Manual of BioAider V1.423

Updated in January 26, 2024

A richly featured desktop platform for analysis of bioinformatics data

Written by Zhou ZJ

Home page: https://github.com/ZhijianZhou01/BioAider

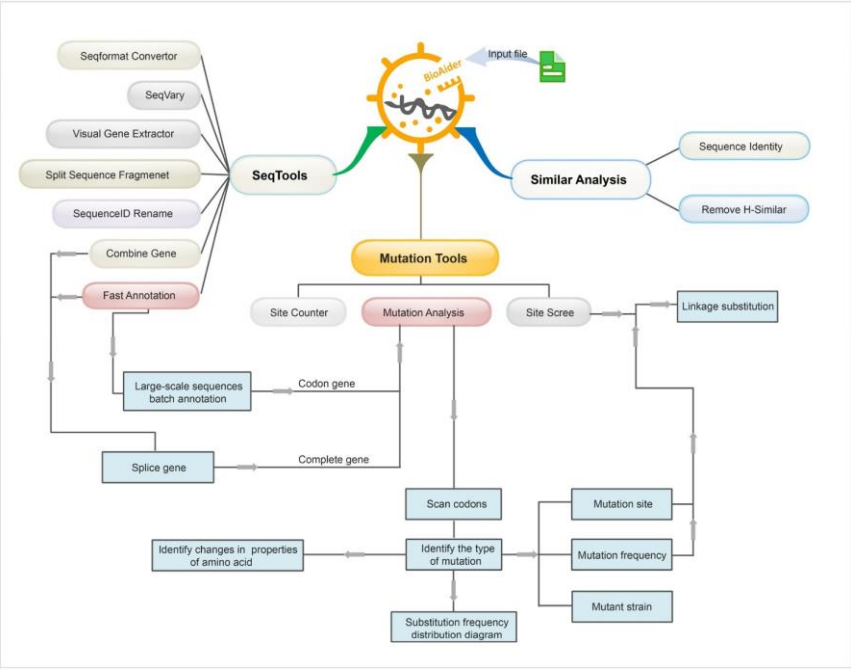
Version 1.423 || Mar 24, 2022

Content

1. Introduction

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

BioAider is developed based on Python3, which is a user-friendly program with GUI-interface. As a desktop platform, the design concept of BioAider is that simplicity of operation and high summary of analysis results, which could save a lot of time for researchers.



2. Download and install

BioAider and all the updated versions is freely available for non-commercial user at https://github.com/ZhijianZhou01/BioAider/releases. After obtaining the program,

users could directly run the program in Windows, MacOS or Linux (Ubuntu 16.04 or more) system without installation.

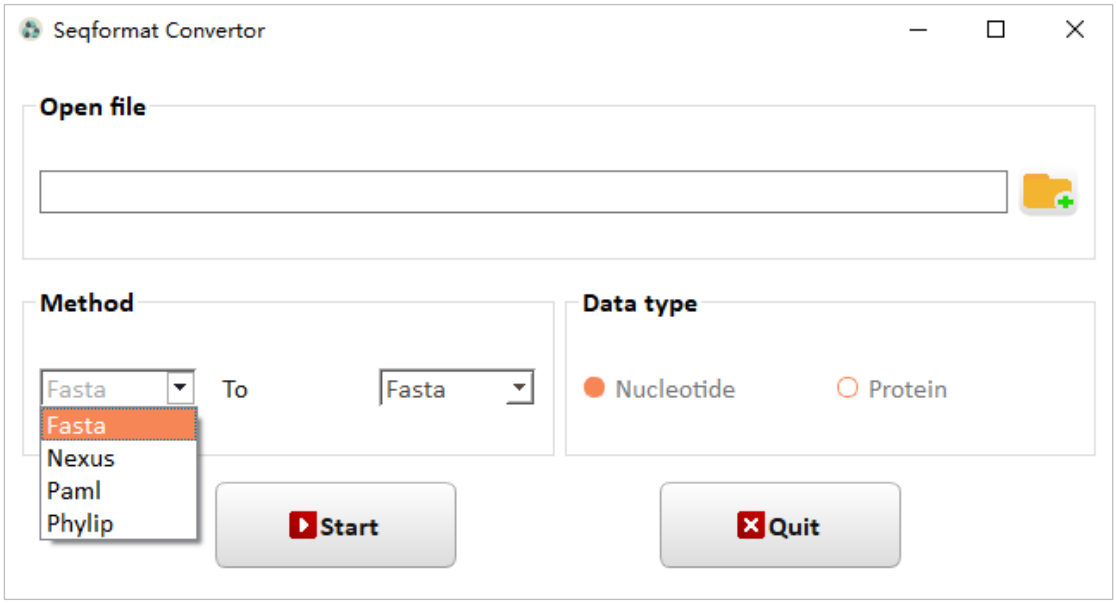
BioAider will be in the long-term update, this document briefly introduces some of its current commonly functions. In V1.423, we've beautified the interface again and added a variety of interface themes to make BioAider easier to use.

