Hans Journal of Computational Biology 计算生物学, 2019, 9(3), 39-47
Published Online September 2019 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2019.93006

Hans汉斯

The Visual Analysis of Variables Statistic Based on Arabidopsis DNA Sequences

Yingping Zhou, Jia Luo, Xin Yuan, Guochen Qiu, Hengyi Qu, Zhijie Zheng

School of Software, Yunnan University, Kunming Yunnan
Email: yingpingzhou@126.com

Received: Aug. 14th, 2019; accepted: Aug. 28th, 2019; published: Sep. 4th, 2019

Abstract

With the development of plant genome sequencing, Arabidopsis gene sequencing has received great attention from all over the world for its great scientific research value. At the same time, the research on the sequence of Arabidopsis gene has been greatly advanced. Because of the specific and complex of traditional genetic representations, we try to use a more intuitive and universal genetic variable mapping system. Using this model, after preprocessing the DNA sequence obtained from the gene pool, we count the number of bases A, C, G, and T and calculate the number of AG, and AT. Then the quantity is projected into a two-dimensional or three-dimensional image which can observe the characteristics of the DNA sequence of Arabidopsis. As can be seen from the series of diagrams given, the new statistical distribution results provide basic information for subsequent similarity analysis of gene sequences and higher dimensional studies.

Keywords

Arabidopsis, DNA Sequence, Variable Diagram, Visualization

基于拟南芥DNA序列的变值统计可视化分析

周影平，罗　佳，袁　鑫，邱国宸，曲恒熠，郑智捷

云南大学软件学院，云南　昆明
Email: yingpingzhou@126.com

收稿日期：2019年8月14日；录用日期：2019年8月28日；发布日期：2019年9月4日

摘　要

随着植物基因组测序的展开，拟南芥基因测序因其巨大的科研价值得到了世界各国的重视。拟南芥基因

文章引用: 周影平, 罗佳, 袁鑫, 邱国宸, 曲恒熠, 郑智捷. 基于拟南芥 DNA 序列的变值统计可视化分析[J]. 计算生物学, 2019, 9(3): 39-47. DOI: 10.12677/hjcb.2019.93006

序列的研究得到了极大的进展。由于传统的基因图示非常特殊和复杂，我们试图利用一种更为直观和普适的基因变值图示系统。利用该类模式，对从基因库获取到的DNA序列进行预处理后，统计碱基A，C，G，T的数量，计算AG，AT的数量，然后将数量投影到二维或者三维图像中，从而观察拟南芥的DNA序列的特征。从给出的系列图示可以看到，新的统计分布结果对后续为基因序列的相似性分析和更高维的特征研究提供基础信息。

关键词

拟南芥，DNA序列，变值图示，可视化

1. 引言

拟南芥，学名 Arabidopsis。拟南芥是一种实验友好的植物，遗传背景简单清晰，在分子生物学的研究上有着很大的优势。植物基因研究人员们将其视为植物中的“果蝇”。基因序列的测序工作在各国工程组的努力下正在逐渐完善，更多生物种类的 DNA，RNA 以及蛋白质的序列正在填充各个基因库环境下，拟南芥基因组全序列在 2000 年底完全测定并公开发表。拟南芥是第一种完成全基因组测序的有花植物、第三种完成测序的多细胞生物、第四种完成测序的真核生物，而人类基因组测序在 2001 年第一次完成草图，2003 年完成测序的最终版本，至今尚未完成真正的全基因组测序。在拟南芥之前完成测序的真核生物，只有酿酒酵母(1996 年)、秀丽隐杆线虫(1998 年)和果蝇(2000 年)。拟南芥的基因组由 5 对染色体组成，总大小只有 1.35 亿碱基对；作为对比，人类有 23 对染色体，32 亿碱基对，是拟南芥的 23.7 倍。从中可以窥见，拟南芥在植物基因序列研究有着极大的科研价值。

随着植物基因组测序的展开，拟南芥基因测序因其巨大的科研价值得到了世界各国的重视，拟南芥基因的序列的研究得到了极大的进展。由于传统的基因图示非常特殊和复杂，本文试图利用一种更为直观和普适的基因变值图示系统给出基因图示。利用该类模式，对从基因库获取到的 DNA 序列进行预处理后，统计碱基 A，C，G，T 的数量，计算 AG，AT 的数量，然后将数量投影到二维或者三维图像中，从而观察拟南芥的 DNA 序列的特征[1]。从给出的系列图示可以看到，新的统计分布结果为后续的基因序列的相似性分析和更高维的特征研究提供基础信息。

