

BioAider V1.0

- 1. Introduction
- 2. Download and install
- 3. Functions
 - 3.1. SeqTools
 - 3.11. Seqformat Convertor
 - 3.12. SeqVary
 - 3.13. SequenceID Rename
 - 3.14. Split Sequence Fragment
 - 3.15. Combine Gene
 - 3.16. Visual Gene Extractor
 - 3.17. Fast Annotation
 - 3.2. Similar Analysis
 - 3.2.1 Sequence Identity Matrix
 - 3.2.2 Remove H-Similar Sequence
 - 3.3. Mutation Tools
 - 3.3.1. Mutation Analysis
 - 3.3.2. Site Counter
 - 3.3.3. Site Scree
 - 3.4 Plugin

1. Introduction

With the development of sequencing technology, a large amount of genomic sequenced data has been accumulated. Analyzing these data will help us understand their genetic variation at the molecular level. However, processing a large-scale sequence is difficult for biological or clinical experts without bioinformatics and programming skills. Besides, the needs are also diverse due to different research purposes. Therefore, simplicity of operation and diversity of function are needed.

BioAider is developed based on Python and R, which is a user-friendly GUI-interface program. As a desktop platform for genomic sequencing data studies, BioAider is designed to simplicity of operation and high summary of analysis results, which could save a lot of time for researchers.

2. Download and install

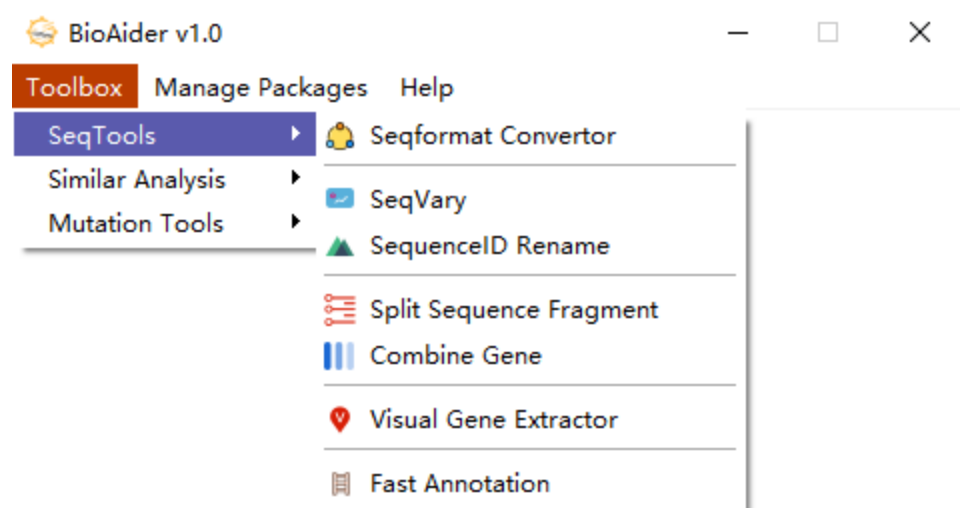
BioAider is **free available** from <https://github.com/ZhijianZhou01/BioAider/releases>.

Currently, the available version is only for windows system. After obtaining the program, users could directly run the program by clicking **BioAider.exe** without installing.

3. Functions

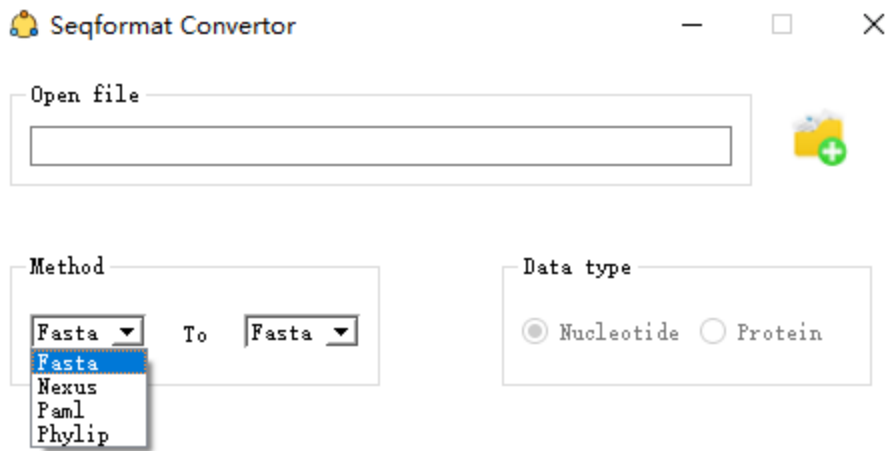
The functions of BioAider are contained in different sectors. In general, the input format of the sequence is in fasta. At present, there are 3 functional sectors.