3. Main

3.1. *SeqTools*

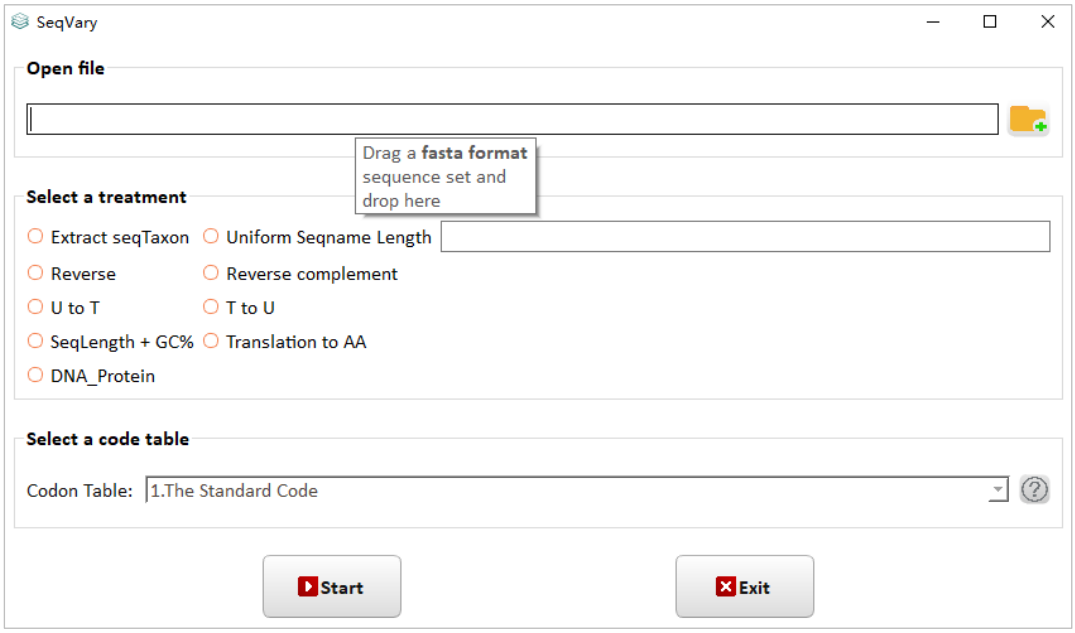
3.1.1. *Seqformat Convertor*

BioAider 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".



3.1.2. *SeqVary*

The ***"SeqVary"*** option of BioAider provides some small functions for sequence preprocessing. For example, ***"SeqLength+GC%"*** is used to batch calculate sequence length and content of GC.