Hamori 和 Ruskin 提出了 G 曲线和 H 曲线表示方法[2]。G 曲线是在 5 维空间下，分别以 A、C、G、T 四种核苷酸以及 DNA 序列中核苷酸的位置为坐标。H 曲线在 G 曲线的基础上，为了便于人们理解，将五维降为 3 维。为了更直观的表示，在此之后，Gates，Nandy，Leong 和 Morgenthaler 分别独立发现了 DNA 序列有退化性的 2 维图形表示方法[3] [4] [5]。为了克服 2 维图示较高的退化性，Xiaofeng Guo [6]等人提出了较低退化的 2 维表示方法。为了提供一种简单直接的图形方法，既可以消除退化性又可以将序列展示出来，Yonghui Wu [7]等人提出了 DB-曲线。DB-曲线在平面上一次表示 2 个核酸基的性质。但是 DB-曲线不能保证任意一条 DNA 序列和它的图形表示之间的映射是一一对应的。而为了克服一一对应关系，Stephen S. T. Yau [8]提出了一种新的 2 维 DNA 序列图形表示方法，但是计算冗杂。而本文的变值图示在满足了消除退化性和一一对应的基础上，更为直观和便于计算。

2. 模型处理

2.1. 相关介绍

染色体的主要化学成分和组成基因的材料是脱氧核糖核酸(DNA)编码遗传信息的生物大分子。DNA的结构是由一对多核苷酸链相互盘绕组成的双螺旋。DNA 序列组成的四种核苷酸分别是腺嘌呤(A)，鸟漂呤(G)，胞嘧啶(C)，胸腺嘧啶(T)，这四种核苷酸无间隔的排列在一起构成序列。任意长度大于 4 的一串核苷酸被称作一个序列。近二十年来，序列的图形表示方法不断发展和演变，在研究序列局部和整体的比较分析中起到了很大的作用。序列图形表示方法的主要思想是用 4 个向量去表示 DNA 的 4 种核苷酸，将其映射成 3 维空间或 2 维平面上的曲线图形，由此提取序列的数值特征，从而用于 DNA 序列的相似性分析。

2.2. 研究模型和方法

模型处理过程如图 1 所示：

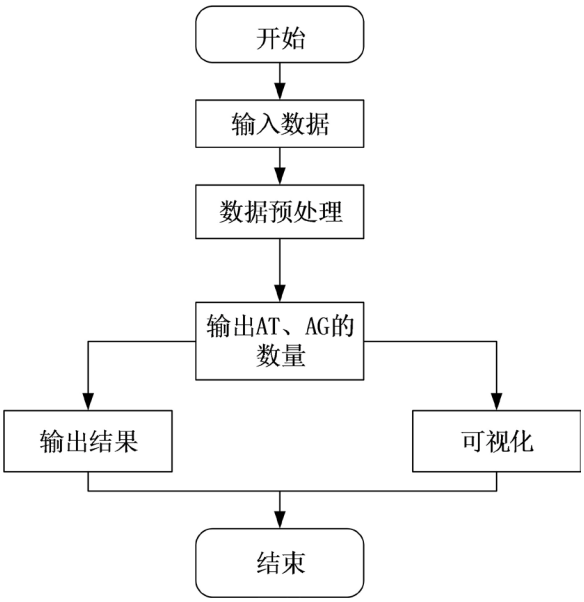

Figure 1. Processing flow
图 1. 处理流程

方法：

输入：从 TAIR 官网[9]获取 TAIR10_genome_release 版本的拟南芥 1 号染色体前 36 万个碱基的 DNA 序列。

数据预处理：首先去除基因序列文件中的信息标注；然后将文件格式转换为需要的数据格式，本文将其转换为纯文本文件格式；最后根据需要决定是否拆分数据文件，本文根据需要拆分得到前 36 万个碱基序列文件。

数据统计：分段处理和统计基因序列，分别以 50，60，70，80，90，100，110，120，130，140，150，160 和 170 为数据长度单位对 DNA 序列中的四个碱基(A, C, G, T)数量进行统计，并计算得到 AT 和 AG 的数量，作为可视化模块的输入。

输出：分段统计的 DNA 序列每段中四个碱基 A、G、C、T 的数量以及 AT 和 AG 的数量。

可视化：利用 MATLAB，根据统计的数据进行绘图。

2.3. 变量说明

NUM：数据分段长度，本文中的分段长度采用距离为 10 的十组长度，其中 NUM $\propto$ {50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170}。