3.1. SeqTools



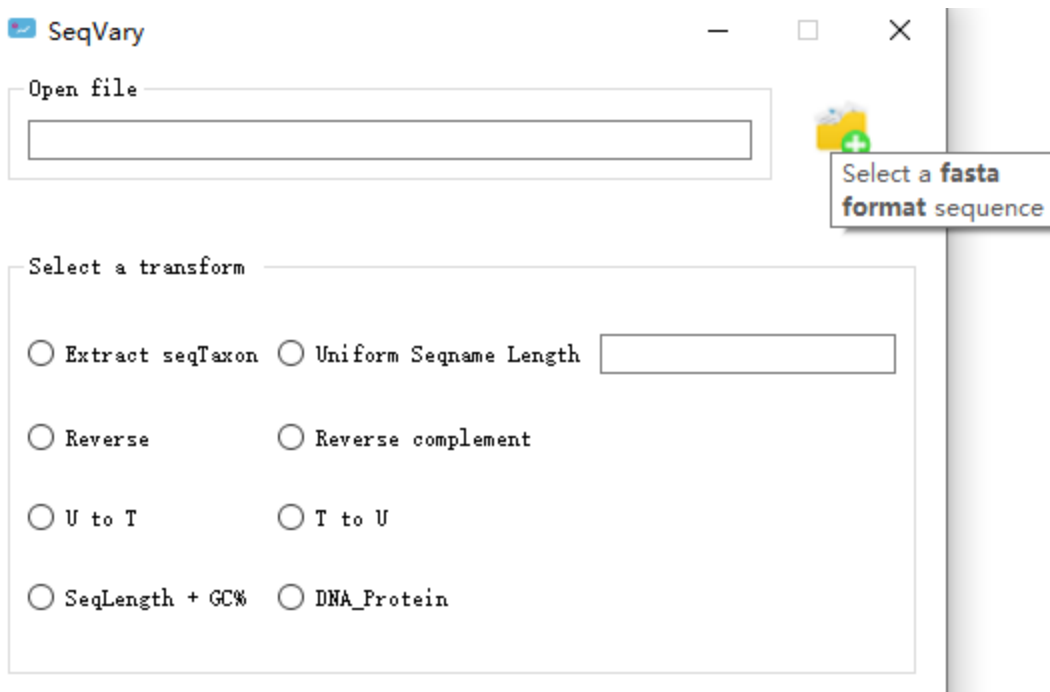
There are 6 sub-menus under the menu bar of the SeqTools Library, which are **Seqformat Convertor**, **SeqVary**, **SequenceID Rename**, **Split Sequence Fragment**, **Combine Gene**, **Visual Gene Extractor** and **Fast Annotation**.

