

SAtoolkit V1.0

User manual

1. Software introduction

SAtoolkit V1.0 software is a toolkit for biological sequence alignment (nucleic acid sequence and amino acid sequence) written in C and Python3 languages. It integrates three mainstream biological sequence alignment algorithms through our own implementation logic. With a simple operation, users can quickly analyze their biological sequences through SAtoolkit software, and output information such as the location of similar sequences, and provide a rich graphical display. SAtoolkit V1.0 software is a combination of efficient and convenient software.

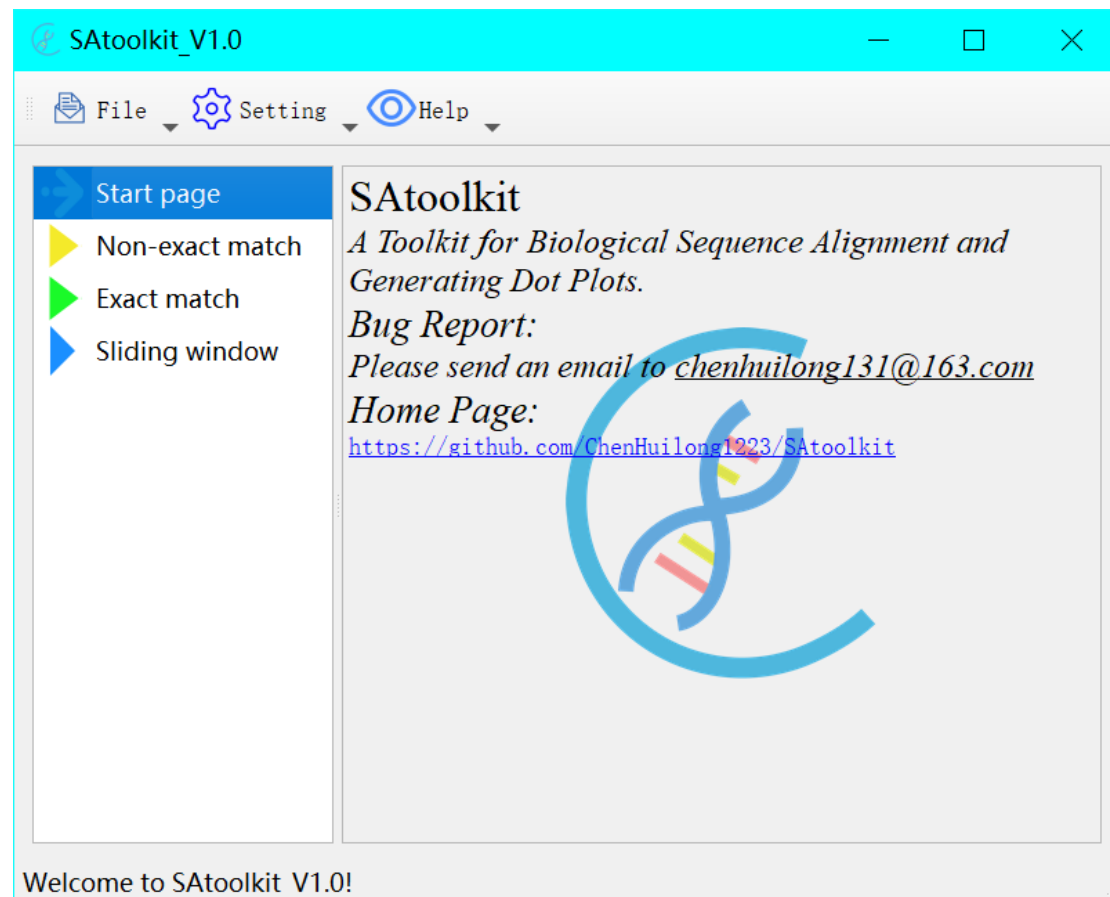
2. Writing purpose

This user manual describes how to use SAtoolkit V1.0. By reading the user manual, you can understand the functions and operations of SAtoolkit V1.0. By following the instructions in the user manual and performing actual operations, Users can quickly learn how to use SAtoolkit V1.0 software.

