文章编号：1000-1573(2019)06-0115-07　　**DOI**：10.13320/j.cnki.jauh.2019.0131

全基因组水平上基因家族分布及进化分析方法研究

刘　宪[1]，于继高[2]，肖志刚[1]，黄　晟[1]

（1. 河北农业大学 机电工程学院，河北 保定 071001；2. 华北理工大学 生命科学学院，河北 唐山 063210）

摘要： 随着众多植物基因组测序计划的相继完成，在全基因组水平上更加高效地解析基因家族进化机制迫在眉睫，本研究提供了 1 种适用于植物基因组水平检测基因家族分布及进化的计算机分析方法。本研究依据物种染色体长度、基因物理位置、候选基因 *Ks* 值等数据，结合三阶贝塞尔曲线算法，利用 Python 语言绘制基因染色体定位图，揭示基因间相似性及进化机制。为验证方法的实用性，本研究以亚洲棉 *DELLA* 基因家族为例进行了进化分析。结果显示在亚洲棉进化过程中 *GA05*、*GA02* 和 *GA13* 有明显的同源 Cluster 被鉴定；*GA05* 染色体上的基因簇更活跃、变化更快；*GA03* 基因簇发生了独立进化；*GA02*、*GA13* 基因出现较晚，可能来源于 *GA05*。本方法是 1 种实用的基因家族分析工具，适用于全基因组水平基因家族分布及进化分析。

关键词： 基因家族；Python；三阶贝塞尔曲线；*DELLA*

中图分类号： S562.035.3　　**开放科学（资源服务）标志码（OSID）：**

文献标志码： A

Research on gene family distribution and evolutionary analysis method at genome-wide level

LIU Xian[1], YU Jigao[2], XIAO Zhigang[1], HUANG Sheng[1]

(1. College of Electrical and Mechanical Engineering, Hebei Agricultural University, Baoding 071001China;
2. College of Life Science, North China University of Science and Technology, Tangshan 063210, China)

Abstract: With the completion of many plant genome sequencing projects, it is urgent to analyze the evolutionary mechanism of gene family more efficiently at the genome-wide level. This study provides a computer analysis method for genome-wide detection of gene family distribution and evolution. Based on the data of species chromosome length, gene physical location and *Ks* value of candidate genes, and combined with the third-order Bessel curve algorithm, the software draws gene chromosome mapping using Python language, and reveals the similarity between genes and evolutionary mechanism. In order to verify practicability of the method, the evolutionary analysis of *DELLA* gene family in *G.aroreum* was carried out. It showed that *GA05*, *GA02* and *GA13* had obvious homologous clusters identified during the evolution of *G.aroreum*; the *GA05* gene cluster was more active and changed faster; the *GA03* gene cluster evolved independently; *GA02* and *GA13* genes appeared late, probably from *GA05*. As a practical tool for gene family analysis, the method is suitable for detection of genome-wide gene family distribution and evolutionary analysis.

Keywords: gene family; Python; third-order Bezier curve; *DELLA*

收稿日期： 2019-03-14

基金项目： 保定市科技计划项目 (18ZN020).

第一作者： 刘　宪（1993-），女，河北石家庄人，硕士研究生，主要从事智能检测与控制技术研究 .
E-mail：1538798913@qq.com
于继高（1992-），男，河南禹州人，硕士研究生，主要从事植物基因组，比较基因组学，基因组进化研究 .
E-mail：yujigao01@163.com；黄晟的贡献等同于第一作者 .

通讯作者： 肖志刚（1970-），男，河北保定人，副教授，主要从事智能检测与控制技术研究 .E-mail：xiaozhg@hebau.edu.cn

探索物种起源与进化机制一直是生命科学研究的核心话题和终极目标。上世纪90年代出现的基因组测序技术为生物进化过程中关键事件的破译提供了新的研究方法[1,2]。

在物种起源与进化研究中发现很多物种经历了基因组加倍、多倍体二倍化等事件，这也加剧了物种基因组的复杂性[3,4]。作为基因组重要组成部分的基因家族，在基因组进化过程也经历了多次复杂事件，最终导致家族基因数目更多、在基因组中分布更广[5-7]，同时也使得系统、科学地认知基因家族更加困难[5,8]。而认知基因家族的进化机制是物种进化研究的重要内容之一，但长期以来进展缓慢。近年来，随着众多植物基因组测序工作的完成，以此为基础揭示基因家族的起源与分子基础已成为可能[8-14]。