3.11. Seqformat Convertor



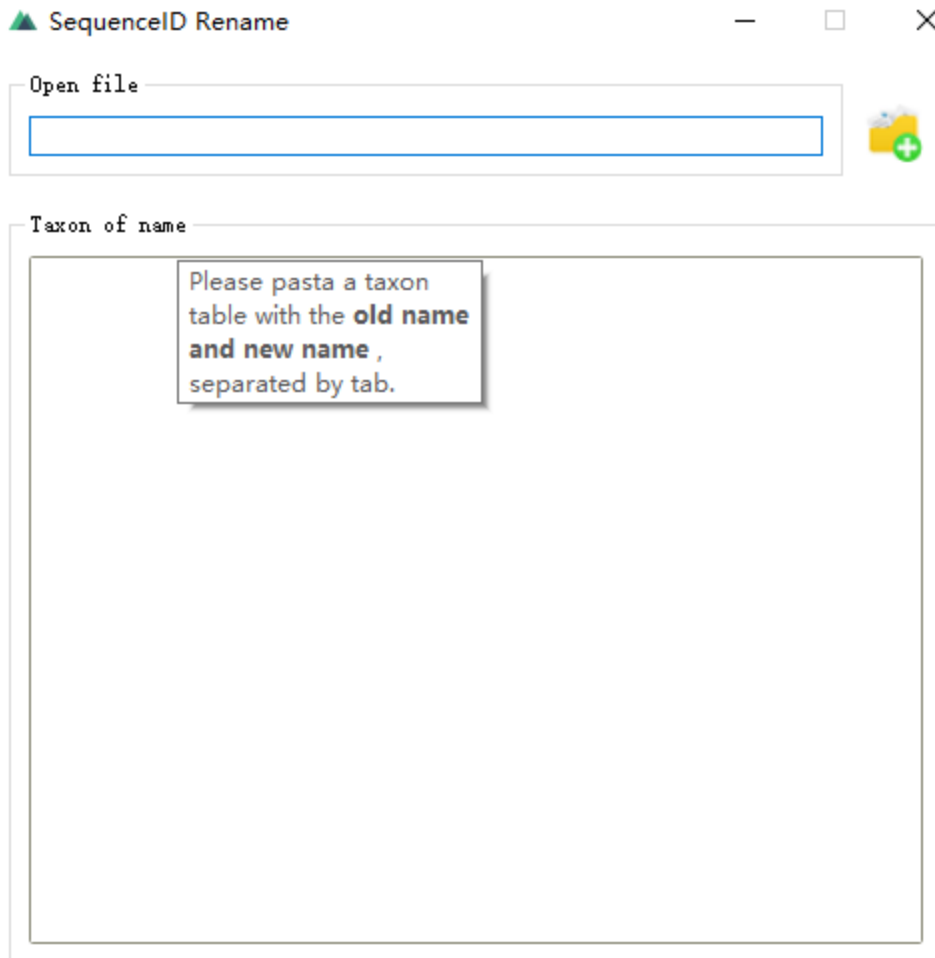
This function provides mutual conversion among several common sequence formats, which are Fasta, Nexus, Paml, and Phylip. Of note, the "Data type" option is only available when the target format is "Nexus".

3.12. SeqVary




The "SeqVary" option of BioAider provides some small functions for sequence preprocessing. "Extract seqTaxon" is used to batch extract sequence names. Note the "DNA_Protein" option requires the gene sequences datas to be aligned based on codons.


3.13. SequenceID Rename



This function can rename the original name in **sequence datas or tree file etc**. In particular, the pictures of the evolutionary tree used for publication often require the taxons of tree to follow a uniform format, so first batch replacement in the tree file saves the trouble of using vector graphics tools to modify later.

3.14. Split Sequence Fragmenet

 Split Sequence Fragment—□×

Open file

Mode of Split

☒ Different range☐ Equal range

Taxon with seqID and regional

Please pasta a taxon table with the **sequence name**, **start and end position of gene**, separated by tab.

This function can batch intercept the specified range of gene fragments, two different modes are available: specified different range (**Different range**) for each sequence, equal range for all sequences (**Equal range**).

