CyDotian_V1.0 使用手册（中文版）

1.介绍

1.1 背景

在我们对谷子 PME 基因的 pro_region 和 PME 结构域的编码序列的比较分析中，发现 pro_region 比 PME 结构域进化更快。通过使用 Dotter 软件[1]发现 pro_region 的编码序列要比 PME 结构域的编码序列内部重复更多，这个基因内部的特性可能是导致 pro_region 比 PME 结构域进化更快的一个原因。然而 Dotter、Dotlet[2]等都是老旧的基于传统的滑动窗口去噪声的原理实现的序列比对的可视化软件，虽然可以全面的反映序列比对的情况，但是没有办法实现输出内部相似区域的具体位置和数量。因此，我们找到另外一款优秀的软件 MUMmer[3]，MUMmer 里面的 repeat-match 功能基于后缀树算法可以实现序列内部重复位置的识别，但是遗憾地是寻找出来的内部重复序列都是完全匹配的，中间没有任何错配的存在。实际上，应该跟 BLAST 软件[4]一样，重复序列应该允许中间有不匹配的情况，这样的重复片段更能反应序列内部真实的重复情况。因此，开发一款既能识别序列内部重复片段数量、位置又能可视化的可交互软件的需求应运而生。

1.2 算法实现原理

改良史密斯-沃特曼算法（Smith-Waterman algorithm）：

传统的滑动窗口方法并没有考虑空位，而且想要实现重复序列具体位置的识别，最多只能实现重复片段中间有错配的不完全匹配结果。所以该算法基于史密斯-沃特

曼算法也不考虑空位的插入，直接强制要求打分矩阵中max来自于左上角。

设要比对的两序列为$A = a_1a_2 \dots a_n, B = b_1b_2 \dots b_m$，其中$n, m$分别为序列$A, B$的长度。

确定置换矩阵

$s(a, b)$-组成序列的元素之间的相似得分

创建得分矩阵H并初始化其首行和首列。该矩阵的大小为$n + 1$行$m + 1$列（注意计数从0开始）。

$$H_{k0} = H_{0l} = 0, (0 \le k \le n, 0 \le l \le m)$$

从左到右，从上到下进行打分，填充得分矩阵H剩余部分。

$$H_{ij} = \max\begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ 0 \end{cases}$$

其中，$H_{i-1,j-1} + s(a_i, b_j)$,表示将$a_i$，$b_j$比对的相似性得分，

0表示a_i，b_j到此为止无相似性。

回溯。从矩阵H中得分最高的元素开始根据得分的来源（改良之后的算法只可能来源于左上角）回溯至上一位置，如此反复直至遇到得分为0的元素。具有局部最高相似性的片段（重复片段）在此过程中产生。具有第二高相似性的片段可以通过从最高相似性回溯过程之外的最高分位置开始回溯，即完成首次回溯之后，从首次回溯区域之外的最高分元素开始回溯，以得到第二个局部相似性片段（重复片段）。同理，寻找第三个局部相似性片段（重复片段）。对这个过程进行迭代，直到找到所有的大于一定长度的相似性片段（重复片段）为止。

1.3 编程实现

为了实现高效的结果输出，我们采用底层语言 C 语言根据算法原理实现，编码 GUI 界面采用的是 Python3 语言中的 PyQt5 库。三种模块用 C 语言分别编程并命名为 bpCyDotian（用来处理 DNA 序列）,aaCyDotian（用来处理氨基酸序列）和 slidingWindow（用传统的滑动窗口原理处理 DNA 或氨基酸序列）。为了实现 C 语言和 Python 语言的混合编程，采用 Cython 进行模块编译。

2.下载和安装

CyDotian 可以免费从 GitHub 网址（https://github.com/ChenHuilong1223/CyDotian）下载。目前，网址可获得的版本仅仅是支持 window 系统，在获取该安装程序后，双击运行安装在用户的 window 系统的电脑相应的位置即可。

3.功能演示

3.1 输入文件

使用 CyDotian 的文件格式为标准的 Fasta 文件。正如下图所示（图 1），CyDotian 有三种输入文件的方式：①点击文件图标，在用户电脑对应的位置进行选择。②可以把文件路径直接输入到文本框中，或者直接把用户想要分析的文件拖进文本框中，自动识别文件路径并添加到文本框中。③直接把用户想要分析的 fasta 格式序列粘贴进 Example 文本框中。