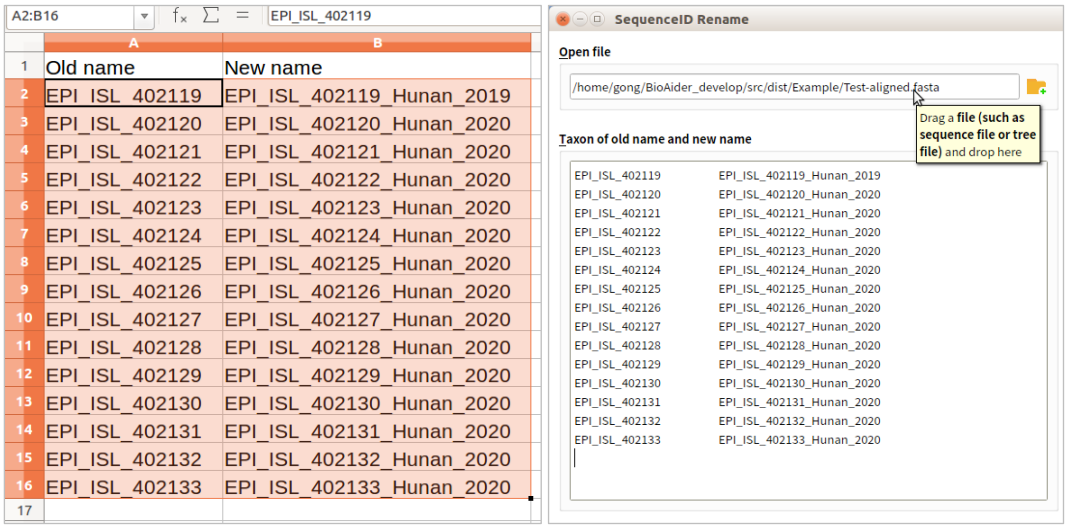


Note: the ***"DNA_Protein"*** option requires the gene sequences data to be aligned based on codons.

3.1.3. *SequenceID Rename*

BioAider could rename the original name in **sequence data or tree file etc**. In particular, the pictures of the evolutionary tree used for publication often require the taxon 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.

First, make a table of **new and old names** in a table editor, then copy and paste them into BioAider:



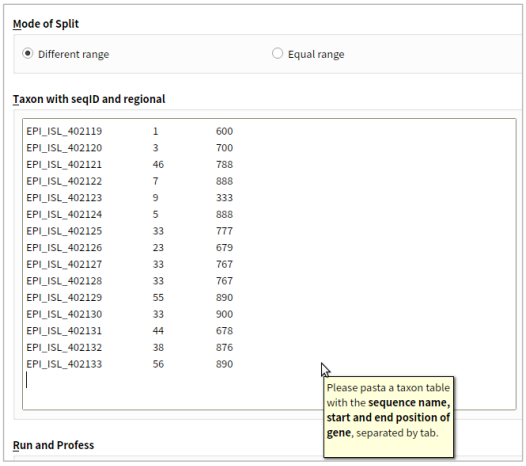
Generally speaking, as long as the input file is a text file, BioAider could successfully perform this work.

3.1.4. *Split Sequence Fragmenet*

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"***).

If you want to use the ***"Different range"*** to split for each sequence, make a table of start and end location firstly, then copy and paste them into BioAider:

	A	B	C
1	Name	Start	End
2	EPI_ISL_402119	1	600
3	EPI_ISL_402120	3	700
4	EPI_ISL_402121	46	788
5	EPI_ISL_402122	7	888
6	EPI_ISL_402123	9	333
7	EPI_ISL_402124	5	888
8	EPI_ISL_402125	33	777
9	EPI_ISL_402126	23	679
10	EPI_ISL_402127	33	767
11	EPI_ISL_402128	33	767
12	EPI_ISL_402129	55	890
13	EPI_ISL_402130	33	900
14	EPI_ISL_402131	44	678
15	EPI_ISL_402132	38	876