3.15. Combine Gene

Combine Gene

Select a dir

C:/Users/zzj/Desktop/conbie

Load file

File path

C:/Users/zzj/Desktop/conbie\S.fas
C:/Users/zzj/Desktop/conbie\ORF3a.fas
C:/Users/zzj/Desktop/conbie\E.fas
C:/Users/zzj/Desktop/conbie\M.fas
C:/Users/zzj/Desktop/conbie\ORF6.fas
C:/Users/zzj/Desktop/conbie\ORF7a.fas
C:/Users/zzj/Desktop/conbie\ORF7b.fas
C:/Users/zzj/Desktop/conbie\ORF8.fas
C:/Users/zzj/Desktop/conbie\N.fas
C:/Users/zzj/Desktop/conbie\ORF10.fas

This function is used to concatenate multiple gene sequences into one. Users can first put different genes dataset files into the same folder, and then drag the folder into the **inputbox**, then click the **"Load file button"** import the file path of each genes datasets into **textbox**. Users can modify the order of genes by manually adjusting the sort of file path in the **textbox**. Of note, all the sequences should be in fasta format.

3.16. Visual Gene Extractor

Upload fasta file

sequence.fasta
browse files

Input save folder path

Gene name

ORF1ab
ORF1ab polyprotein

Run

Visual Gene Extractor

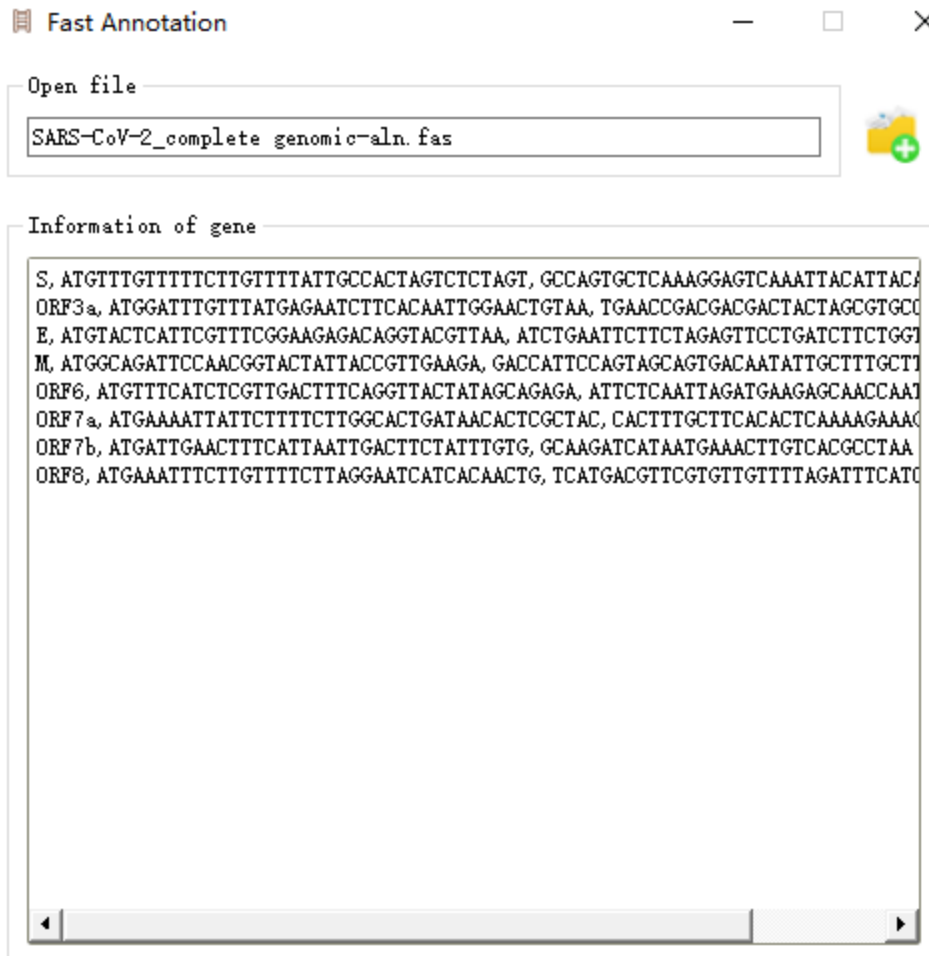
You can view the sequence name here

```
[
  0 :
  "MT419849.1_cds_QJI54274.1_1 [gene=ORF1ab] [protein=ORF1ab polyprotein] [frame=2]
  [partial=5'] [exception=ribosomal slippage] [protein_id=QJI54274.1]
  [location=join(<95..13292,13292..21379)] [gbkey=CDS] "
  1 :
  "MT419849.1_cds_QJI54275.1_2 [gene=ORF1ab] [protein=ORF1a polyprotein] [frame=2]
  [partial=5'] [protein_id=QJI54275.1] [location=<95..13307] [gbkey=CDS] "
  2 :
  "MT419849.1_cds_QJI54276.1_3 [gene=S] [protein=surface glycoprotein]
  [protein_id=QJI54276.1] [location=21387..25208] [gbkey=CDS] "
  3 :
  "MT419849.1_cds_QJI54277.1_4 [gene=ORF3a] [protein=ORF3a protein]
  [protein_id=QJI54277.1] [location=25217..26044] [gbkey=CDS] "
  4 :
  "MT419849.1_cds_QJI54278.1_5 [gene=E] [protein=envelope protein]
  [protein_id=QJI54278.1] [location=26069..26296] [gbkey=CDS] "
```