为在全基因组水平上更加高效地解析基因家族进化机制，本研究依据基因组测序的染色体物理长度、基因的物理位置及基因*Ks*值（同义突变），利用Python语言[15-17]，开发了1个适用于全基因组水平检测基因家族分布及进化分析的软件。为验证软件的适用性，本研究对亚洲棉*DELLA*基因家族进行了初步分析。

1 家族基因同源进化图生成方法

1.1 数据类型

本系统作图需要3种类型文件，以亚洲棉基因组数据为例[12]。一是染色体长度文件，后缀名为.len，文件内容格式为“染色体编号 染色体长度”，例如：“GA01 113 035 596”；二是家族基因染色体位置文件，后缀名为.gff，文件内容格式为“染色体编号 基因ID号 基因起始位点 基因终止位点”，例如：“GA01 GA01^0782.1 11 522 434 11 524 188”；三是基因*Ks*值文件，后缀名为.txt，文件内容格式为“基因ID1 基因ID2 *Ka*值 *Ks*值”，例如：“GA01^0782.1 GA01^1065.1 0.941 226 972 1.961 003 321”。

候选*DELLA*基因家族以拟南芥*DELLA*基因家族数据为基础[10]，通过本地Blast亚洲棉基因组基因库获得82个基因，用于本实验分析。

生成染色体长度文件需要准备的数据为基因组测序的染色体fasta序列文件，后缀名为.fa。基因家族位置文件依据基因家族ID号在基因组测序后基因位置数据库提取，该数据库后缀为.gff，提取后将基因位置文件格式转换为“染色体编号 基因ID号 基因起始位点 基因终止位点 正义/反义”，例如“Chr01 Ga01G0782 11 522 434 11 524 188 +”。生成基因*K*值文件需要基因家族cds序列文件，依据基因组测序得到的cds序列文件，以家族基因ID为关键字提取获得，后缀名为.cds，文件格式为“>基因ID号 cds序列”，例如“>Ga01G0782^pCAGCTGC…… TGTTAGC”（注：^p为换行符）。

1.2 数据预处理

（1）染色体长度文件准备。染色体长度文件是由Python脚本对基因fasta文件所载内容统计得来的，脚本使用了Bio.SeqIO.parse（文件句柄，序列格式）函数来统计每条染色体序号和序列长度，生成的染色体长度文件在fasta文件路径下。

（2）基因家族位置文件准备。基因家族位置文件是以家族基因ID号为关键字在全部位置文件中提取得来，提取出基因家族的染色体编号、基因ID号、基因起始位点和基因终止位点，生成的基因家族的位置文件在原位置文件路径下。

（3）基因家族*K*值文件准备。基因家族*K*值文件是由perl脚本对待研究物种基因家族cds序列文件进行非同义替换率(*Ka*)和同义替换率(*Ks*)计算[18-19]，生成的*K*值文件在cds序列文件路径下。*K*值文件也可通过*KaKs*_Calculator[20-22]获得。

Ka= 发生非同义替换的SNP数/非同义替换位点数

Ks= 发生同义替换的SNP数/同义替换位点数

1.3 家族基因同源进化图绘制

1.3.1 作图区域定义 本系统选取8×8（英寸）为作图区域，文本标注及同源基因连线都在该区域内，横纵坐标均定义为（0~1）。

1.3.2 染色体排序输出

（1）计算染色体图总长度

$$total=\mathrm{sum}[\mathrm{len}(i)]*(1+1/GAP\_RATIO),\quad i=(1,2\ldots\ \ldots,13) \qquad (1)$$

其中*total*为染色体图总长度，*i*为染色体序号，为每条染色体长度，sum[len(*i*)]为第1条到第*i*条染色体的总长度，*GAP_RATIO*为裂隙比，表示染色体与染色体之间空隙的比值，取值为4。