V：两个碱基数据集的集合：V $\propto$ {AT, AG}。

其中：AT 表示 AT 的数量，AT 的数量 = A 的数量 + T 的数量。

AG 表示 AG 的数量，AG 的数量 = A 的数量 + G 的数量。

Pv：两个碱基数据集的比例。

其中：P(X, Y) $\propto$ P(AT, AG)，X，Y 分别为 X 轴和 Y 轴的坐标，通过映射生成二维图形上的点。

3. 图示结果

3.1. 图示生成流程

1) 数据预处理：将下载得到的汇编编译文件转换为纯文本格式，去掉文件中的信息标注。

2) 分段并统计：对应段落中的 AGCT 的数量，分别统计 AG，AT 的数量，以此作为 X、Y 轴的坐标。其中分段的长度分别为 NUM。

3) 统计数据点的深度作为 Z 轴的坐标。

4) 运用 MATLAB 中的绘图函数绘图。

3.2. 图示结果

图示结果分别如下图 2~14 所示：

3.3. 简要分析

1) 当以 50 为统计长度单位时，AT 和 AG 的数量主要分布于 30 到 40 区域，数据特征最为明显。

2) 当以 100 为统计长度单位时，AT 和 AG 的数量主要分布于 70 到 90 区域，数据特征相对明显。

3) 当以 150 为统计长度单位时，AT 和 AG 的数量主要分布于 80 到 120 区域，数据特征开始离散化，但较长度为 200 生成的图仍相对明显。

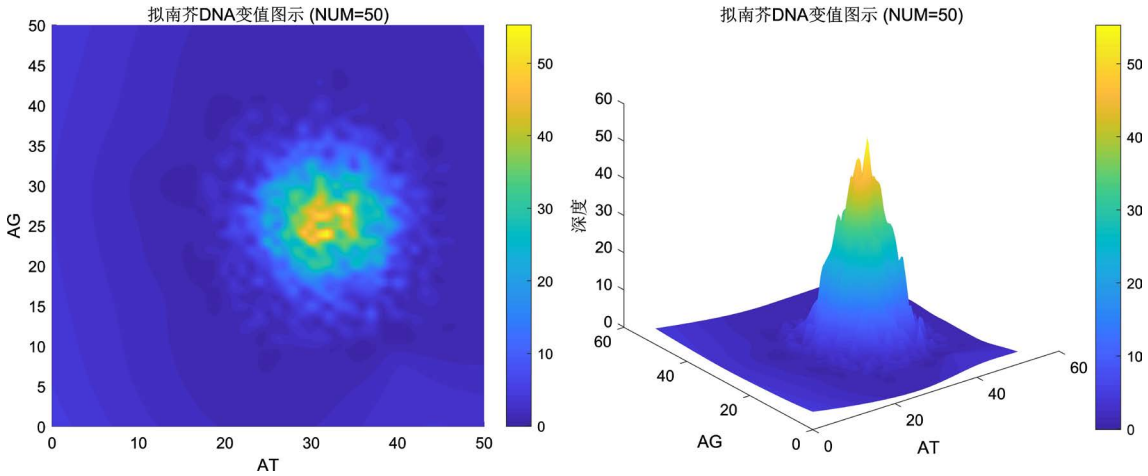