3. System operations

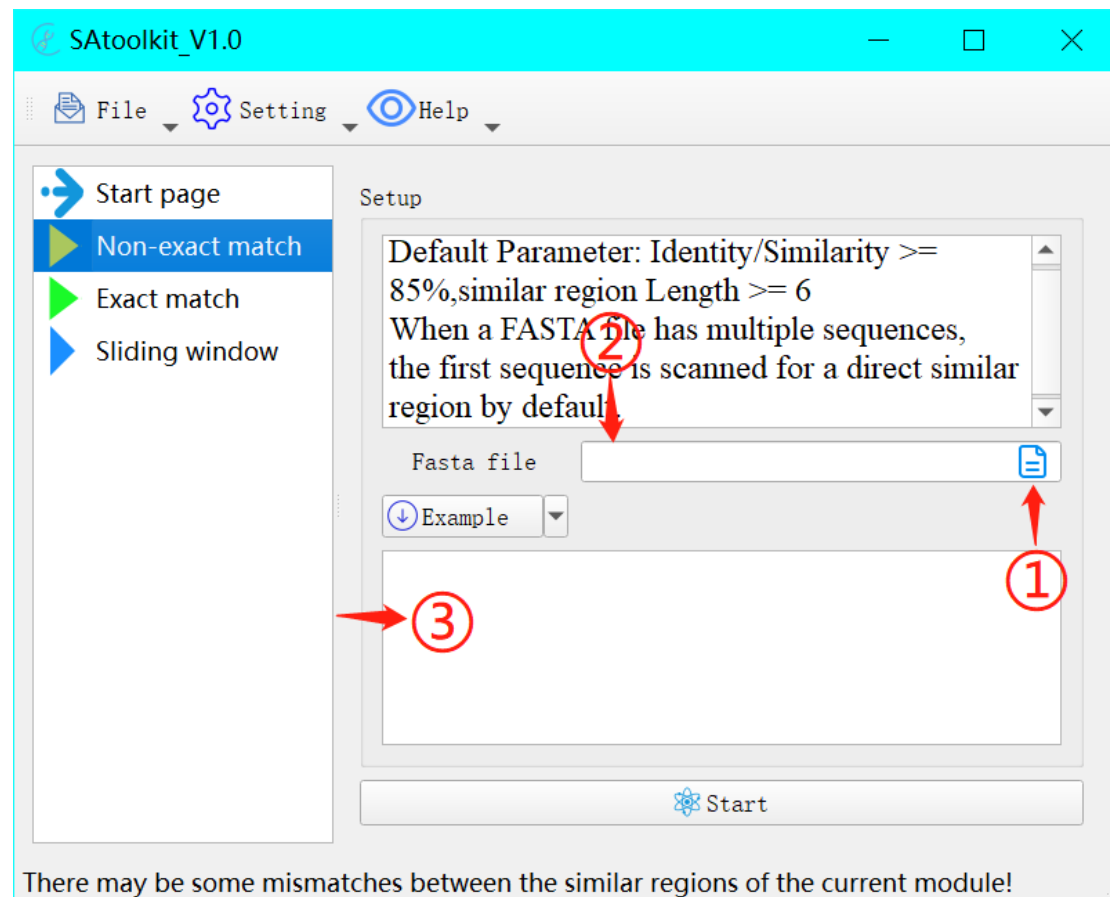
3.1 Home page

Click the icon of SAtoolkit V1.0 on the desktop to enter the home page of SAtoolkit software, as shown in the figure below:



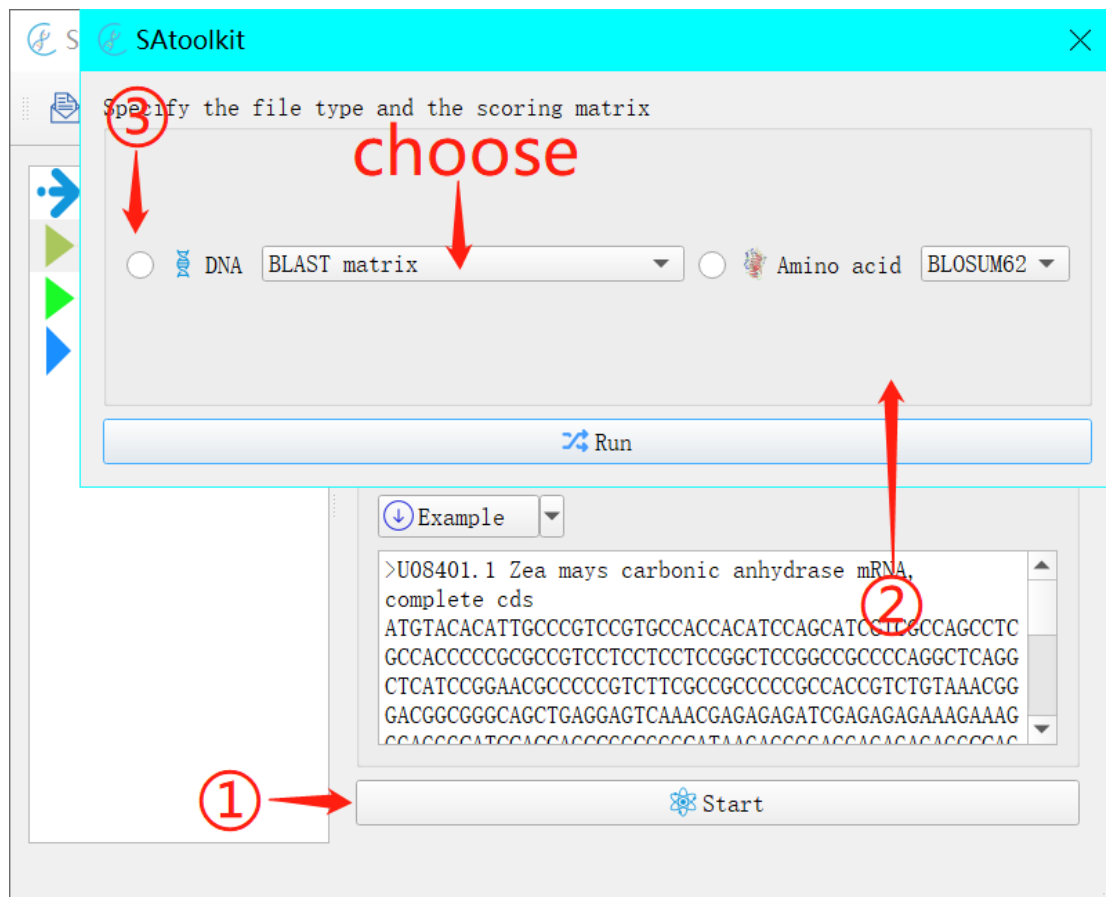
3.2 Operation

The file format for using SAtoolkit V1.0 is a standard fasta file. As shown in the image below, SAtoolkit V1.0 has three ways of entering files: ① Click on the file icon to select the corresponding location on the user's computer. ② You can input the file path directly into the text box, or directly drag the file that the user wants to analyse into the text box, which automatically recognises the file path and adds it to the text box. ③ Paste the fasta format sequence that the user wants to analyse directly into the 'Example' text box.