（2）计算每条染色体起始位置对应弧度

$$rad_{start}=\frac{2\pi*[\mathrm{sum}[len(i-1)]+\frac{total*(i-1)}{(GAP\_RATIO)*13}}{total},$$

$$i=(1,2\ldots\ \ldots,13) \quad (2)$$

其中 rad_{start} 为每条染色体起始位置对应弧度，i=1 时，rad_{start}=0。

（3）计算每条染色体终止位置对应弧度

$$rad_{stop}=\frac{2\pi*[\mathrm{sum}[len(i)]+\frac{total*(i-1)}{(GAP\_RATIO)*13}}{total},$$

$$i=(1,2\ldots\ \ldots,13) \quad (3)$$

其中 rad_{stop} 为每条染色体终止位置对应弧度。

（4）计算每条染色体起始、终止位置对应的坐标

染色体内外圈半径分别为 0.33、0.335，依据极坐标转换直角坐标公式求得直角坐标。

$$\begin{cases}x=r*\cos(rad)\\y=r*\sin(rad)\end{cases} \quad (4)$$

其中 r 为极坐标半径，rad 为极坐标弧度。

（5）画出染色体排序图并标注

根据前面计算结果，做出内外圈相应的弧线，做出起始终止端半圆，并为每条染色体做出文本标注，例如“GR01”。

1.3.3　基因块标记

(1) 计算每个基因所在弧度

$$rad=\frac{2\pi*[start+sum[len(i-1)]+\frac{total*(i-1)}{(GAP\_RATIO)*13}}{total} \quad (5)$$

其中 rad 为每个基因所在弧度，$start$ 为基因家族位置文件中的起始位点，i=1 时，$rad=\frac{2\pi*start}{total}$。

(2) 做出基因块标记

据公式（4）将每个基因的极坐标转换为直角坐标，并以该坐标为中心做出长 0.01、宽 0.006 的基因标记块，给每个基因做出文本标注，例如“0782”。

1.3.4　同源基因鉴定和显示　本研究运用了三阶贝塞尔曲线拟合圆弧的方法[23-26]，依据物种基因组及基因 Ks 值[27-28]，鉴定基因同源性及进化机制、以线性形式输出。

三阶贝塞尔曲线公式：

$$B(t)=P_0(1-t)^3+3P_1t(1-t)^2+3P_2t^2(1-t)+P_3t^3,\ t\in[0.1] \quad (6)$$

其中 $B(t)$ 为曲线方程，t 取值为 0~1，步长为 0.01，P_0、P_1、P_2、P_3 分别为 4 个控制点，其坐标确定如下：$P_0(x_1, y_1)$，$P_3(x_2, y_2)$；若 2 个同源基因不在同 1 条染色体上，取控制点 $P_1(0.5x_1, 0.5y_1)$，$P_2(0.5x_2, 0.5y_2)$ 若 2 个同源基因在同 1 条染色体上取控制点 $P_1(x_1+0.5(y_2-y_1), y_1+0.5(x_1-x_2))$，$P_2(x_2+0.5(y_2-y_1), y_2+0.5(x_1-x_2))$。

(x_1, y_1)，(x_2, y_2) 分别为 2 个同源基因坐标，由同源基因极坐标转换为直角坐标公式确定

$$\begin{cases}(x_1, y_1)=(r_1\cos(rad_1), r_1\sin(rad_1))\\(x_2, y_2)=(r_2\cos(rad_2), r_2\sin(rad_2))\end{cases} \quad (7)$$

其中 r_1，r_2 分别为 2 个同源基因的半径，rad_1，rad_2 分别为 2 个同源基因所在的弧度。

将公式（6）变形为：

$$B(t)=at^3+bt^2+ct+P_0 \quad (8)$$

$$\text{其中}\begin{cases}c=3(P_1-P_0)\\b=3(P_2-P_1)-c\\a=P_3-P_0-c-b\end{cases} \quad (9)$$

（1）绘制基因组不同染色体基因同源图

不在同一染色体上同源基因 4 个控制点矩阵为：

$$\begin{bmatrix}x\\y\end{bmatrix}=\begin{bmatrix}x_1 & 0.5x_1 & 0.5x_2 & x_2\\y_1 & 0.5y_1 & 0.5y_2 & y_2\end{bmatrix} \quad (10)$$

依据公式（9）、（10）可以推导出

$$\begin{cases}x(t)=(x_1-x_2)t^3+1.5x_2t^3-x_1t+x_1\\y(t)=(y_1-y_2)t^3+1.5y_2t^3-y_1t+y_1\end{cases} \quad (11)$$