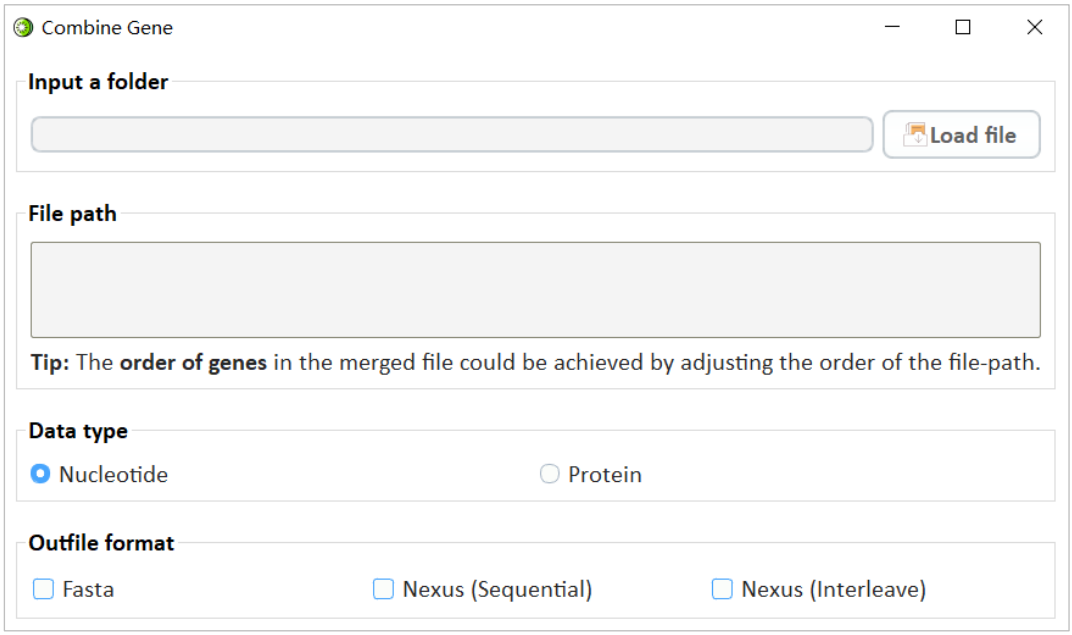


If users choose the options of ***"Equal range"***, BioAider will split all the sequences according to the same specified range.

3.1.5. *Combine Gene (Tandem Gene)*

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***.

It should be pointed out that **the sequence names in different gene data sets should be consistent**, otherwise BioAider cannot be associated with them, but BioAider allows some data in a certain gene dataset to be missing and will represent them by gaps ("-").



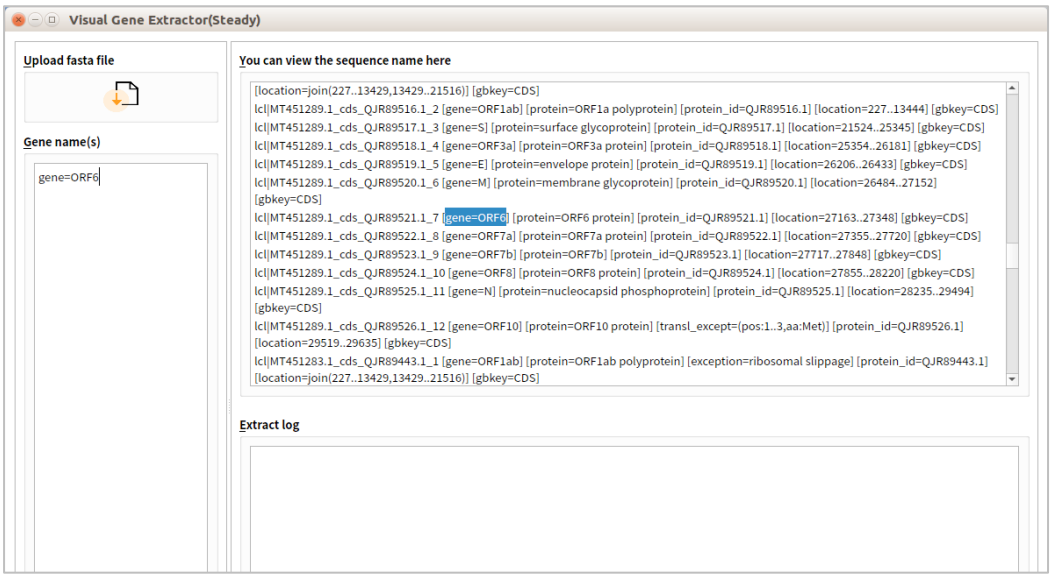
Note, users can **modify the order of genes** in tandemy sequence by adjusting the sort of inputfile path in the ***textbox.*** **Of note**, all the sequences which are used for combined should be fasta format.