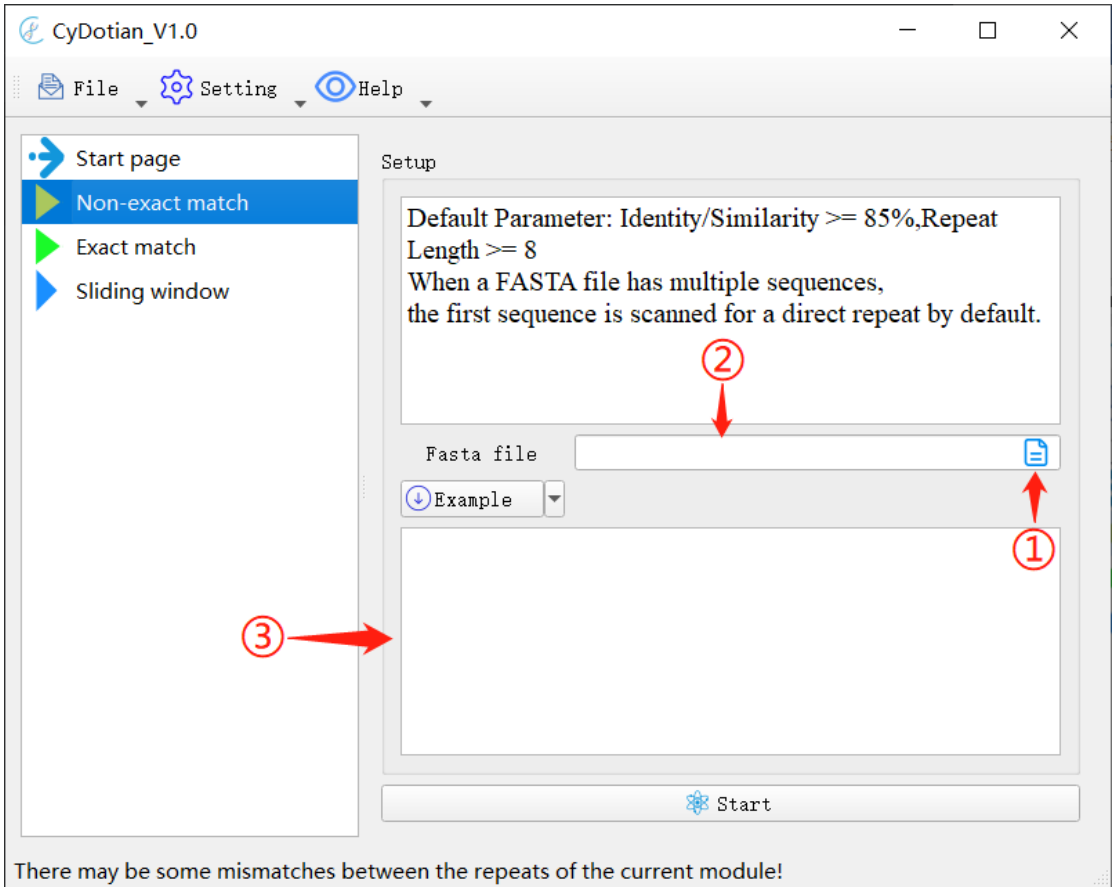

图 1 文件输入页面

3.2 三种模式

3.2.1 Start page

打开软件后，用户会看到下面演示的界面（图 2），如果用户在使用过程中发现

bug,请发送该 bug 到该页面显示的两个邮箱的任意一个地方，方便我们进行修正。

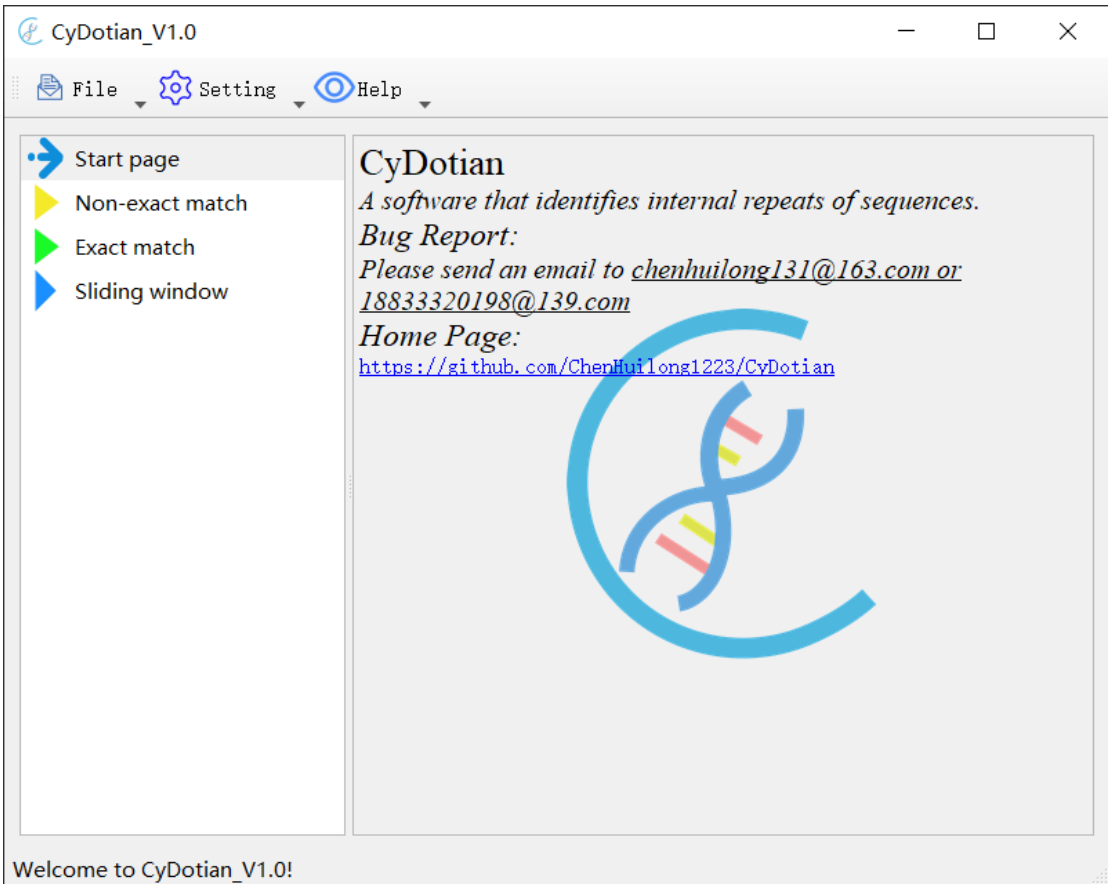

图 2 CyDotian_V1.0 起始页面

3.2.2 Non-exact match

该模式正是 CyDotian 主打的算法实现模块，用户根据自己的喜好输入正确的 fasta 文件后，点击 Start 按钮，进入到指定用户序列类型的对话框。根据用户输入的序列类型指定对应的类型（DNA 或者氨基酸）。DNA 支持两种矩阵（BLAST 和 Transition-transversion 矩阵）。氨基酸支持 BLOSUM45、BLOSUM62、BLOSUM80、BLOSUM90、PAM30、PAM70 和 PAM250 矩阵（图 3）。选择好对应的矩阵之后，点击 Run 按钮，将会弹出一个控制面板（左边）和一个点阵图窗口（右边）（图 4）。