Figure 2. Calculate with 50 as the data length unit
图 2. 以 50 为数据长度单位统计

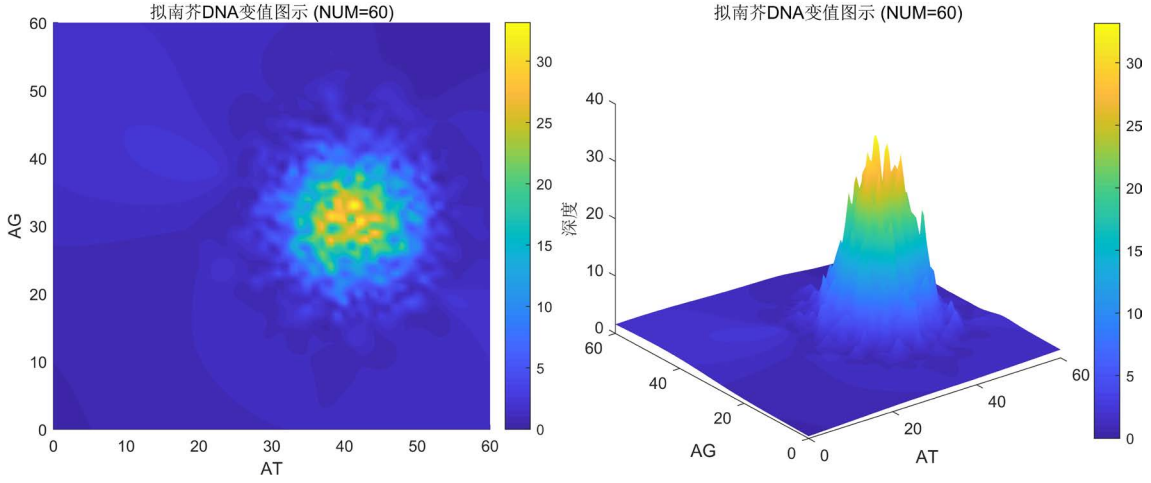

Figure 3. Calculate with 60 as the data length unit
图 3. 以 60 为数据长度单位统计

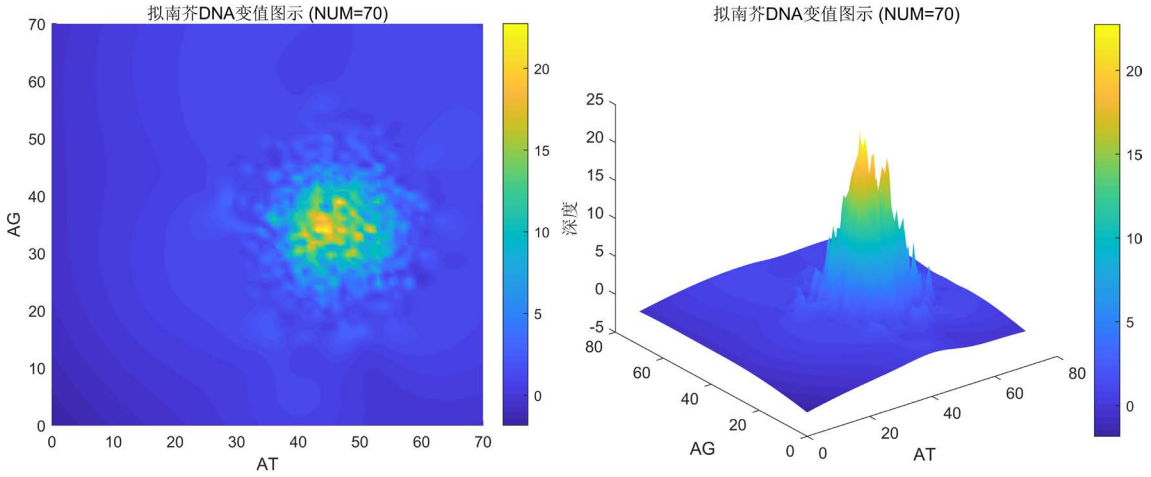

Figure 4. Calculate with 70 as the data length unit
图 4. 以 70 为数据长度单位统计

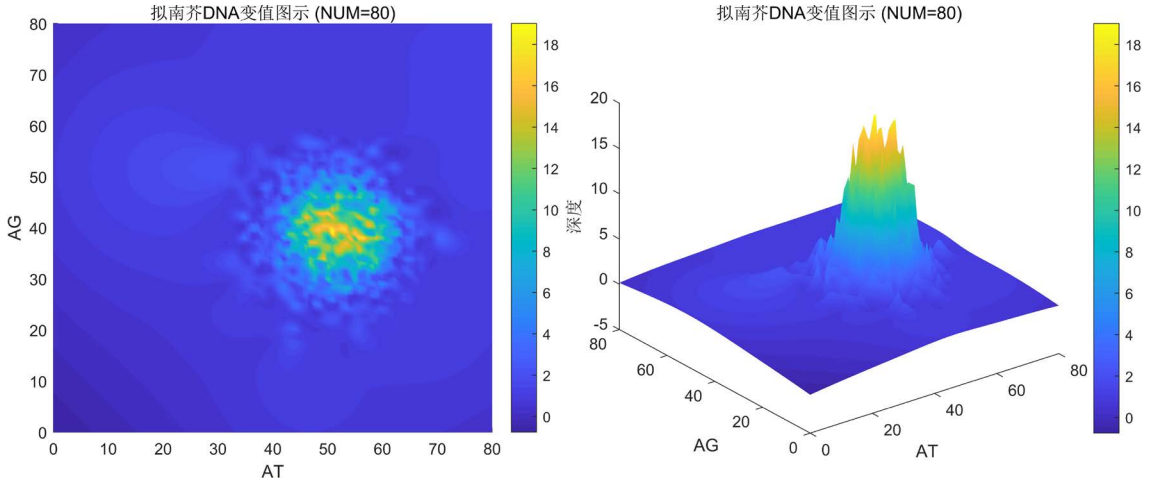

Figure 5. Calculate with 80 as the data length unit
图 5. 以 80 为数据长度单位统计

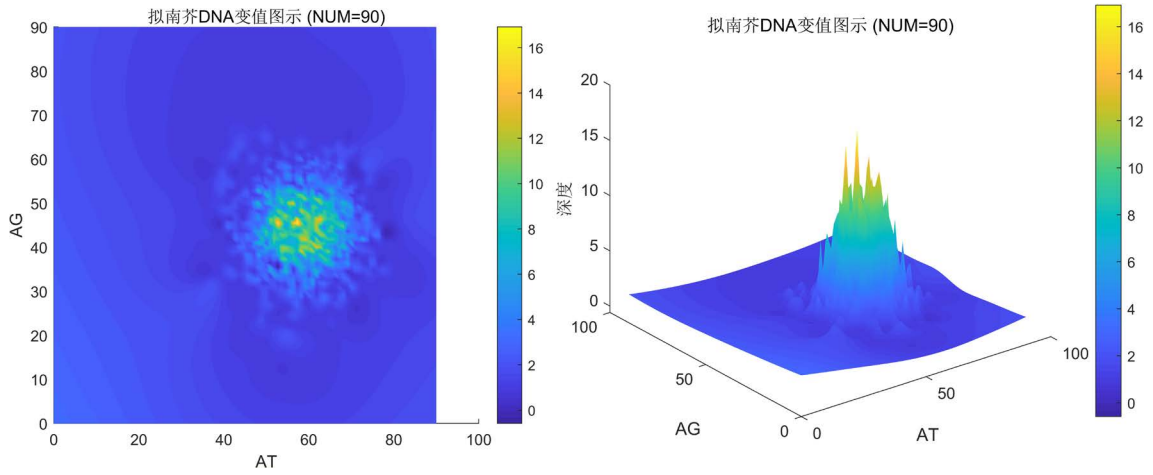

Figure 6. Calculate with 90 as the data length unit
图 6. 以 90 为数据长度单位统计

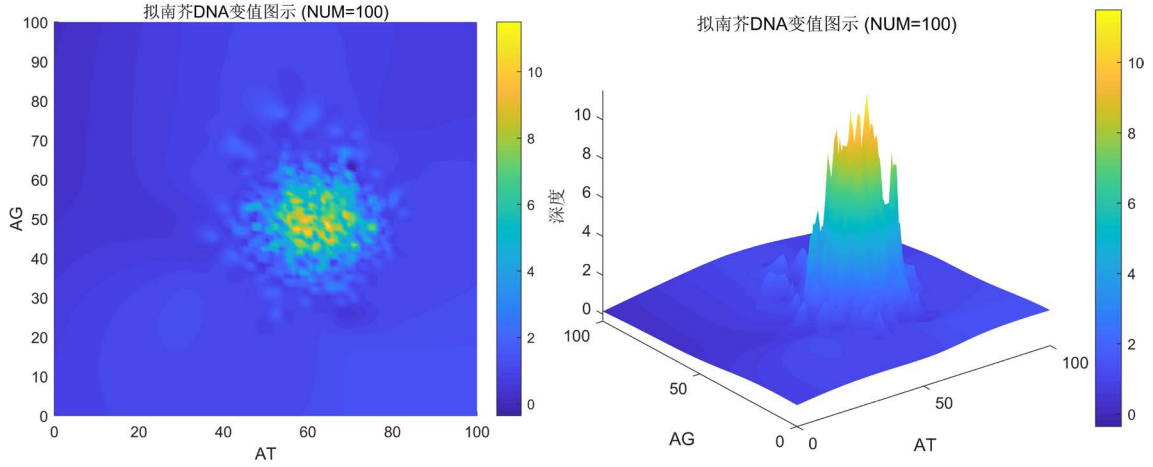

Figure 7. Calculate with100 as the data length unit
图 7. 以 100 为数据长度单位统计