3.1.6. *Visual Gene Extractor*

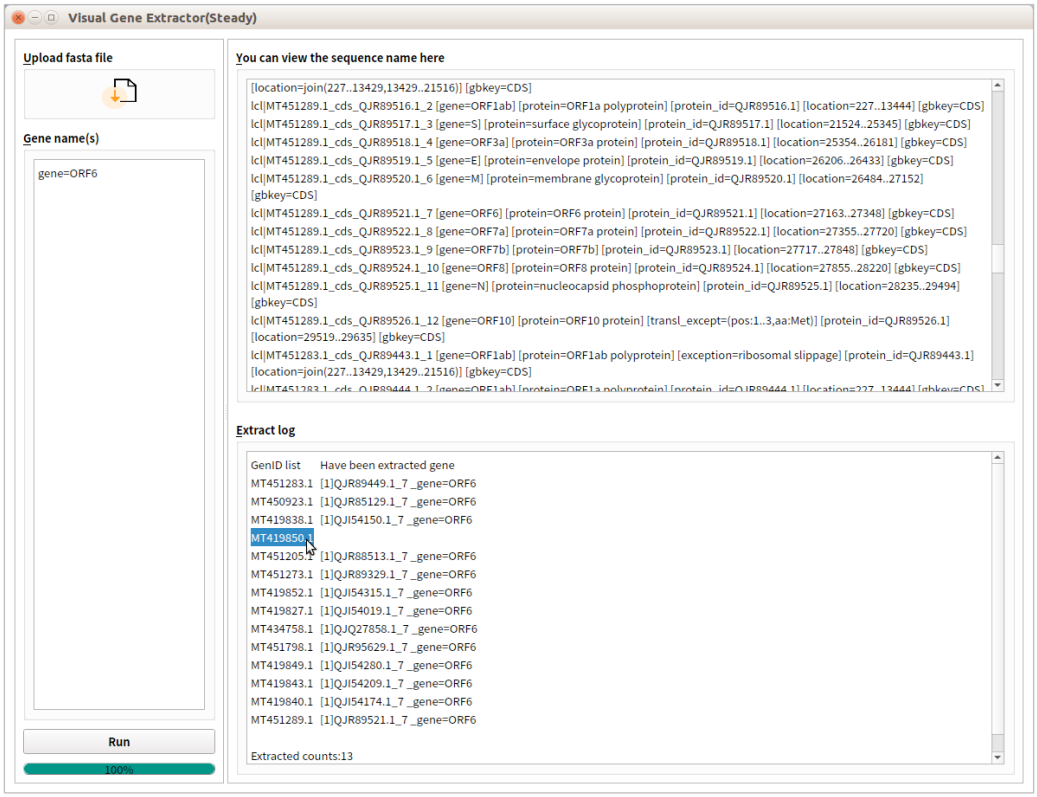
This function is used to extract the sequences included specified gene from mixed coding gene sequence set, especially when these sequences data are downloaded from NCBI database. Given that the same gene may have different manifestations in different studies, the textbox of ***"Gene name"*** could enter multiple names, and BioAider will extract the corresponding gene sequence which contain these gene names.

Next, we demonstrate the use of ***Visual Gene Extractor(Steady)***.

Example 1 (The sequences datas are directly downloaded from NCBI database, including some gene fragments of SARS-CoV-2):