This function is used to extract the specified gene sequences from mixed coding gene sequence set which was downloaded from NCBI nucleotide database. Given that the same gene may have different manifestations in different studies, the **textbox of Gene name** could

enter multiple names, and BioAider will extract the corresponding gene sequence which contain these gene names.

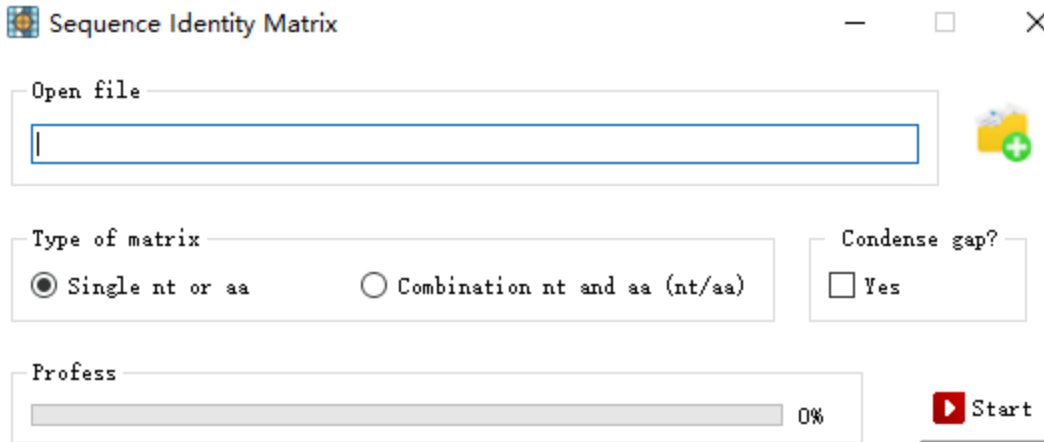
3.17. Fast Annotation



For these strain sequences from the same or highly related species, their nucleotide identity is usually relatively higher. Therefore, the sequences annotation could be based on the gene information of the reference sequence after multi-sequence alignment. BioAider provides a quickly sequence annotation function, users can import the aligned gene sequence set (fasta format file), and adjust the reference sequence for annotation to the forefront of the file. Paste the gene information of reference sequence, name, starting string and end string into the textbox, separated by ",". Then batch abstract genes. Note that the start string or end string of the gene is not limited in length, but it is required to be unique in the reference sequence. Besides, the higher of similarity among sequences, the higher accuracy of the annotation.

3.2. Similar Analysis

3.2.1 Sequence Identity Matrix



Sequence Identity Matrix

Open file

Type of matrix

☒ Single nt or aa ☐ Combination nt and aa (nt/aa)

Condense gap?