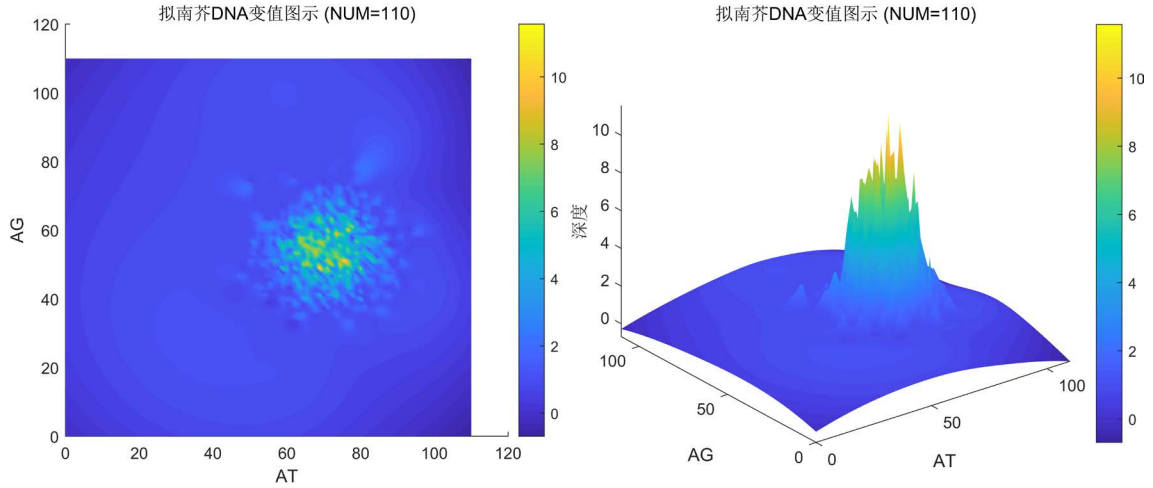

Figure 8. Calculate with 110 as the data length unit
图 8. 以 110 为数据长度单位统计

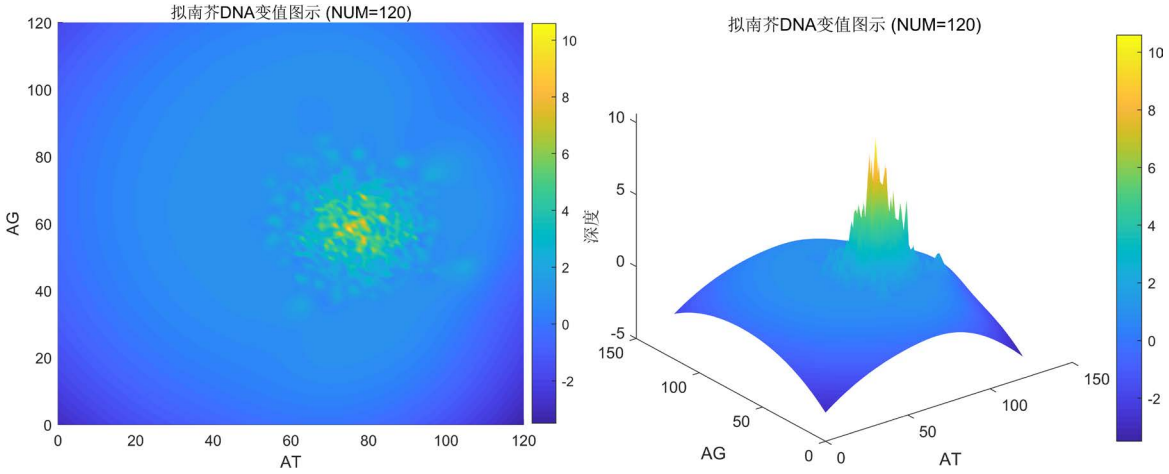

Figure 9. Calculate with 120 as the data length unit
图 9. 以 120 为数据长度单位统计

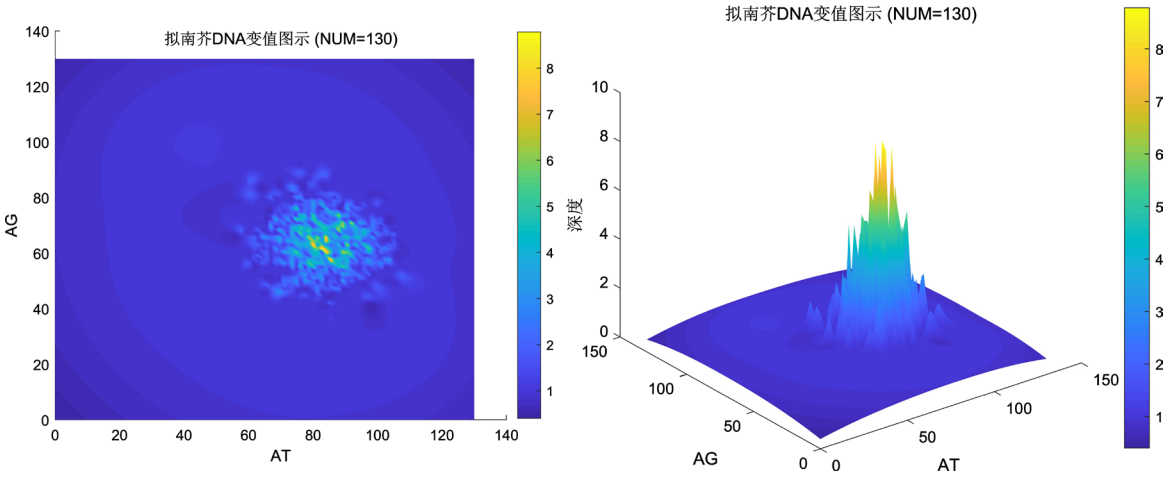

Figure 10. Calculate with 130 as the data length unit
图 10. 以 130 为数据长度单位统计

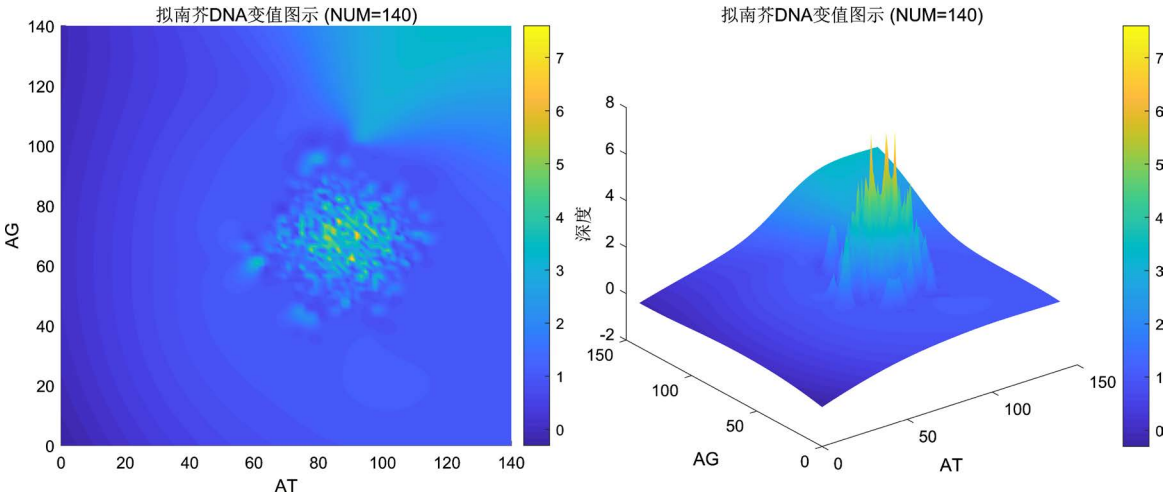

Figure 11. Calculate with140 as the data length unit
图 11. 以 140 为数据长度单位统计

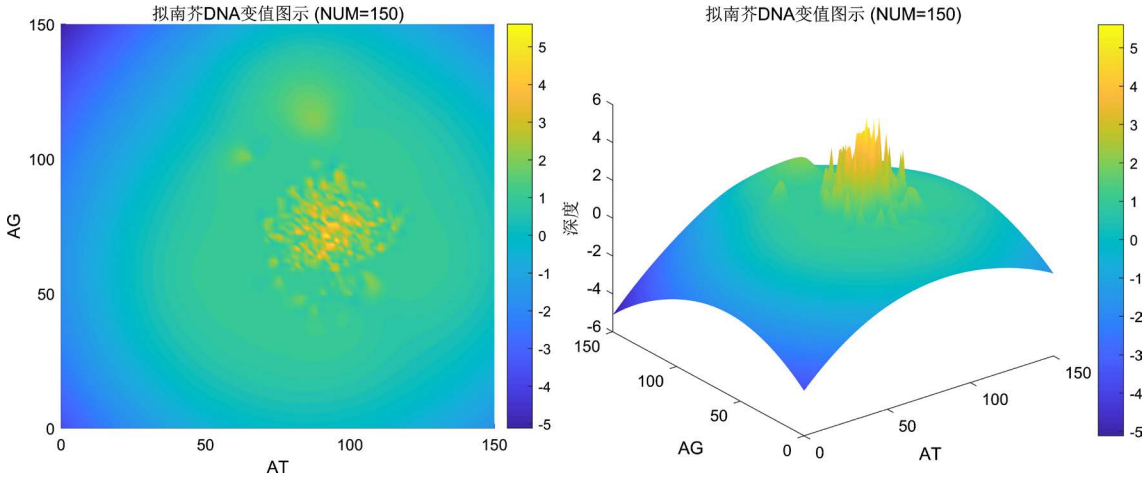

Figure 12. Calculate with 150 as the data length unit
图 12. 以 150 为数据长度单位统计

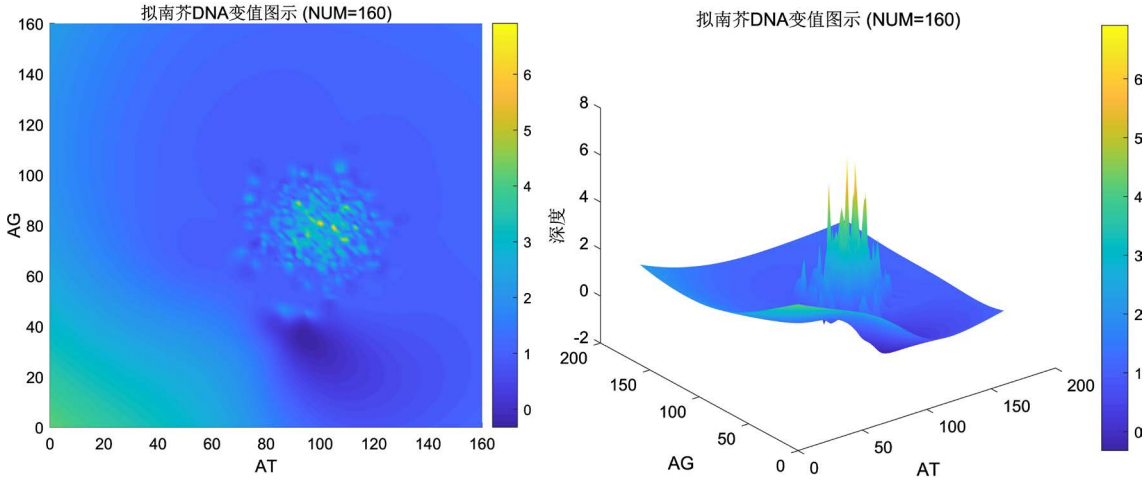

Figure 13. Calculate with 160 as the data length unit
