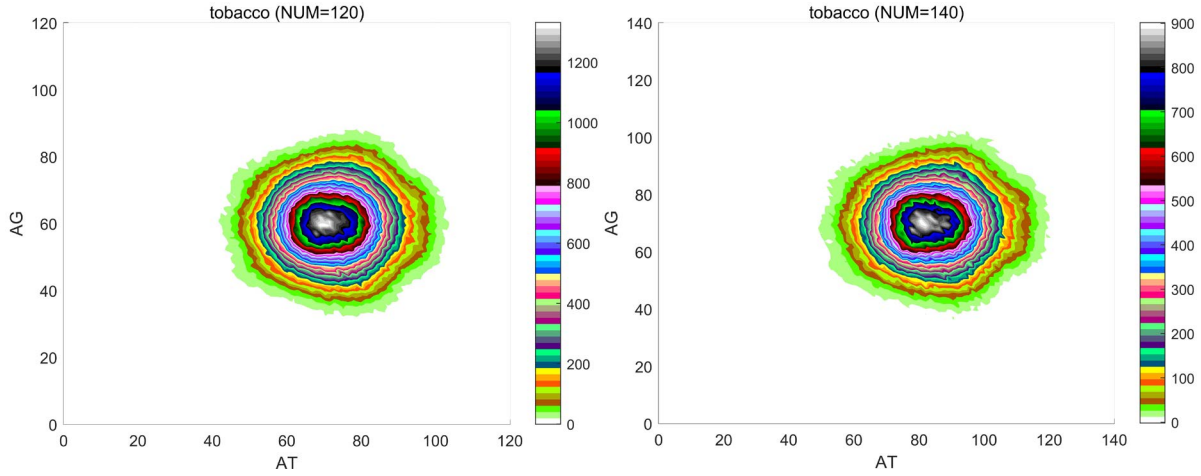
图中分别显示了当 m = {80, 100, 120, 140, 160, 180, 200, 220, 240, 260}不同分组长度时的图像，在图