根据变量 t 不同取值求出 $x(t)$、$y(t)$，并在图上以（$x(t)$，$y(t)$）为坐标找出这些点并连接。

（2）绘制基因组同一染色体基因同源图

同一染色体上同源基因 4 个控制点矩阵为：

$$\begin{bmatrix}x\\y\end{bmatrix}=\begin{bmatrix}x_1 & x_1+0.5(y_2-y_1) & x_2+0.5(y_2-y_1) & x_2\\y_1 & y_1+0.5(x_1-x_2) & y_2+0.5(x_1-x_2) & y_2\end{bmatrix} \quad (12)$$

依据公式（9）、（12）可以推导出：

$$\begin{cases}x(t)=2t^3(x_1-x_2)+1.5t^2[2(x_2-x_1)+(y_1-y_2)]+1.5t(y_2-y_1)+x_1\\y(t)=2t^3(y_1-y_2)+1.5t^2[2(y_2-y_1)+(x_2-x_1)]+1.5t(x_1-x_2)+y_1\end{cases} \quad (13)$$

其余步骤与不同染色体上同源基因图相同。

1.4 程序流程图

生成家族基因同源进化图的 Python 脚本程序流程图如图 1 所示。

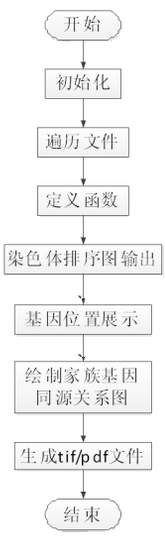

图 1 生成家族基因同源进化图的程序流程图

Fig.1 Program flow chart for generating homologous evolution map of family genes

初始化：规定作图区域，设定坐标轴区间，设定染色体裂隙比、染色体内外圈半径以及标记块大小。

遍历文件：读取指定文件内容，例如染色体长度文件中每条染色体的长度，位置文件中每个基因 ID 号及其在染色体上的起始终止位点，以及 *K* 值文件中每组同源基因的 *Ka* 和 *Ks* 值。如果文件格式与程序要求不一致，程序将会提示键值错误，导致程序中断执行。

定义函数：定义计算每条染色体对应的圆心角角度和圆弧长度函数；定义设置染色体图的文本标注位置和字体大小函数；定义设置基因块标记的大小形状以及颜色函数；定义设置基因块标记的文本标注位置和字体大小的函数；定义将染色体位置转换为轴坐标的函数，定义计算连接 1 组同源基因的圆弧半径函数。

做染色体排序图：计算出染色体图总长度，以极坐标 0° 为起始位置，按染色体编号从小到大逆时针排序。依据每条染色体长度确定与之相对应的圆弧长度，每条染色体由 4 部分构成，依次画出靠近圆心的圆弧、远离圆心的圆弧、染色体始端的半圆、终端的半圆。

基因位置展示：根据位置文件中基因的起始终止位点信息确定每个基因在染色体上的位置，并用蓝色方块标示，注释基因序号。

绘制家族基因同源关系图：本研究中 *Ks* 取值范围的依据是物种基因组 *Ks* 值区分的进化年代[5,10,27-29]。*Ks* 出现负值时设置程序不报错，也不连线，即对不符合要求的数据不会误操作，只标示基因位置。

生成文件：文件格式可根据用户要求改变，生成 tif 或者 pdf 格式的文件。

2 结果与分析

2.1 实验结果

利用本程序绘制的亚洲棉 *DELLA* 基因家族同源进化图如图 2，显示 *DELLA* 基因家族分布在除 *GA08* 和 *GA10* 外的所有染色体上；*GA05*、*GA02* 和 *GA*13 有明显的同源 cluster 被鉴定，涉及 18 个基因（其中 *GA02* 基因簇 3 个，*GA05* 基因簇 10 个，*GA13* 基因簇 5 个），*GA05* 染色体上 1 个基因簇更活跃、变化更快；*GA03* 基因簇发生了独立进化；*GA02*、*GA13* 基因出现较晚，可能来源于 *GA05*；*GA01* 上 2293、2294 和 2295，*GA*05 上 4048 和 4053、4049 和 4051 用蓝线展示，*Ks* 值大于 1.3，出现时间最早，放大后的 *GA05* 基因簇连线如图 3 所示。此外，图 2 中基因间连线反应了同源基因，线色反应了基因形成的年代：1.3<*Ks* 用蓝线显示，0.52 ≤ *Ks* ≤ 1.3 用黄线显示，0.11 ≤ *Ks*<0.52 用粉线显示，0 ≤ *Ks*<0.11 用红线显示。依据已报道的棉属物种进化过程：*γ* 事件发生在约 1.5 亿年前，*β* 事件发生在约 5 000 ～ 6 000 万年前，棉属物种多倍体二倍化发生在约 1 300 ～ 1 400 万年前，公式为 1.5 亿年 /1.3=*t*/*Ks* [5,8,27,30,31]。