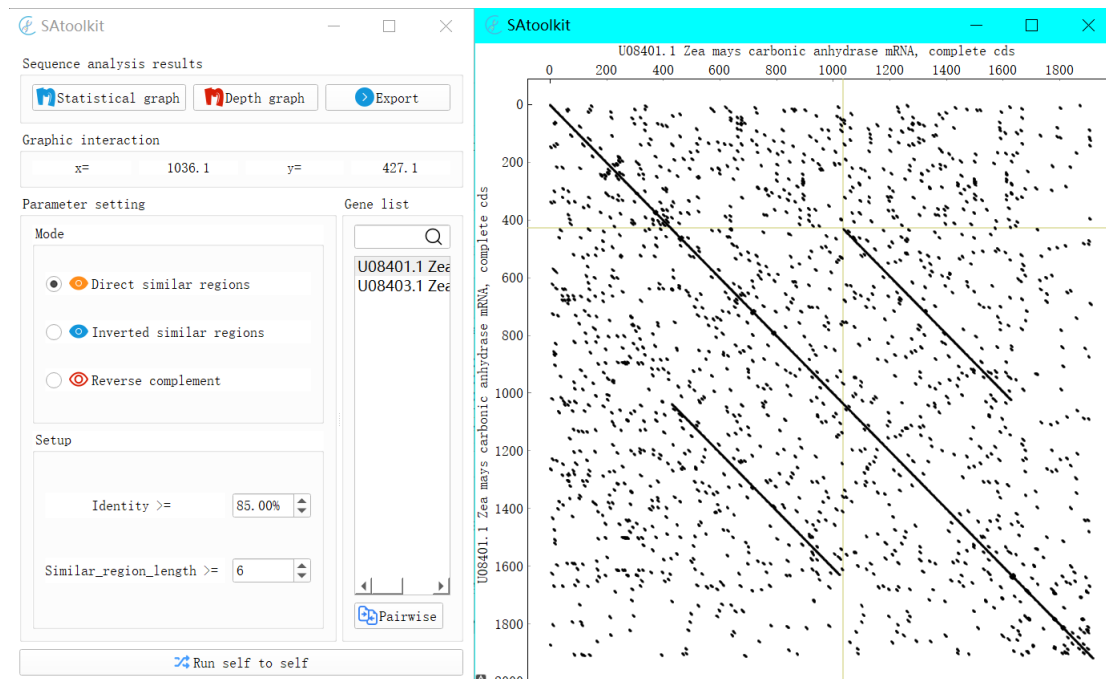


SAtoolkit integrates CyDotian algorithm (Non-exact-match and Exact match) and sliding window algorithm for sequence matching. All based on our own logic for implementation. There are three modes. The following is a detailed explanation of the software operation with the first type of Non-exact match as an example.

This mode is based on the CyDotian algorithm implemented to allow mismatched sequence matching effects in sequence matching. After the user enters the correct fasta file according to his preference, he clicks the Start button to enter the dialog box specifying the user's sequence type. DNA supports two types of matrices (BLAST and Transition-transversion matrices). Amino acids support BLOSUM45, BLOSUM62, BLOSUM80, BLOSUM90, PAM30, PAM70, and PAM250 matrices.



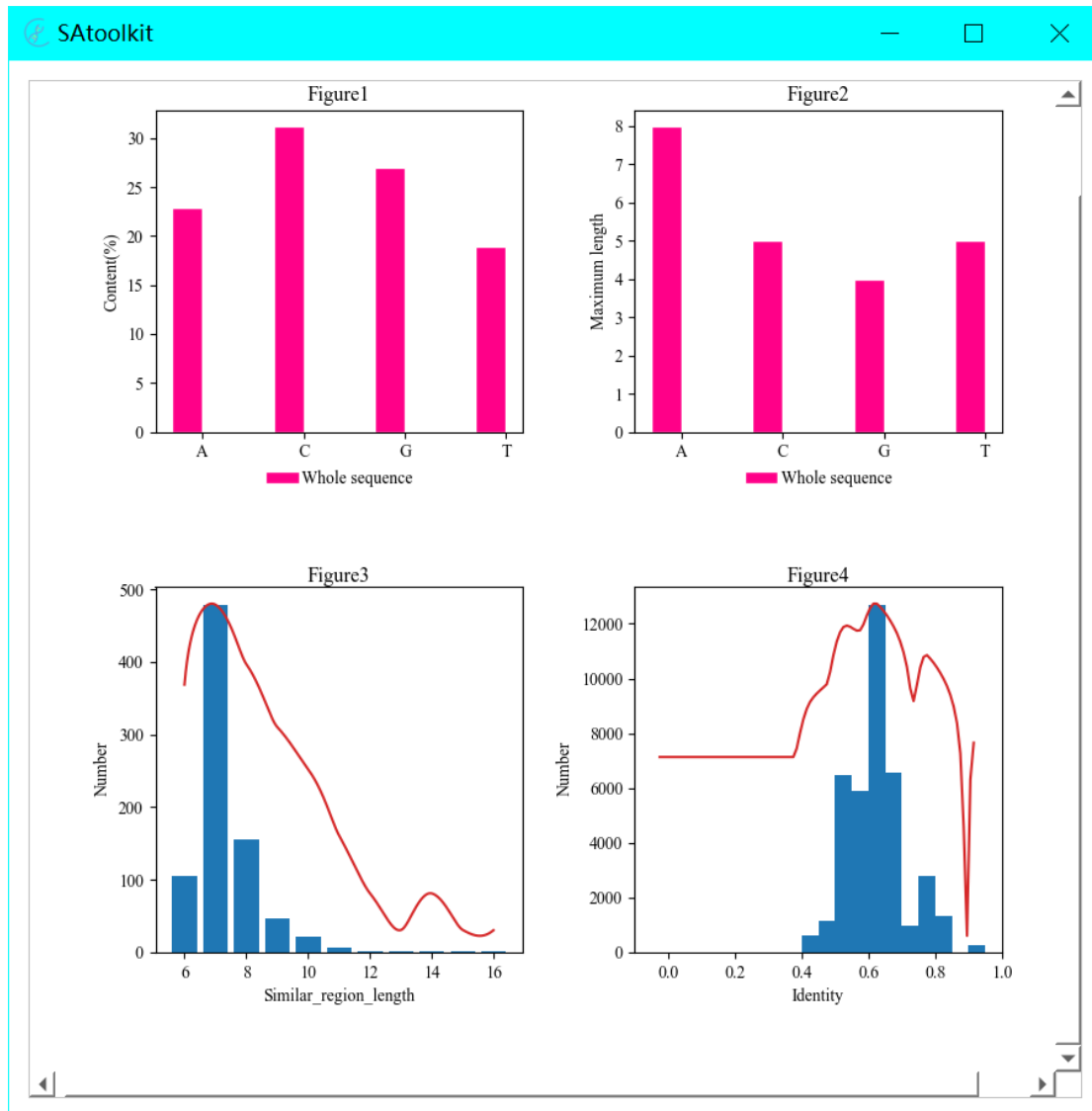
After selecting the corresponding matrix, click the Run button and a control panel (left) and a dot plot window (right) will pop up.