像中颜色的相似的位置表示重叠相差不多的投影数量，重叠投影数量由彩色图中心处向四周递减。如图4 不同分组长度的彩色图(m = 20~260)所示。



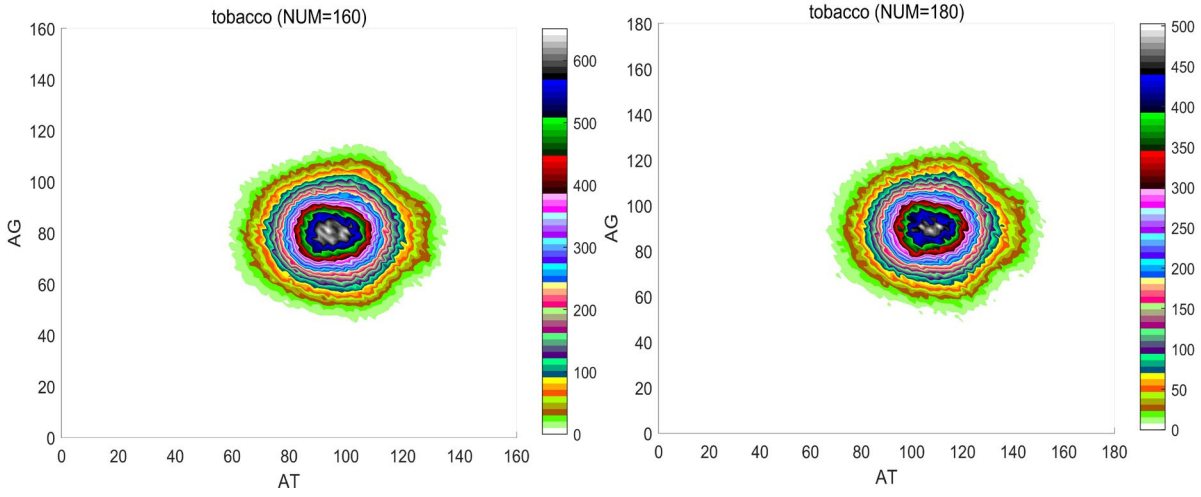
(a) m = 80

(b) m = 100



(c) m = 120

(d) m = 140



(e) m = 160

(f) m = 180

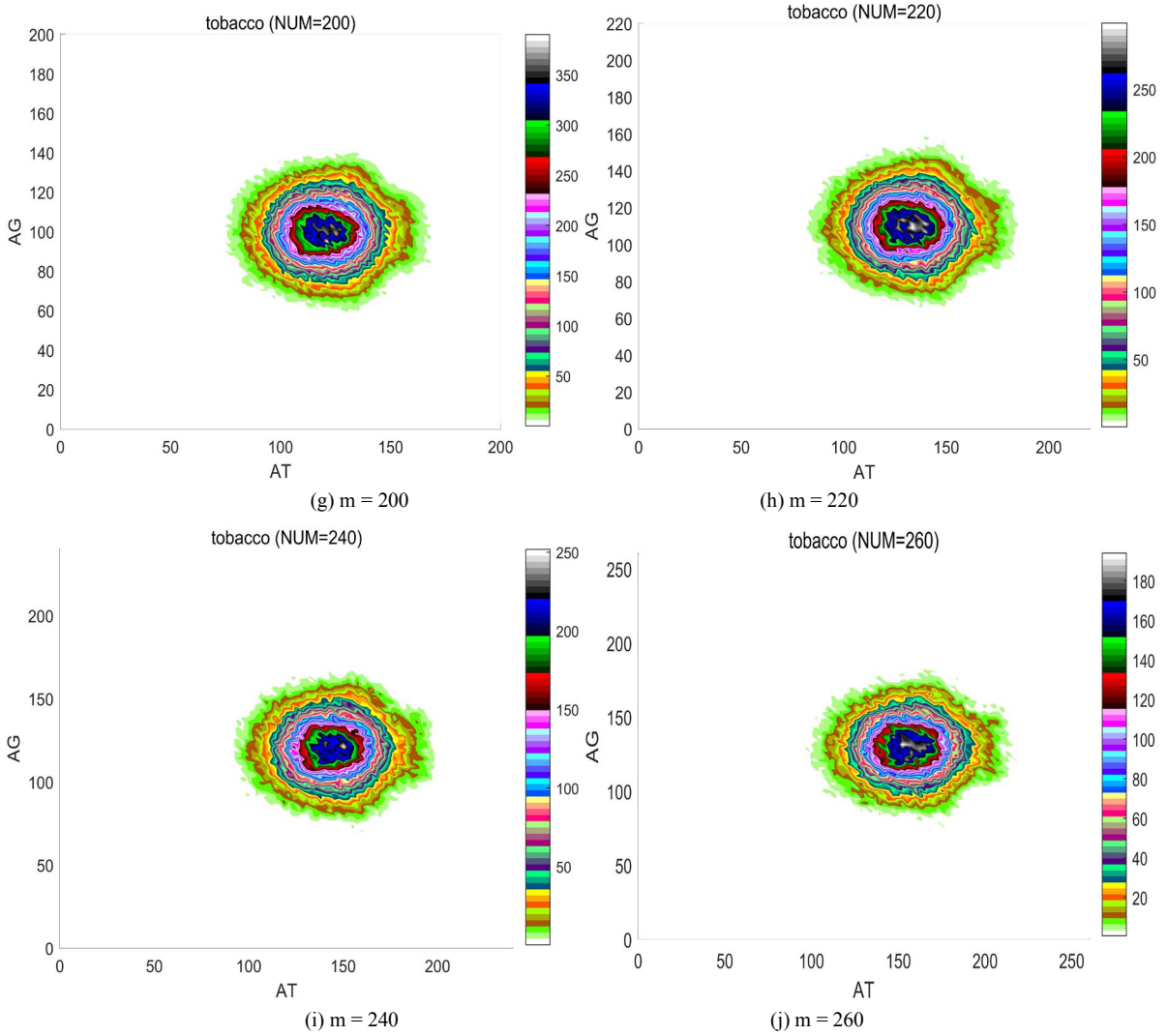


Figure 4. Color maps of different group lengths (m = 80~260)
图 4. 不同分组长度的彩色图(m = 80~260)