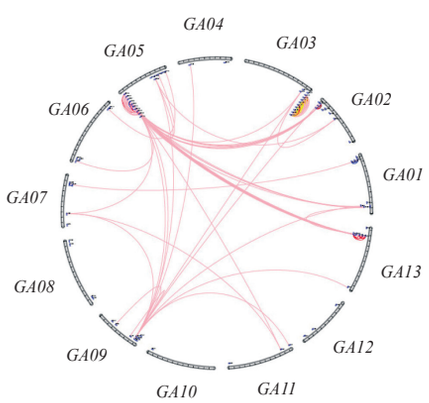

图 2 亚洲棉 *DELLA* 基因家族同源进化图

Fig. 2 *DELLA* gene family homologous evolutionary map of *G.aroreum*

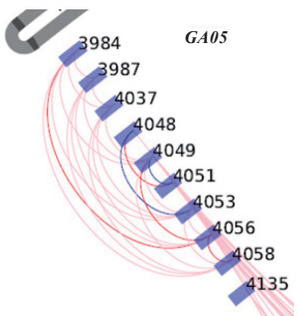

图 3　亚洲棉 GA05 连线

Fig.3　*GA05* chromosome linkage in *G.aroreum*

2.2　不同权重结果分析

由贝塞尔曲线方程可知，权重不同所做出的曲线弧度也不同，即权重大小影响同源基因连线的弧度，与物种并无直接关系，在程序中进行赋值即可。具体关系可由图 4 ～ 9（*step*=0.01）看出。

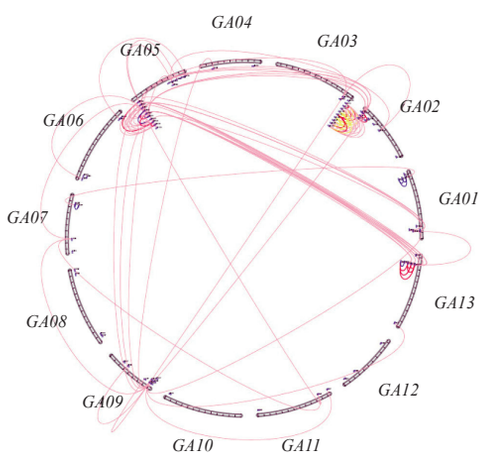

图 4　*ratio*=1.5 时，亚洲棉 *DELLA* 基因家族同源进化图

Fig.4　*DELLA* gene family homologous evolutionary map of *G.aroreum* (*ratio*=1.5)

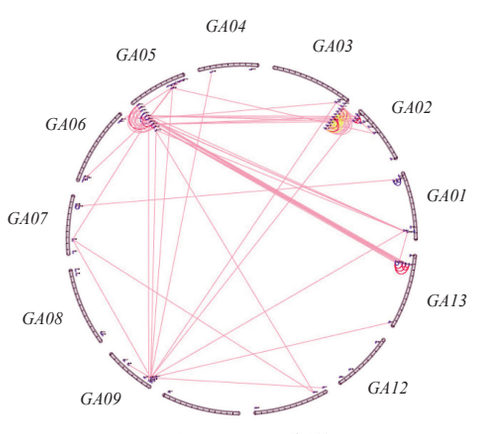

图 5　*ratio*=1 时，亚洲棉 *DELLA* 基因家族同源进化图

Fig.5　*DELLA* gene family homologous evolutionary map of *G.aroreum* (*ratio*=1)