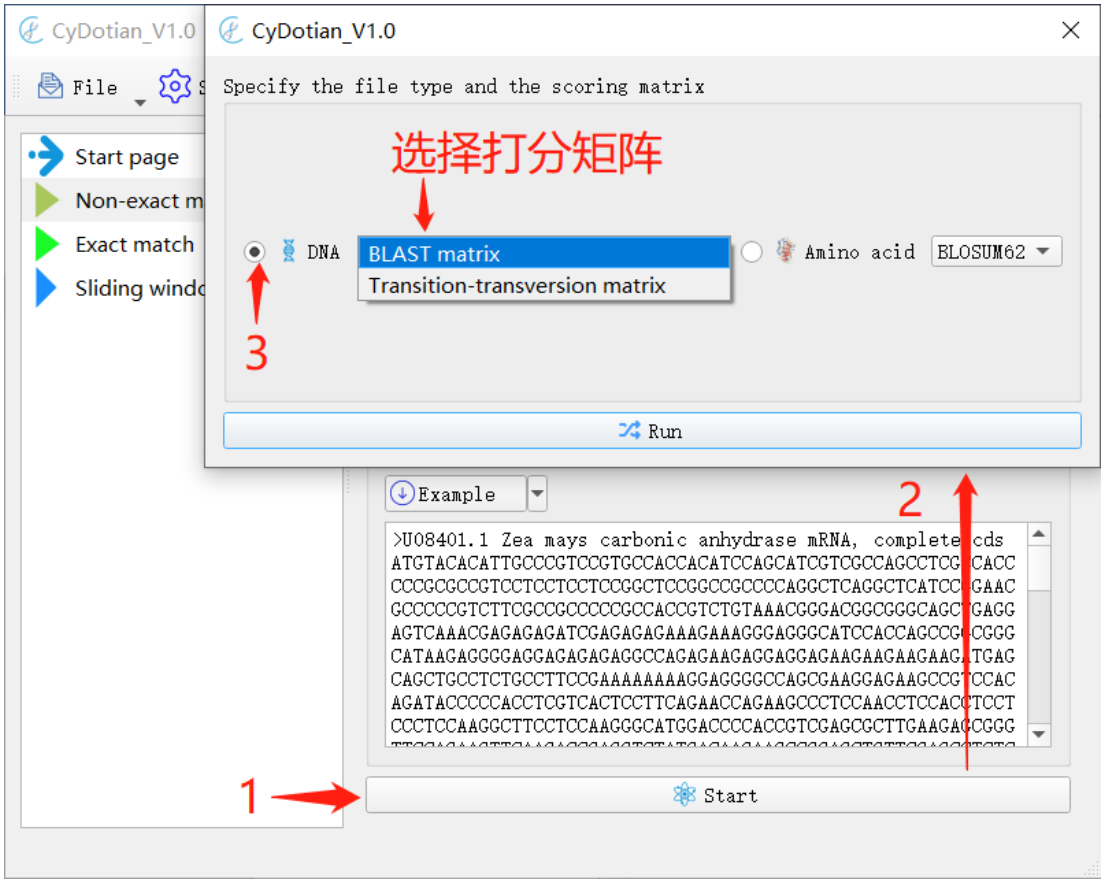

图 3 指定序列类型和选择矩阵窗口

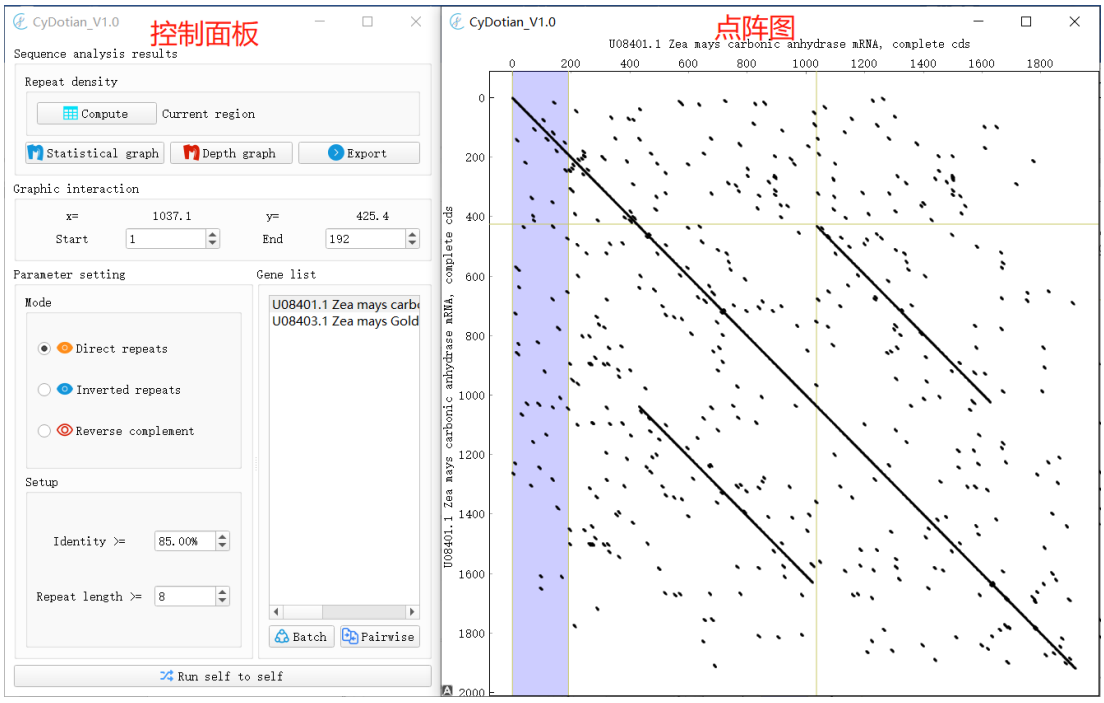

图 4 控制面板和点阵图

默认的结果是基因列表（Gene list）中第一条序列的自己对自己的比对结果。控制面板一共划分为 4 个部分：Sequence analysis results、Graphic interaction、Parameter