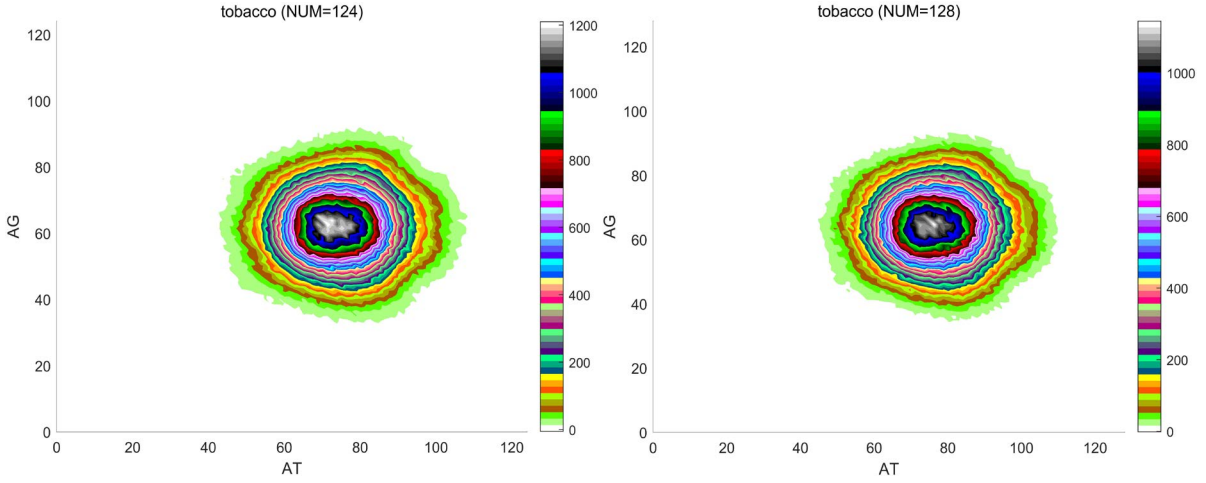
3.3. 简要分析

生成图像的质量取决于图像中投影的数量，因此质量较好的图像需要包含大量的投影点。相同投影形成叠加，在生成的彩色图像中，相同地方叠加的投影点越多则该地方的颜色就越鲜艳。包含大量投影点，则会生成颜色分明的彩色图。这里我们使用的是烟草的 1 号颜色体中的 DNA 序列，1 号染色体有大量的 DNA 序列，从而得到了大量的投影点和漂亮的投影结果。当 m 的值为 120 到 160，所形成的图像右边凸出明显，并且图像效果最佳。如图 5 不同分组长度的彩色图(m = 124~160)所示，我们对 m = {124, 128, 132, 136, 140, 144, 148, 152, 156, 160}进行精细的划分来观看图像的变化，发现图像之间仅存在一些细微的差距，但所有的图像都保留了右边凸出的特点，烟草的 DNA 序列的一部分特性在投影图上得到体现。

4. 总结

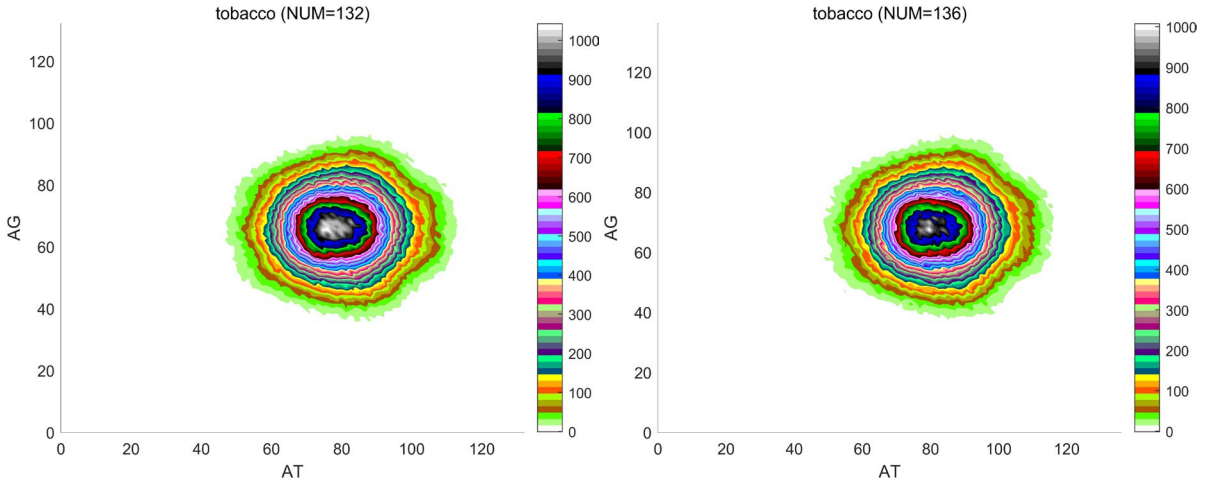
本文通过对烟草 DNA 序列分组分段处理，利用相关的映射方案，将烟草的 1 号染色体中的完整的 DNA 序列转换为具有显著视觉特征的彩色图。除了投影方法和 DNA 序列处理方法之外，还包括通过调