图 13. 以 160 为数据长度单位统计

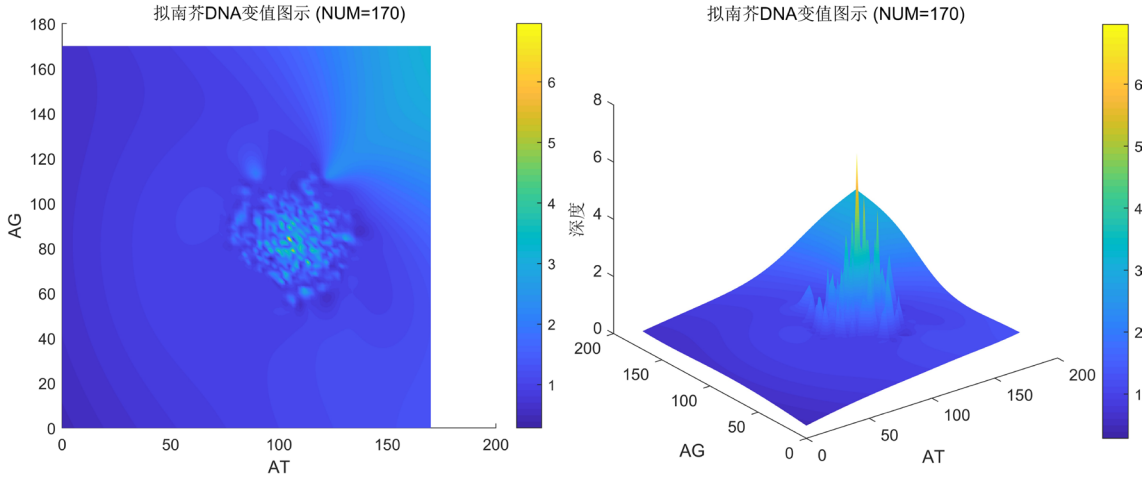

Figure 14. Calculate with 170 as the data length unit
图 14. 以 170 为数据长度单位统计

4) 数据中心点与数据分段长度有关，体现了一一对应关系，避免了因不同碱基序列得到相同图示的问题。

4. 总结

拟南芥的基因在植物基因研究上的作用类似于果蝇在真核生物研究上的作用。本文所做的拟南芥变值图示系统可以更方便和清晰地看到拟南芥 DNA 序列的特征。随着植物基因序列的研究领域里更多新型测序和研究方法的出现，变值图示系统在将来也许能为植物基因的高维研究与序列分析提供更有效的参考。

致　谢

感谢郑智捷教授对本文工作的悉心指导，感谢云南大学软件学院和云南省重点工程实验室的支持。

参考文献

[1] Zheng, J. (2018) Variant Construction from Theoretical Foundation to Applications. Springer, Berlin. https://doi.org/10.1007/978-981-13-2282-2

[2] Hamori, E. and Ruskin, J. (1983) H Curves, a Novel Method of Representation of Nucleotide Series Especially Suited for Long DNA Sequences. *The Journal of Biological Chemistry*, **258**, 1318-1327.