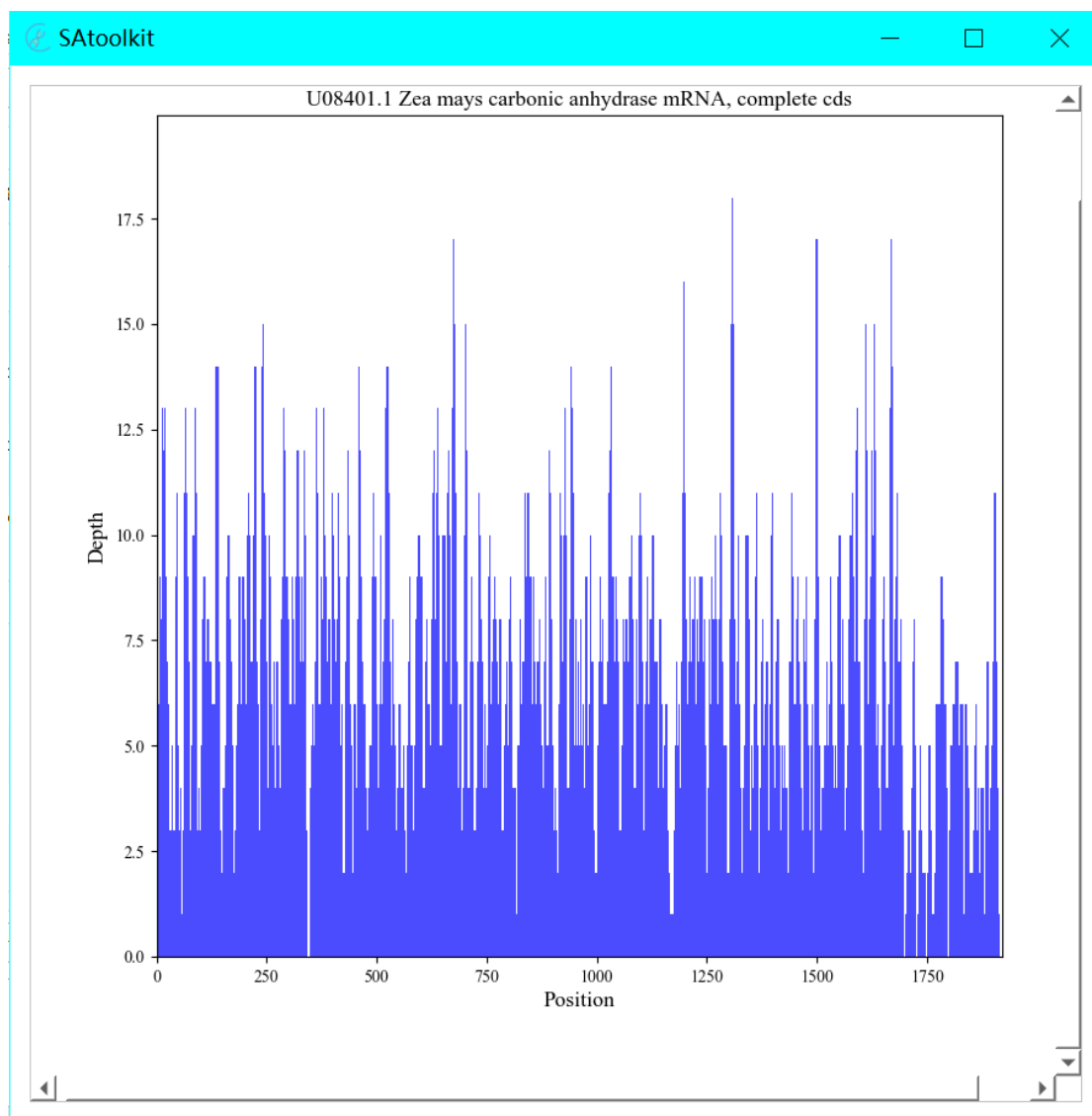
The default result is a self-to-self comparison of the first sequence in the Gene list. The control panel is divided into 4 parts: Sequence analysis results, Graphic interaction, Parameter