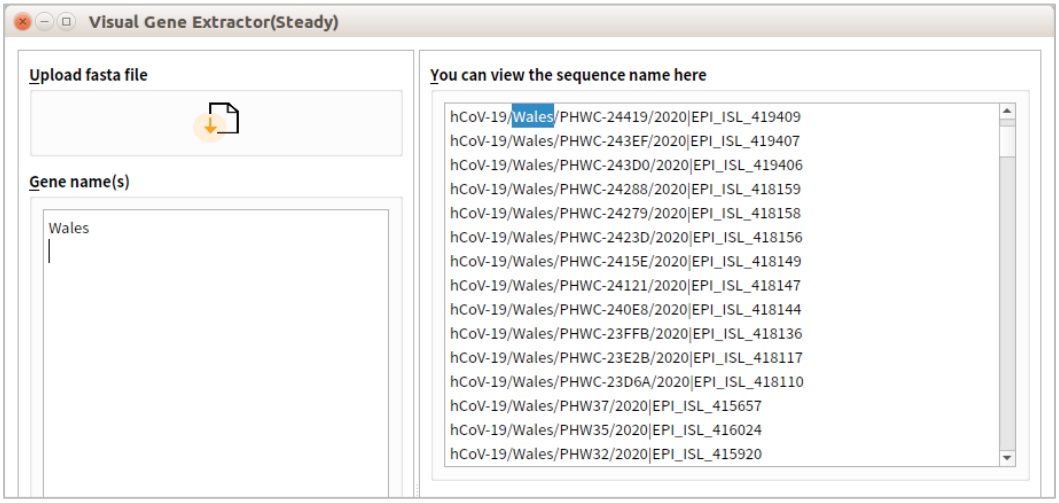
After uploading the sequence to BioAidrs as above, then we extract ORF6 gene sequence of SARS-CoV-2. Input a string **containing at least the gene name** to textbox of ***"Gene name(s)"***, then click button of ***"Run"***, run log as follows:



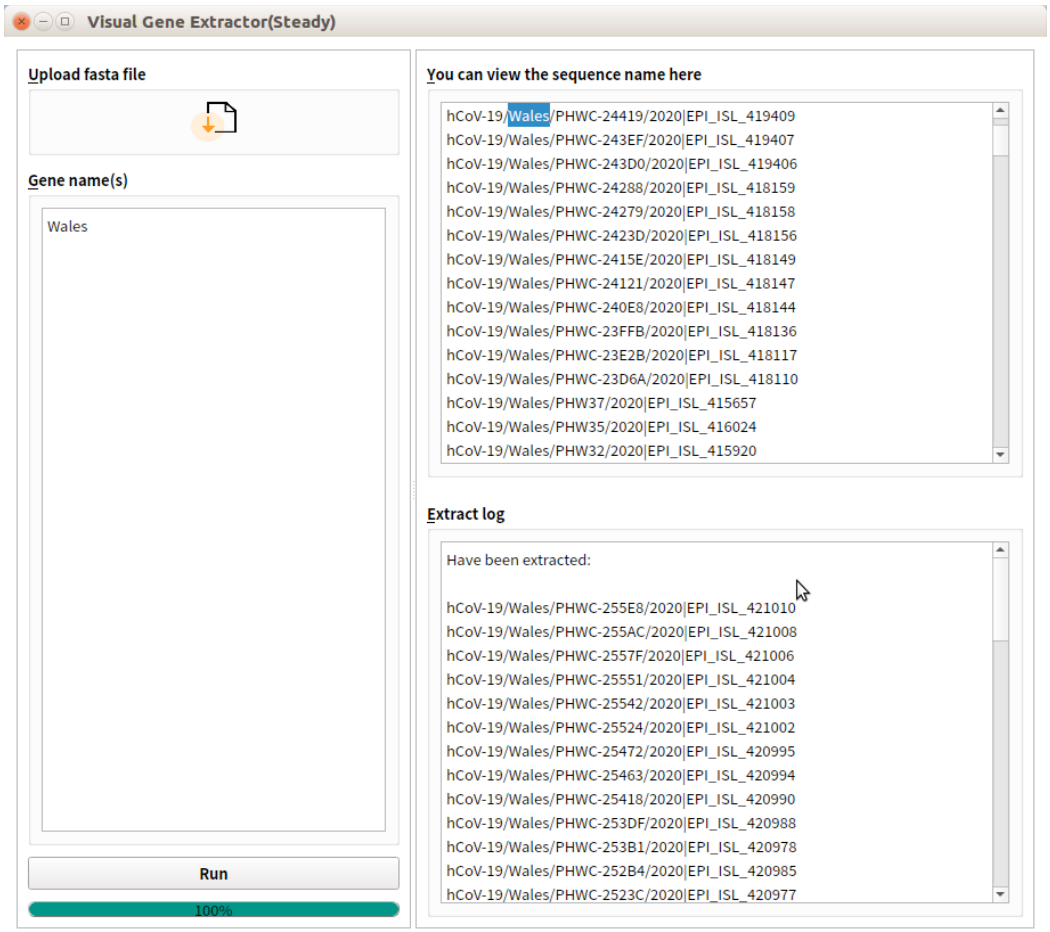
If some gene fragments are not extracted (as shown by the arrow), the possible reason is that the strain does not contain this gene fragment or the gene owns other names in