整段长度 m 来进行投影，从而形成不同的投影分布。通过制作的彩色图以及结果，可以发现一些有意义的信息。比如 AG 以 AT 在烟草 DNA 序列中的分布。通过彩色图表示方法描述基因序列的分布具有直观



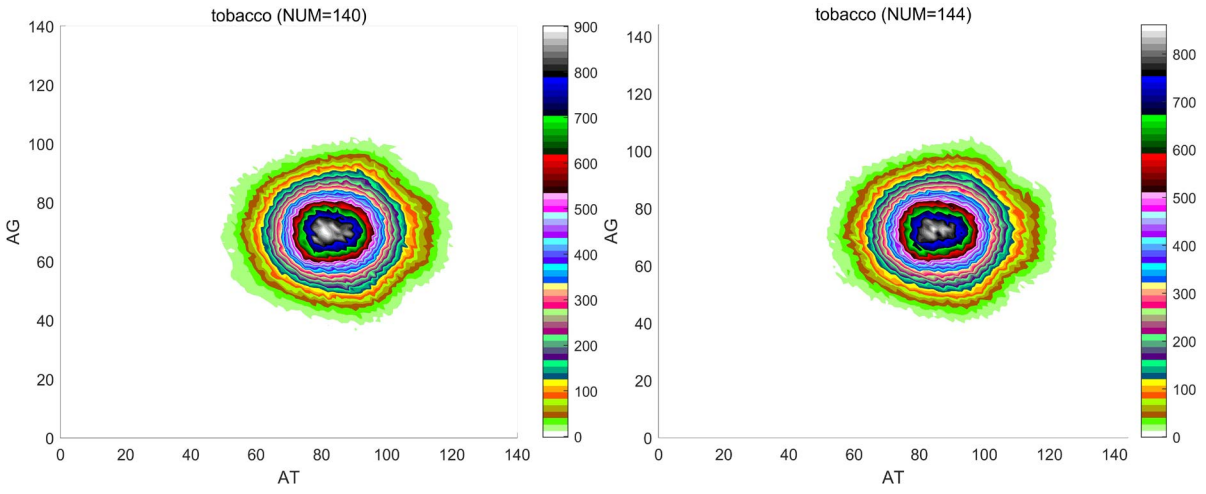
(a) m = 124

(b) m = 128



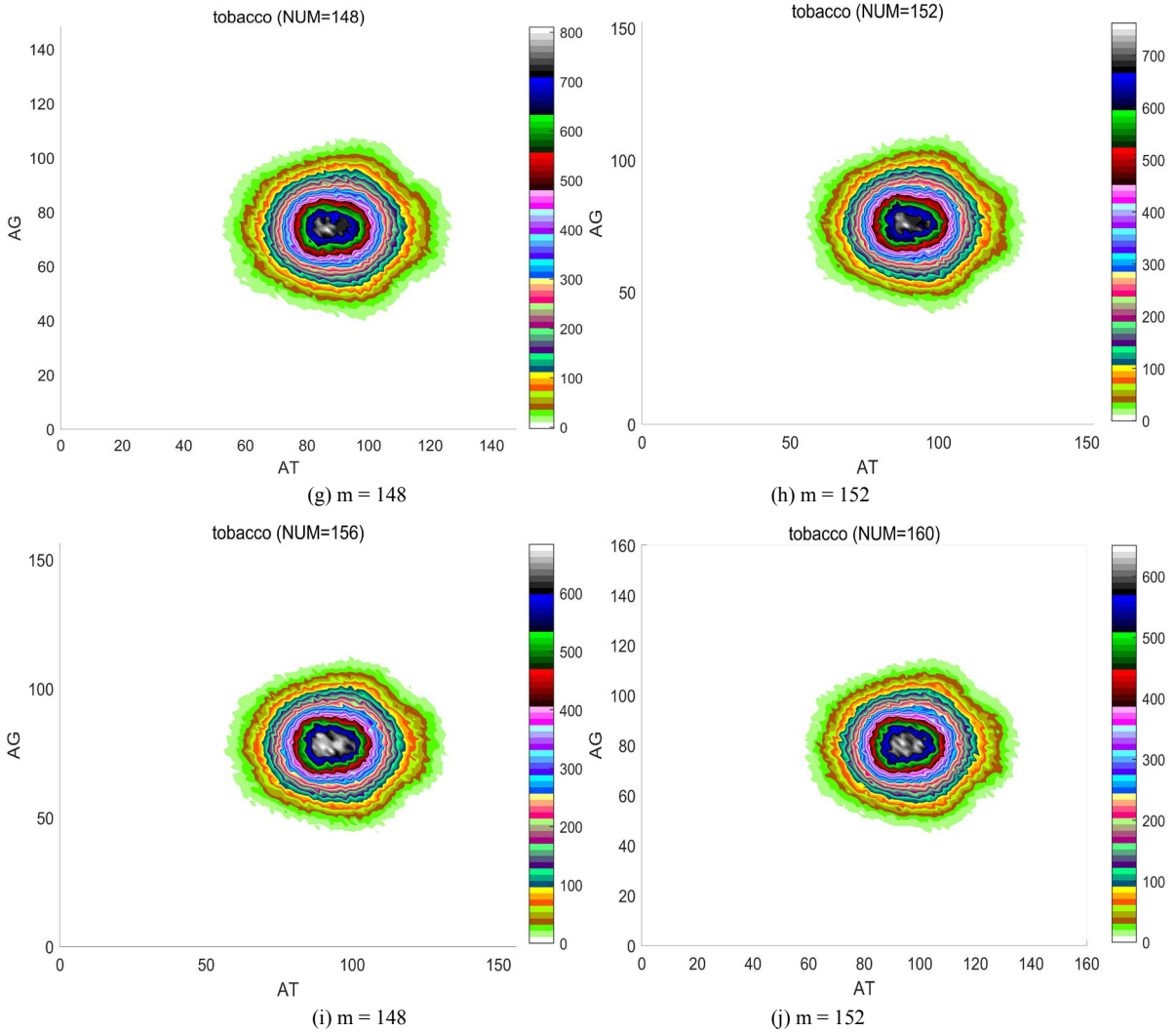
(c) m = 132

(d) m = 136



(e) m = 140

(f) m = 144



(g) m = 148 (h) m = 152

(i) m = 148 (j) m = 152

Figure 5. Color maps of different group lengths (m = 124~160)
图 5. 不同分组长度的彩色图(m = 124~160)

性和计算简单等优点，相较于传统的研究方法，缩短研究的进程，可以为生物信息、生命科学等方面提供一定的研究基础。

致　谢

感谢郑智捷教授对这篇文章的细心指导，感谢云南大学软件学院、云南省软件工程重点实验室的支持。

参考文献

[1] Berger, J.A., Mitra, S.K., Carli, M. and Neri, A. (2004) Visualization and Analysis of DNA Sequences Using DNA Walks. *Journal of the Franklin Institute*, **341**, 37-53. https://doi.org/10.1016/j.jfranklin.2003.12.002

[2] Pitt, J.N., Rajapakse, I. and Ferre-D'Amare, A.R. (2010) SEWAL: An Open-Source Platform for Next-Generation Sequence Analysis and Visualization. *PMC*, **38**, 7908-7915. https://doi.org/10.1093/nar/gkq661