setting、Gene list。

Sequece analysis results 部分一共四个按钮，Compute 按钮（如果可用的情况下）点击之后将会计算出点阵图窗口的淡蓝色区域的序列内部重复密度并将结果显示在 Current region 标签的右边空白区域。Statistical graph 按钮点击之后将会弹出一个显示该序列整体情况的四种统计图（图 5）。Depth graph（如果可用的情况下）按钮点击之后将会弹出一个显示序列整体内部点图深度的一个条形图（图 6）。Export 按钮点击之后将会弹出一个布有用户可选择想要输出对应结果的勾选控件的对话框（图 7），用户选择之后，即可选择对应的输出文件夹路径并点击 Export 按钮输出即可。

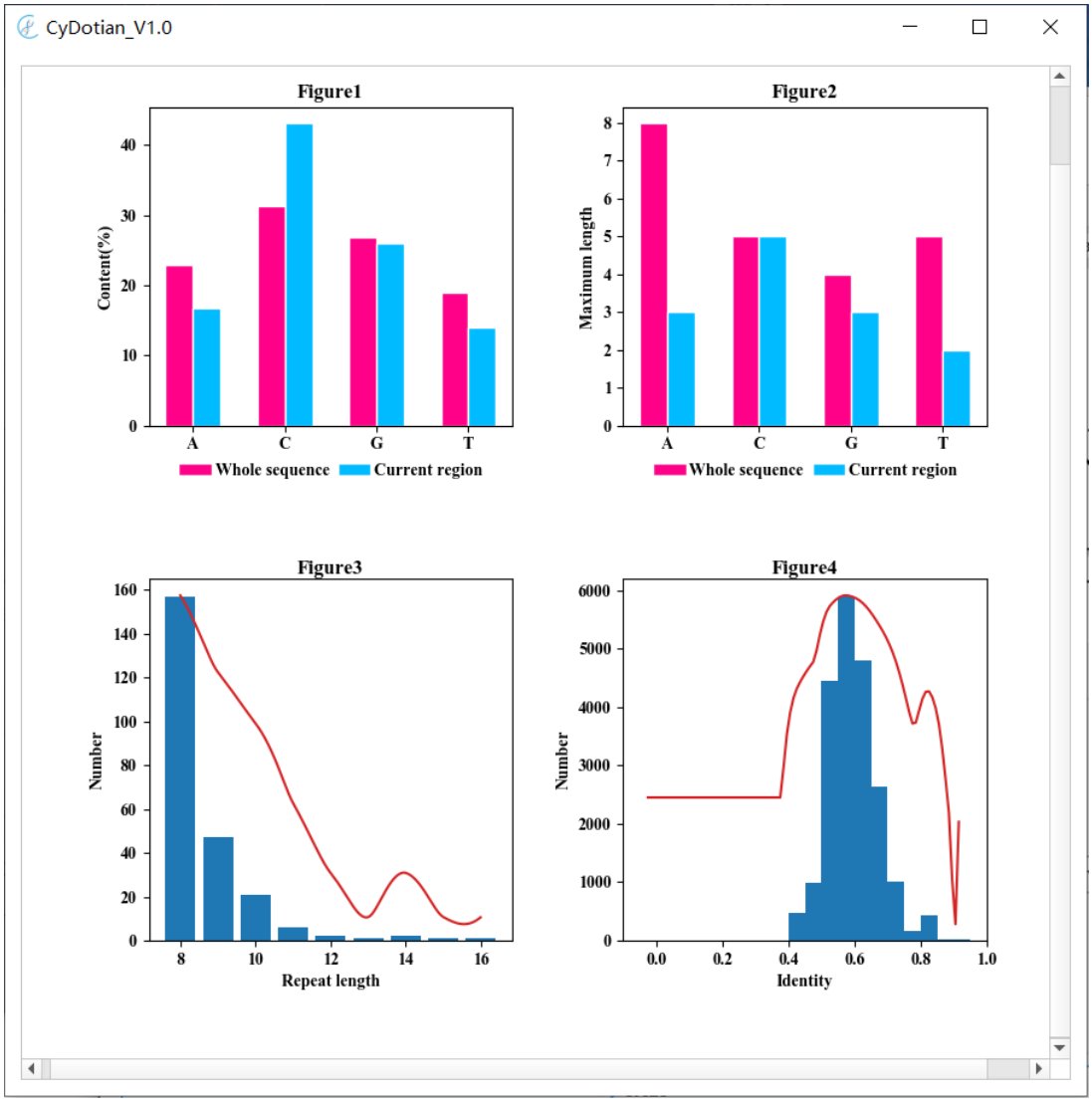

图 5 四种统计图窗口

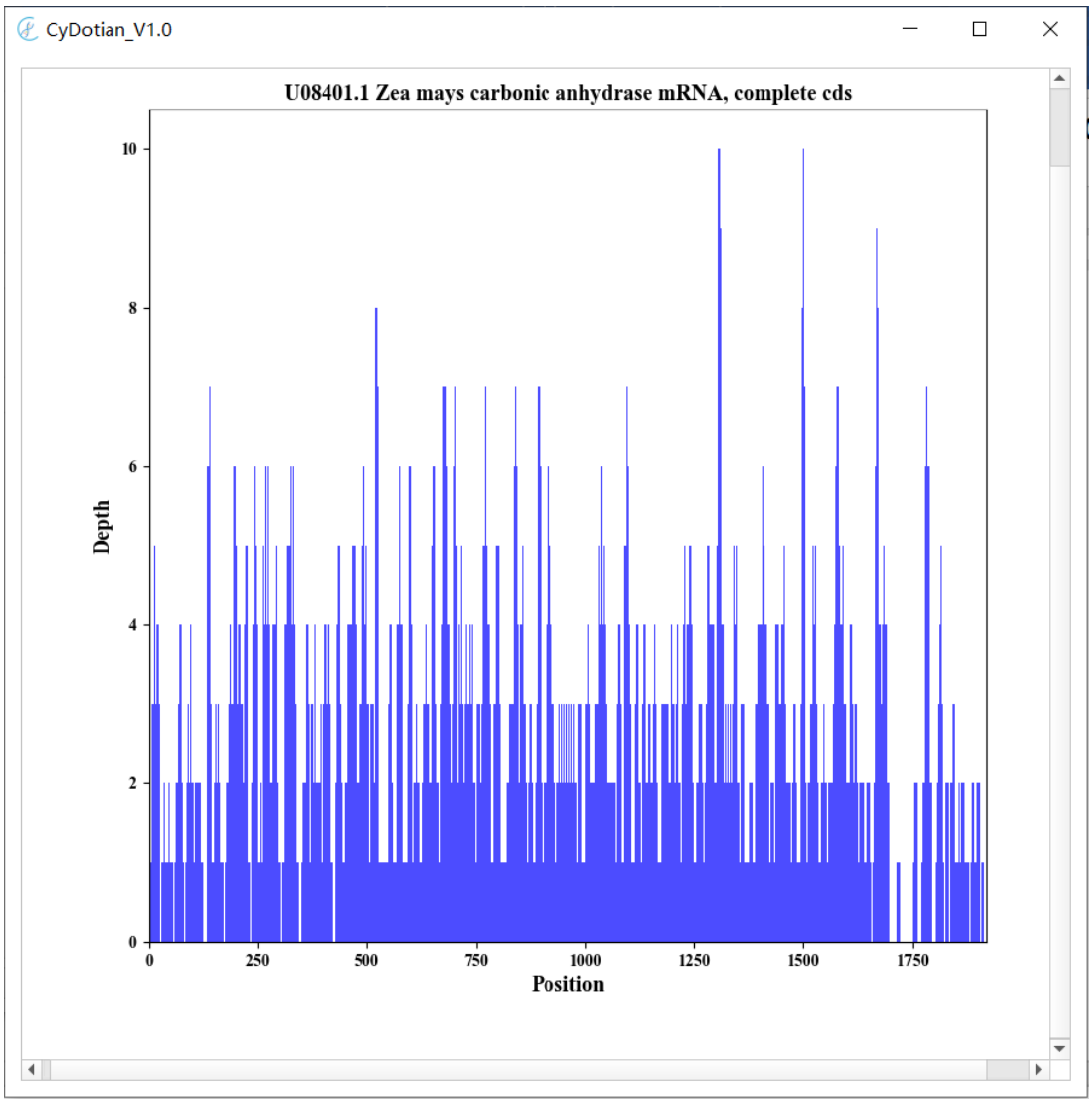

图 6 深度条形图窗口