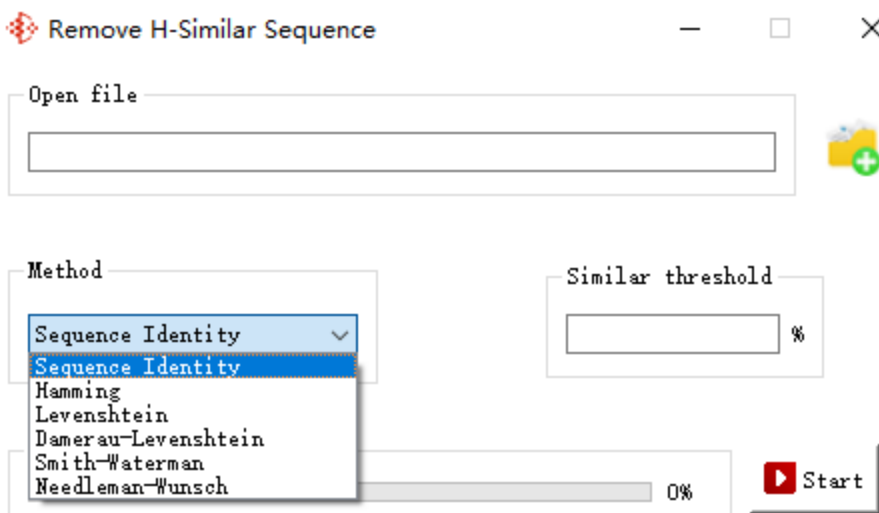
☐ Yes

Progress 0%

Start

By inputting the aligned sequence datasets in fasta format, and a pairwise sequence identity matrix can be generated. This function contains two different modes: nucleotide or amino acid sequence identity matrix (Single nt or aa), nucleotide plus amino acid sequence identity matrix (Combination nt and aa). It should be noted that if the "Combination nt and aa" is selected, the inputted sequences should be aligned based on codon method. In order to better fit the variation characteristics, BioAider provides the "Condense gap" function. If the option was selected, the program will treat every 3 consecutive inserted or deleted bases as one.

3.2.2 Remove H-Similar Sequence



Remove H-Similar Sequence

Open file

Method

Sequence Identity

Sequence Identity

Hamming

Levenshtein

Damerau-Levenshtein

Smith-Waterman

Needleman-Wunsch

Similar threshold

%

Progress 0%

Start

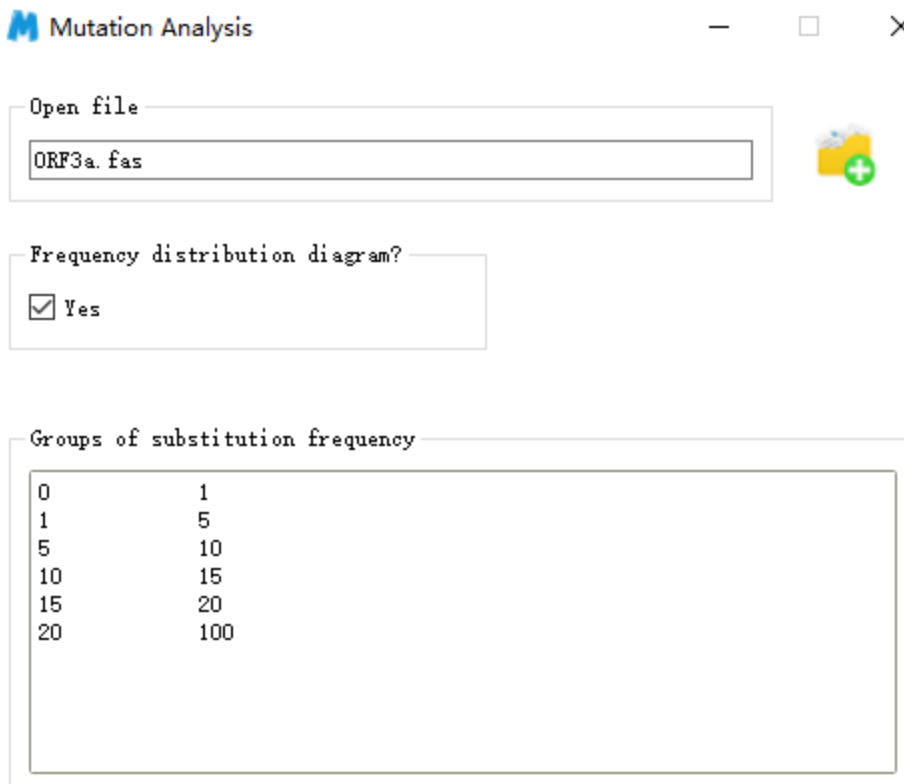
This function could remove highly similar sequences and keep one by specifying the threshold of similarity (Similar threshold). BioAider provides 6 different methods for calculating the similarity of sequences. It should be noted that the "Sequence Identity" and "Hamming" methods require the input sequences data are aligned, and we suggest that the sequences