some sequences. If it is the second case, you can append other names of this gene to the next line of ***"Gene name(s)"*** textbox.

Example 2 (Arbitrary fasta sequence):



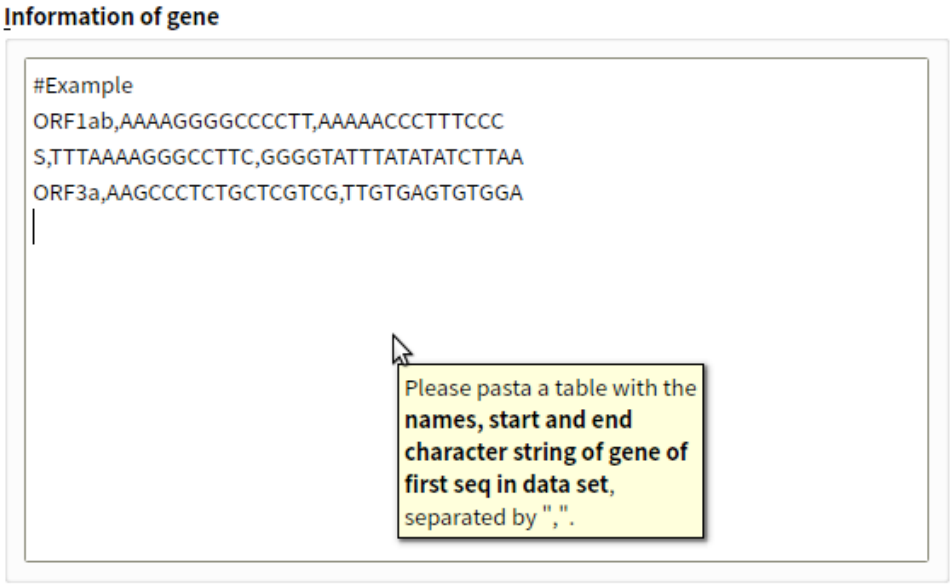
As shown above, if you want to extract these sequences which containing the tags of ***"Wales"***, BioAider could accomplish it well.



The extracted sequence will be saved in the directory where the input file is located.

3.1.7. *Fast Annotation*

For different strain sequences from the same virus, 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.