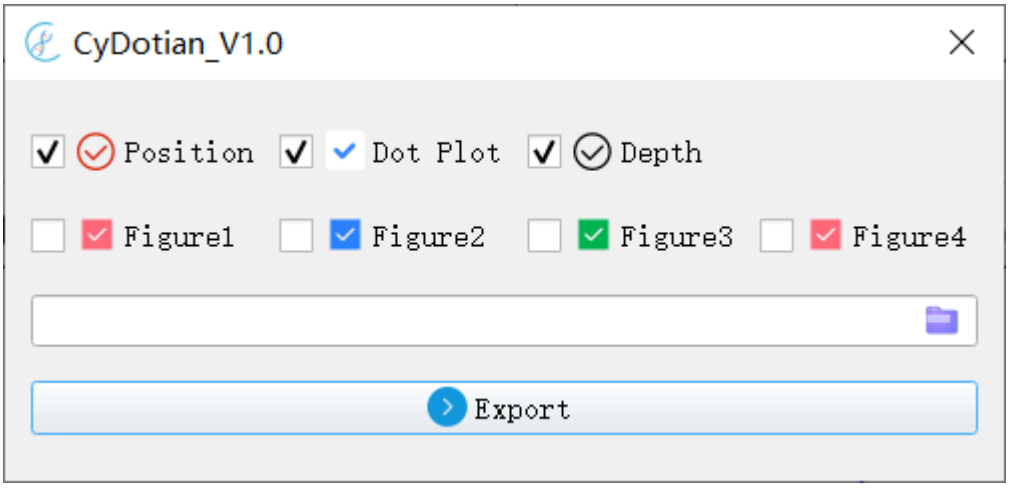

图 7 输出窗口

Graphic interaction 部分用于显示右边点阵图交互的实时信息，用户也可以在 Start 和 End 控件处手动设置想要查看的序列区域。

Parameter setting 是影响输出结果的一个重要部分，区域中 Mode 部分的的三个控件从上到下分别表示 Direct repeats（寻找正向重复）、Inverted repeats（寻找反向重复）和 Reverse complement（寻找反向互补）——当用户输入序列类型为 DNA 的时候，才会出现 Reverse complement 的控件。Setup 部分是表示用户设置的阈值，表示输出结果为 Identity/Similarity 大于等于当前阈值和 Repeat length 大于等于该长度的结果。

Gene list 部分用于显示用户输入序列的 ID 列表，用户选中那个 ID，在 Parameter setting 部分设置完参数之后即可点击 Run self to self 按钮，之后输出的结果将是这个 ID 对应序列的分析结果。该部分的 Batch 和 Pairwise 按钮的作用分别表示批量和两两比对模式。当用户点击 Batch 按钮，表示进入批量模式，将会弹出一个布有用户可选择想要输出对应结果的勾选控件的对话框（图 8），用户选择之后，即可选择对应的输出文件夹路径并点击 Export 按钮输出即可，该模式下，输出的结果为基因列表（Gene list）中所有基因的所有情况的结果。用户点击 Pairwise 按钮之后，表示进入两两比对模式，用户通过选择 Gene list 里的基因名字，将用户选择的基因 ID 显示在 Vertical 和 Horizontal 知识的文本框中（用户最后一次点击的 ID 放置在 Horizontal（水平），倒数第二次点击的 ID 放置在 Vertical（垂直））（图 9），确认无误后，点击 Run pairwise comparision 运行即可出相应的结果（图 10）。