datasets for remaining 4 methods better not be pre-aligned, because these algorithm own alignment function. If the **Similar threshold** is seted to 100, the function of excluding repeated sequences will be turned on.

If you want to obtain the sequence similarity matrix calculated by the above 6 methods, you can click the **right button of mouse** in any region of the program interface to call up the functional menu.

3.3. Mutation Tools

3.3.1. Mutation Analysis



Mutation Analysis

Open file

ORF3a.fas

Frequency distribution diagram?

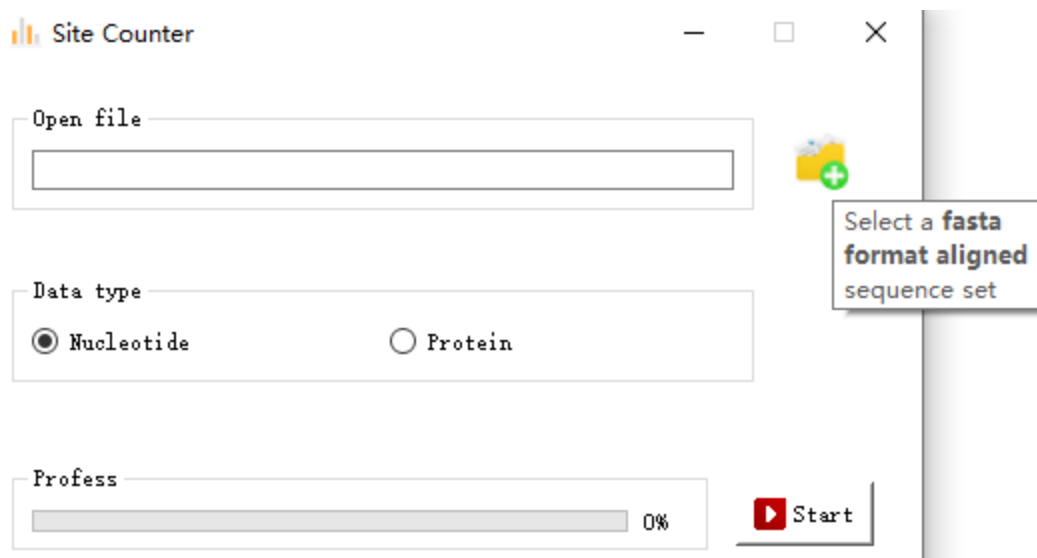
☒ Yes

Groups of substitution frequency

0	1
1	5
5	10
10	15
15	20
20	100

This function is used for analyzing the mutations characteristicson of single or combined coding gene. BioAider will scan each condon sites in aligned sequence datasets, and identifies the type of mutation, including synonymous, non-synonymous, insertions and deletions. Finally, BioAider will automatically summarize and output the relevant analysis results. It should be noted that the input sequence is required to be pre-aligned based on standard codon method. The each line of **Groups of substitution frequency** represents one group of substitution frequency for synonymous or nonsynonymous sites , and start value and end value are separated by tab symbol. Note, the start value of each group is not included in the range of frequency. Especially, the substitution frequency distribution image is drawn by R program, so you have to make sure the **Rscript.exe** have been added into BioAider throung the plugin function.