setting, Gene list.

There are three buttons in the Sequence analysis results section, and the Statistical graph button will bring up a popup of four statistical graphs showing the overall situation of the sequence when clicked.



The Depth graph (if available) button, when clicked, will bring up a bar graph showing the depth of the overall internal point map of the sequence.



After the Export button is clicked, a dialog box will pop up with a checkbox for the corresponding result you want to export, and then you can select the corresponding output folder path and click the Export button to export it.

The figure is a screenshot of the SAtoolkit application window showing the Export dialog box. The title bar reads "SAtoolkit". The dialog box contains the following options:

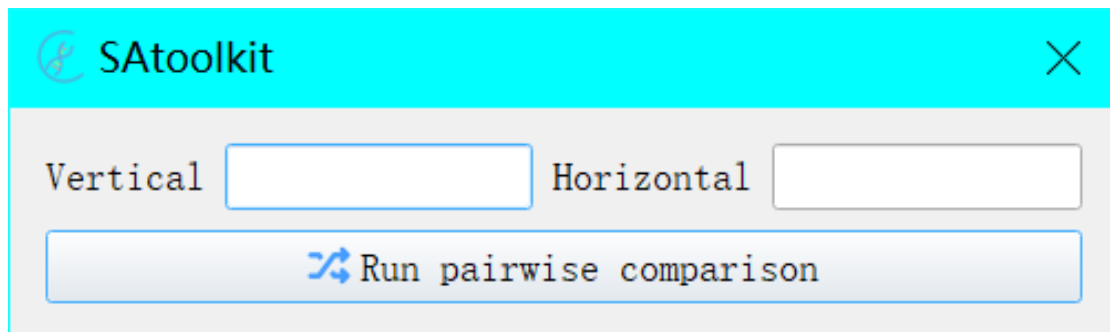
- ☒ Position
- ☒ Dot Plot
- ☒ Depth
- ☐ Figure1
- ☒ Figure2
- ☐ Figure3
- ☒ Figure4

Below the options is a text input field for the output folder path, with a folder icon on the right. At the bottom is a large button labeled "Export" with a right arrow icon.

The Graphic interaction section is used to display real-time information about the coordinates of the cross cursor position for the right dot matrix interaction.