[3] Gates, M.A. (1986) A Simple Way to Look at DNA. *Journal of Theoretical Biology*, **119**, 319-328. https://doi.org/10.1016/S0022-5193(86)80144-8

[4] Nandy, A. (1996) A New Graphical Representation and Analysis of DNA Sequence Structure I. Methodology and Application to Globin Genes. *Current Science*, **70**, 611-668.

[5] Leong, P.M. and Morgenthaler, S. (1995) Random Walk and Gap Plots of DNA Sequences. *Computer Applications in the Biosciences*, **11**, 503-507. https://doi.org/10.1093/bioinformatics/11.5.503

[6] Guo, X., Randic, M. and Basak, S.C. (2002) A Novel 2-D Graphical Representation of DNA Sequences of Low Degeneracy. *Chemical Physics Letters*, **350**, 106-112. https://doi.org/10.1016/S0009-2614(01)01246-5

[7] Wu, Y., Liew, A.W., Yan, H. and Yang, M. (2003) DB-Curve: A Novel 2D Method of DNA Sequence Visualization and Representation. *Chemical Physics Letters*, **367**, 170-176. https://doi.org/10.1016/S0009-2614(02)01684-6

[8] Yau, S.S.S., Wang, J., Niknejad, A., Lu, C., Jin, N. and Ho, Y. (2003) DNA Sequence Representation without Degeneracy. *Nucleic Acid Research*, **31**, 3078-3080. https://doi.org/10.1093/nar/gkg432

[9] https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release