如图 4，*ratio*=1.5 时，同 1 条染色体上同源基因连线弧度很大，不同染色体上同源基因连线从图外连接，超出染色体图范围，图形混乱。

如图 5，*ratio*=1 时，同 1 条染色体上同源基因连线为弧度较大的曲线，图形效果不受影响，但不同染色体上同源基因之间连线为直线，出线方向不一致，图形凌乱。

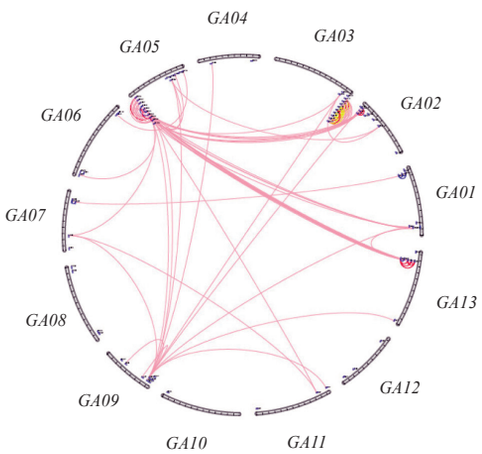

图 6　*ratio*=0.7 时，亚洲棉 *DELLA* 基因家族同源进化图

Fig.6　*DELLA* gene family homologous evolutionary map of *G.aroreum* (*ratio*=0.7)

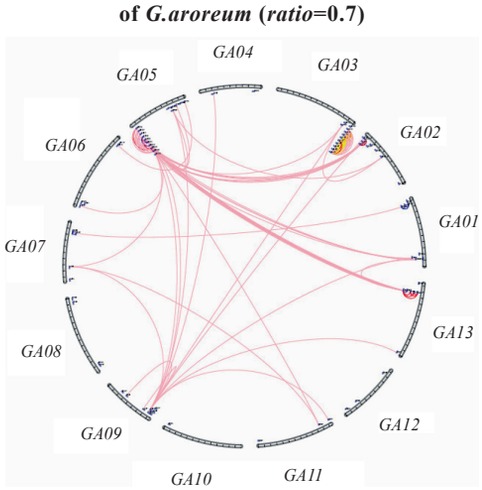

图 7　*ratio*=0.6 时，亚洲棉 *DELLA* 基因家族同源进化图

Fig.7　*DELLA* gene family homologous evolutionary map of *G.aroreum* (*ratio*=0.6)

如图 6，*ratio*=0.7 时，*GA02* 与 *GA05* 之间同源基因连线同 *GA03* 上同源基因连线会出现交叉，但是 *ratio*=0.6 时（图 7），不会出现这样的现象，所以权重为 0.6 比 0.7 更合适。

由 *ratio*=0.4 时（图 8）和 *ratio*=0.3 时（图 9）比较，可见 *ratio*=0.3 时弧度过小，使得相同染色体上同源基因连线过于密集，不容易分辨，但是 *ratio*=0.4 时，连线清晰直观，所以权重为 0.4 比 0.3 更合适。

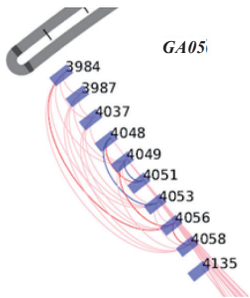

图8 *ratio*=0.4时，亚洲棉*GA05*连线

Fig.8 *GA05* chromosome linkage in *G.aroreum* (*ratio*=0.4)

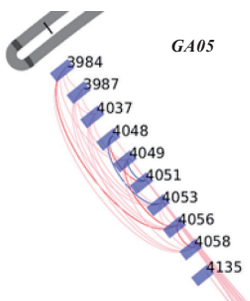

图9 *ratio*=0.3时，亚洲棉*GA05*连线

Fig.9 *GA05* chromosome linkage in *G.aroreum* (*ratio*=0.3)

通过图2～9对不同权重进化图结果的比较与讨论，*ratio*在0.4～0.6之间随机取值时图形曲线美观清晰，并且能够显示出同源基因之间的关系。可取中间值0.5为权重，在程序中进行赋值。