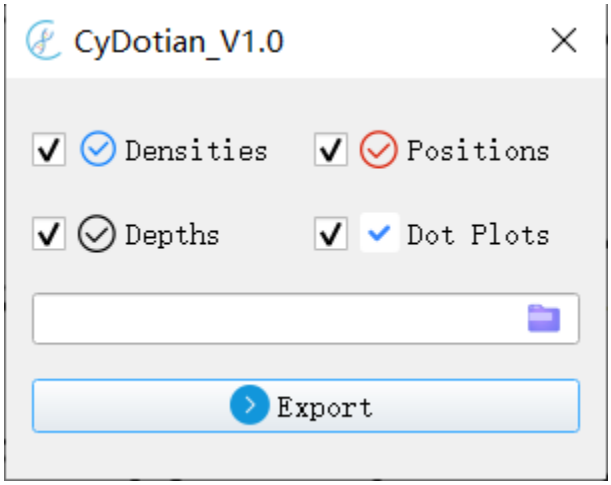

图 8 批量模式输出窗口

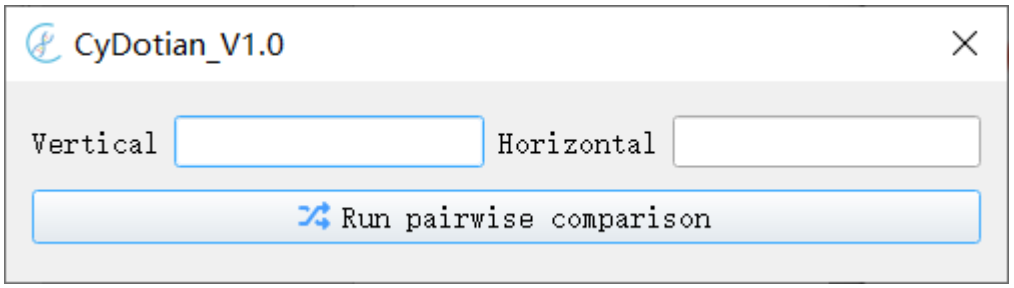

图 9 两两比对模式设置窗口

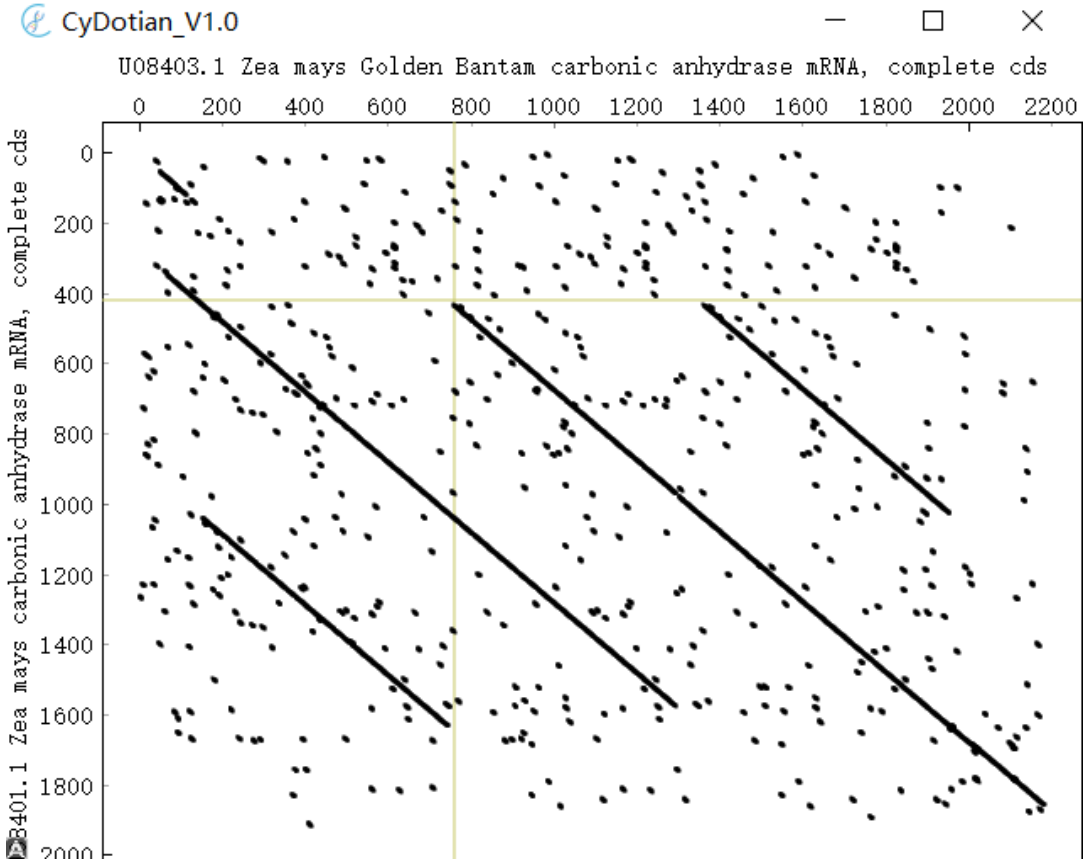

图 10 两两比对模式下不同基因比对结果的点阵图

3.2.3 Exact match

使用方法同 Non-Exact match

3.3.4 Sliding window

使用方法同 Non-Exact match

4 参考文献

1. Sonnhammer E L L, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence

analysis. Gene, 1995, 167(1–2):GC1.

2. Thomas J, Marco P. Dotlet: diagonal plots in a Web browser. Bioinformatics(2): 178-179.
3. Marçais G, Delcher A L, Phillippy A M, et al. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol, 2018, 14(1): e1005944.
4. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. Bmc Bioinformatics, 2009, 10(1):421.