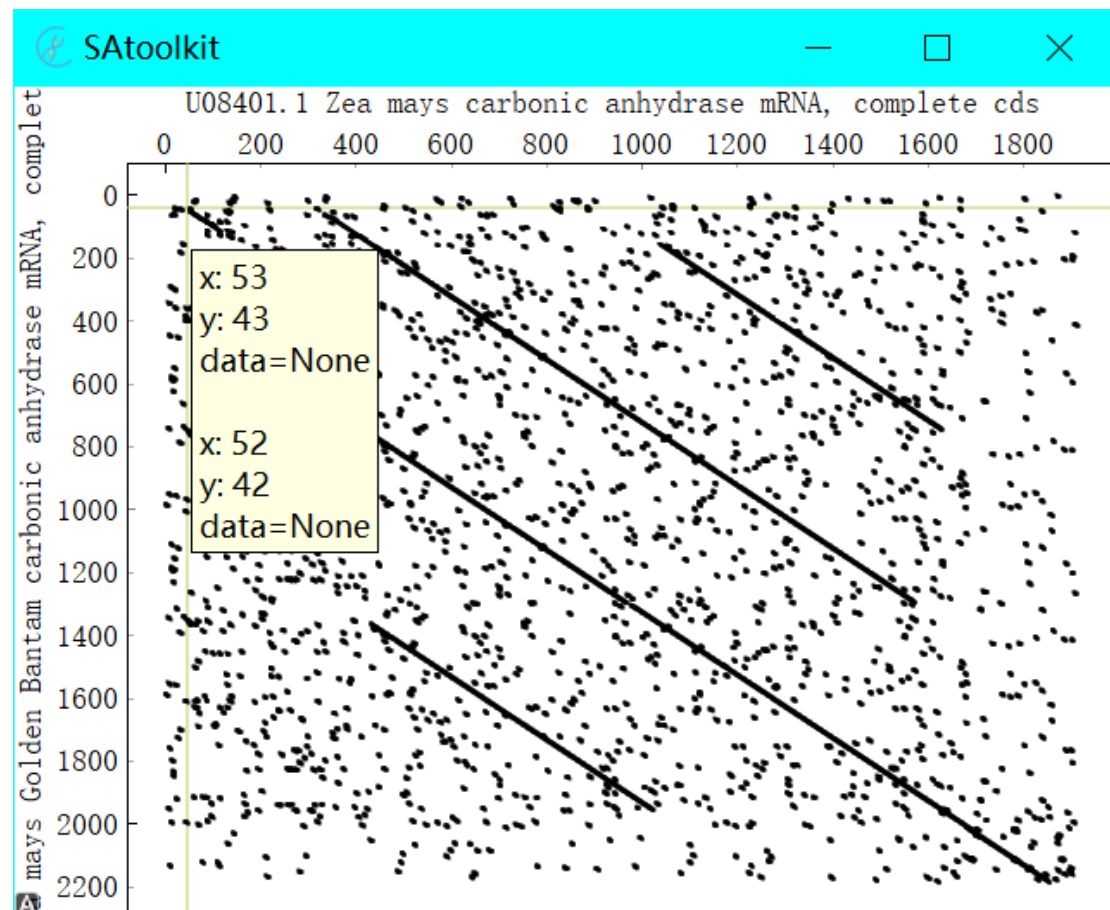
Parameter setting is an important part that affects the output results. The three controls in the Mode section of the region, from top to bottom, represent Direct repeats (finding forward similarity), Inverted repeats (finding reverse similarity) and Reverse complement (finding reverse complement) - the Reverse complement control appears only when the user inputs a DNA type. The Setup section represents the threshold set by the user, indicating that the output is Identity/Similarity greater than or equal to the current threshold and similar region length is greater than or equal to this length.

The Gene list section is used to display the list of IDs of the user input sequences, the user selects the ID and clicks the Run self to self button after setting the parameters in the Parameter setting section, then the output result will be the analysis result of the sequence corresponding to this ID. The role of the Pairwise button in this section indicates the two-by-two comparison mode respectively. When the user clicks on the Pairwise button, the user enters the two-by-two comparison mode, and the user selects the gene name in the Gene list, the gene ID selected by the user is displayed in the text box of Vertical and Horizontal knowledge (the last ID clicked by the user is placed in Horizontal and the penultimate ID clicked is placed in Vertical).



The screenshot shows a software window titled "SAtoolkit" with a blue header bar containing a logo and a close button. The main area is white and contains two input fields labeled "Vertical" and "Horizontal". Below these fields is a large button with a blue icon and the text "Run pairwise comparison".

After confirming that there are no errors, click Run pairwise comparison to run the corresponding results.



In addition, SAtoolkit has many simple and convenient small functions that users can quickly get started by characters. In a word, SAtoolkit is a very convenient tool for biological sequence comparison.