2.3 不同步长结果分析

步长不同所选取的中间点个数不同，即步长大小影响同源基因连线的平滑程度，与物种无直接关系，步长取值应为能整除1的小数，与中间点个数关系见表1。

表1 步长与中间点个数关系

Table 1 Relation between step and number of intermediate points

步长 Step	中间点个数 Number of intermediate points	曲线平滑程度 Level of curve smothness
0.5	1	同一条染色体上同源基因连线重合部分多，不方便读取信息
0.1	9	如图10（*ratio*=0.5，*step*=0.1），能看出明显的折线痕迹，曲线不平滑
0.05	19	图形放大后，能看出折线痕迹，图形不够美观
0.02	49	图形放大后，会看出折线痕迹，但不是很明显
0.01	99	如图3（*ratio*=0.5，*step*=0.01），图形放大后视觉上辨别不出折线痕迹

通过对不同步长进化图结果的比较（图3，图10），*step*取值越小图形曲线越平滑，经过验证*step*=0.01取99个中间点时曲线足够平滑，图像经过多倍放大后看不出折线痕迹（图3）。取更多的点会使程序运算复杂，耗时较长，因此，程序中设置步长*step*=0.01。

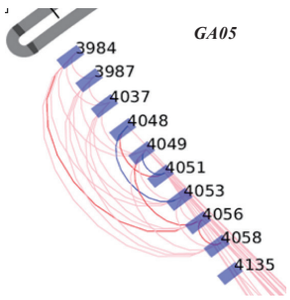

图10 *step*=0.1时，亚洲棉*GA05*连线

Fig.10 *GA05* chromosome linkage in *G.aroreum* (*step*=0.1)

3 结论

通过对不同权重以及不同步长同源进化图结果的比较与分析，最终将同源基因连线权重设定为0.5，步长设定为0.01。此参数为系统自身参数，与选取基因家族和物种无关。以亚洲棉*DELLA*基因家族为例，验证了方法的实用性。通过对基因家族数据信息的整合和分析，绘制的基因家族同源进化图能够显示大致的进化进制，帮助棉花工作者初步认识亚洲棉*DELLA*基因家族的进化规律，使今后有效认知基因家族进化规律成为可能。

作为1种实用的分析工具，本方法适用于全基因组水平基因家族分布及进化分析，但在进化过程中各个物种所经历的事件不同，需要根据待研究物种进化发生的时间设置相应的年代区间，进行同源基因连线。

参考文献：

[1] Murat Florent, Armero Alix, Pont Caroline, et al. Reconstructing the genome of the most recent common ancestor of flowering plants [J]. Nature Genetics, 2017, 49(4): 490-496.

[2] Popp Magnus, Erixon Per, Eggens Frida, et al. Origin and Evolution of a Circumpolar Polyploid Species Complex in Silene (Caryophyllaceae) Inferred from Low Copy Nuclear RNA Polymerase Introns, rDNA, and Chloroplast DNA[J]. Systematic Botany, 2005, 30(2): 302-313.

[3] Renny-Byfield Simon, Gallagher Joseph P, Grover Corrinne E, et al. Ancient Gene Duplicates in *Gossypium* (Cotton) Exhibit Near-Complete Expression Divergence [J]. Genome Biology and Evolution, 2014, 6(3): 559-571.