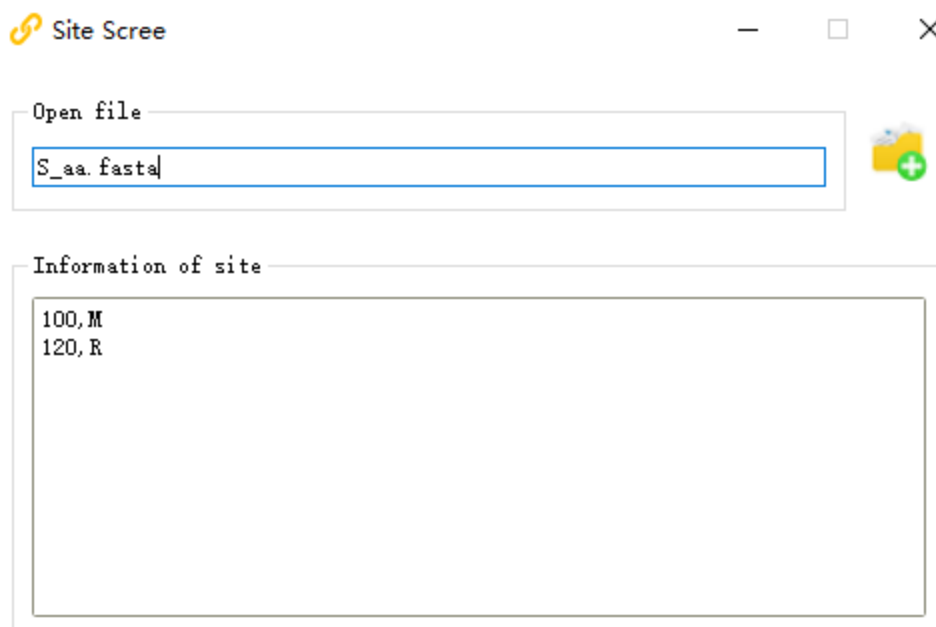
3.3.2. Site Counter



The Site Counter interface includes a title bar with a bar chart icon and the text 'Site Counter'. It features three main sections: an 'Open file' section with a text input field and a folder icon; a 'Data type' section with two radio buttons labeled 'Nucleotide' (selected) and 'Protein'; and a 'Progress' section with a progress bar at 0% and a red 'Start' button. A tooltip on the right side of the folder icon reads: 'Select a **fasta** format aligned sequence set'.

This function could summary the type, count and proportion of bases (or amino acids) at each site for the aligned sequence datasets. In addition, BioAider will output a consensus sequence based on the highest proportion base (or amino acid) in each site.

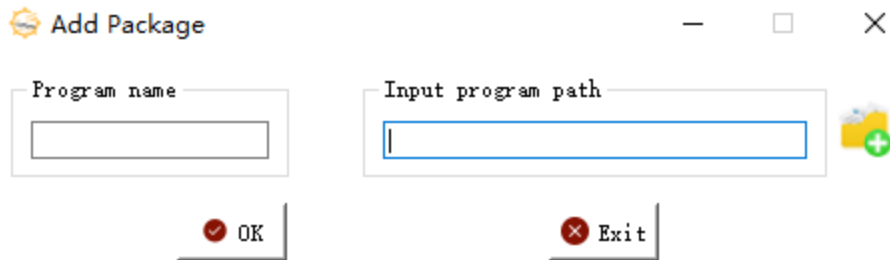
3.3.3. Site Scree



The Site Scree interface includes a title bar with a chain link icon and the text 'Site Scree'. It features two main sections: an 'Open file' section with a text input field containing 'S_aa.fasta' and a folder icon; and an 'Information of site' section with a large text area containing the text '100, M' and '120, R'.

This function is used to extract the sequences with corresponding base or amino acid in specified one or more site. It is very useful for studying whether there is linkage inheritance among different gene sites.

3.4 Plugin



BioAider opens the plug-in function, users can add executable programs into BioAider by click the menu bar of **Manage Packages** and select **Add packages**.