[3] Bernstein, B.E., Birney, E., Dunham, I., *et al*. (2012) An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*, **489**, 57-74. https://doi.org/10.1038/nature11247

[4] Pennisi, E. (2012) Genomics. ENCODE Project Writes Eulogy for Junk DNA. *Science*, **337**, 1159-1161. https://doi.org/10.1126/science.337.6099.1159

[5] 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*, Kunming, 64-69.

[6] 张巍琼, 郑智捷. 基于不同产生机制的伪随机序列和 DNA 序列的随机性测量[J]. 成都信息工程学院学报, 2012, 27(6): 548-555.

[7] 刘玉倩, 郑智捷. 编码和非编码 DNA 序列的可视化分析[J]. 计算生物学, 2014, 4(2): 20-31.

[8] 完竹, 郑智捷. DNA 序列一维分段测量分布可视化[J]. 云南大学学报(自然科学版), 2013(35): 1-6.

[9] 杜磊, 郑智捷. 在非线性函数下的 DNA 概率测量聚类分布[J]. 软件工程与应用, 2014, 3(3): 41-49.

[10] 李彦平, 丁燕芳, 孙焕, 李雪君, 朱景伟, 段旺军, 马浩波, 郭芳阳. 我国烟草育种现状及思考[J]. 河南农业科学, 2010(9): 148-150+156.

[11] http://plants.ensembl.org/Nicotiana_attenuata/Info/Index

[12] Xu, S., Brockmöller, T., Navarro-Quezada, A., *et al*. (2017) Wild Tobacco Genomes Reveal the Evolution of Nicotine Biosynthesis. *Proceedings of the National Academy of Sciences*, **114**, 6133-6138. https://doi.org/10.1073/pnas.1700073114