[4] Wang Kunbo, Wang Zhiwen, Li Fuguang, et al. The draft genome of a diploid cotton *Gossypium raimondii* [J]. Nature Genetics, 2012, 44(10): 1098-1103.
[5] Wang Xiyin, Guo Hui, Wang Jinpeng, et al. Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation [J]. New Phytol, 2016, 209(3): 1252-1263.
[6] Wang Xiyin, Shi Xiaoli, Hao Bailin, et al. Duplication and DNA Segmental Loss in the Rice Genome: Implications for Diploidization [J]. The New Phytologist, 2005, 165(3): 937-946.
[7] Paterson A H, Bowers J E, Chapman B A, et al. Ancient polyploidization predating divergence of the cereals, and Its consequences for comparative genomics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(26): 9903-9908.
[8] Paterson Andrew H, Wendel Jonathan F, Gundlach Heidrun, et al. Repeated polyploidization of *Gossypium genomes* and the evolution of spinnable cotton fibres [J]. Nature, 2012, 492(7429): 423-427.
[9] Wang Maojun, Tu Lili, Yuan Daojun, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense* [J]. Nature Genetics, 2019, 51(2): 224-229.
[10] Du Xiongming, Huang Gai, He Shoupu, et al. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits [J]. Nature Genetics, 2018, 50(6): 796-802.
[11] Yuan Daojun, Tang Zhonghui, Wang Maojun, et al. The genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres [J]. Scientific Reports, 2016, 5: 17662.
[12] Li Fuguang, Fan Guangyi, Lu Cairui, et al. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution [J]. Nature Biotechnology, 2015, 33(5): 524-530.
[13] Zhang Tianzhen, Hu Yan, Jiang Wenkai, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement [J]. Nature Biotechnology, 2015, 33(5): 531-537.
[14] Li Fuguang, Fan Guangyi, Wang Kunbo, et al. Genome sequence of the cultivated cotton *Gossypium arboreum* [J]. Nature Genetics, 2014, 46(6): 567-572.
[15] Cock P J A, Antao T, Chang J T, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics [J]. Bioinformatics, 2009, 25(11): 1422-1423.
[16] Hirashima Akinori, Rafaeli Ada, Gileadi Carina, et al. Three-dimensional pharmacophore hypotheses of octopamine receptor responsible for the inhibition of sex-pheromone production in *Helicoverpa armigera* [J]. Journal of Molecular Graphics and Modelling, 1999, 17(1): 43, 50-53, 54.
[17] Oliphant, Travis E. Python for Scientific Computing[J]. COMPUTING IN SCIENCE & ENGINEERING, 2007, 9(3): 10-20.
[18] 李雪娟 , 黄原 , 雷富民 . 山鹛鸠属鸟类线粒体基因组的比较及系统发育研究 [J]. 遗传 , 2014, 36(9): 912-920.
[19] Zhang Yuanji. miRU: an automated plant miRNA target prediction server [J]. Nucleic Acids Research, 2005, 33 (Web Server): W701-W704.
[20] Wang Dapeng, Zhang Yubin, Zhang Zhang, et al. *KaKs_* Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies [J]. Genomics Proteomics Bioinformatics, 2010, 8(1): 77-80.
[21] Wang Dapeng, Zhang Song, He Fuhong, et al. How Do Variable Substitution Rates Influence Ka and Ks Calculations [J]. Genomics Proteomics Bioinformatics, 2009, 7(3): 116-127.
[22] Zhang Zhang, Li Jun, Zhao Xiao-Qian, et al. *KaKs* Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging [J]. Geno Prot Bioinfo, 2006, 4(4): 259-263.
[23] 陈鸣芳 , 陈哲 , 何炎平 , 等 . 贝塞尔曲线在浮式风力机模型试验中的应用 [J]. 中国设备工程 , 2018(21): 114-115.
[24] 王娟 . 基于 Bezier 曲线的未标定分层重构 [J]. 北京信息科技大学学报：自然科学版 , 2017, 32(6): 48-51.
[25] 宋瑞霞 , 王小春 , 马辉 . 关于曲线拟合的广义 Bezier 方法 [J]. 计算机工程与应用 , 2005(20): 60-63.
[26] Jolly K G, Sreerama Kumar R, Vijayakumar R. A Bezier curve based path planning in a multi-agent robot soccer system without violating the acceleration limits [J]. Robotics and Autonomous Systems, 2009, 57(1): 23-33.
[27] Li Zhikun, Chen Bin, Xiuxin Li, et al. A newly-identified cluster of glutathione S-transferase genes provides *Verticillium wilt* resistance in cotton [J]. The Plant Journal, 2019, 98(2):213-227.
[28] Blanc Guillaume, Wolfe Kenneth H. Widespread Paleopolyploidy in model plant species inferred from age distributions of duplicate genes [J]. THE PLANT CELL ONLINE, 2004, 16(7): 1667-1678.
[29] Wang Jinpeng, Yuan Jiaqing, Yu Jigao, et al. Recursive Paleohexaploidization Shaped the Durian Genome [J]. Plant Physiology, 2019, 179(1): 209-219.
[30] Wang Jinpeng, Sun Pengchuan, Li Yuxian, et al. An Overlooked Paleotetraploidization in Cucurbitaceae [J]. Molecular Biology and Evolution, 2018, 35(1): 16-26.
[31] Wang Jinpeng, Sun Pengchuan, Li Yuxian, et al. Hierarchically aligning 10 Legume genomes establishes a family-level genamice platform [J].plant physiology, 2017, 174(1):284-300.

（编辑：张月清）