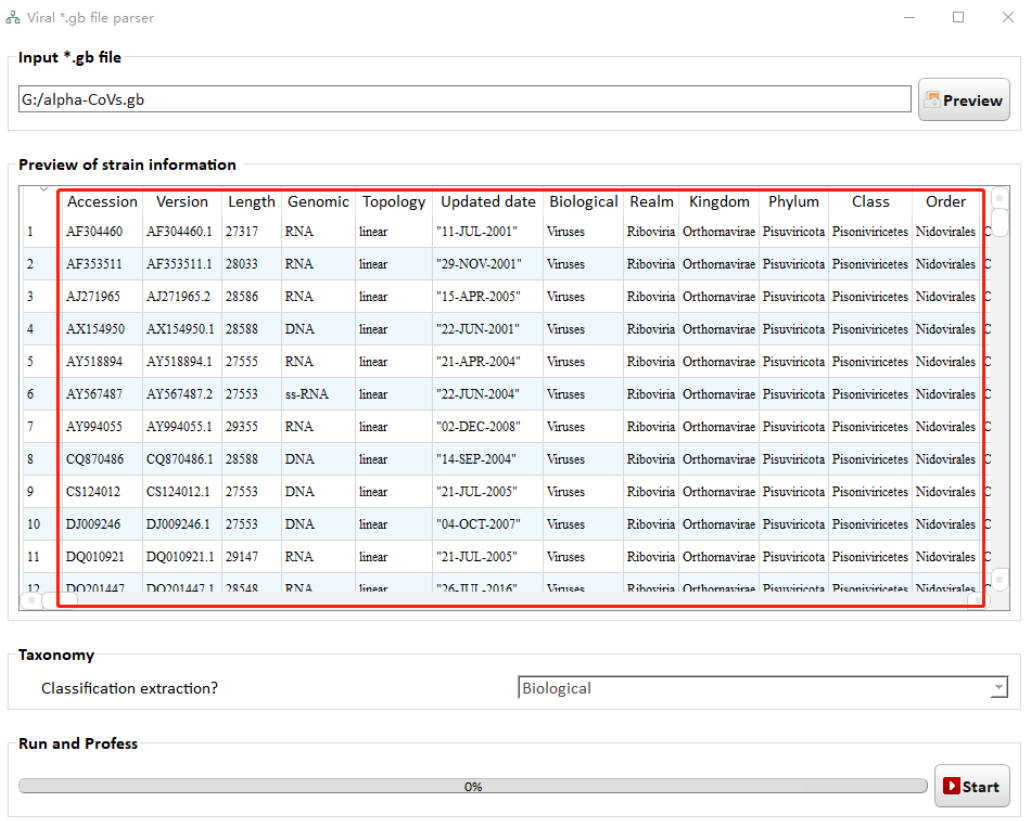


BioAider provides a quickly sequence annotation function, users can import the aligned complete genome sequence set (*.fasta format file), and adjust the reference sequence for annotation to the forefront of the file. Paste the gene information of reference sequence in aligned sets, name, starting string 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.

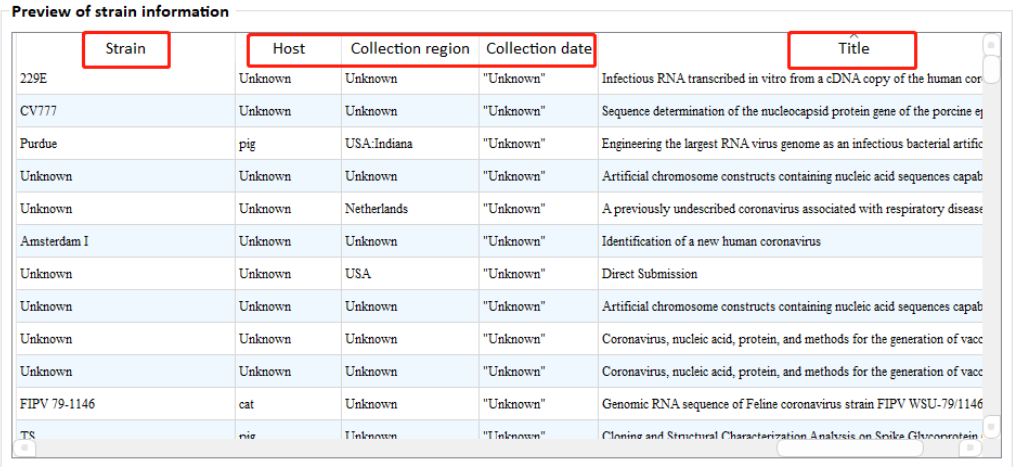
*3.1.8. Vrial *.gb file parser*

A simple function is used to **parse the *.gb file of the virus in a batch method**, then extract the information. Especially, BioAider make relevant optimizations for the coronavirus in the taxonomic unit：



	Accession	Version	Length	Genomic	Topology	Updated date	Biological	Realm	Kingdom	Phylum	Class	Order
1	AF304460	AF304460.1	27317	RNA	linear	"11-JUL-2001"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
2	AF353511	AF353511.1	28033	RNA	linear	"29-NOV-2001"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
3	AJ271965	AJ271965.2	28586	RNA	linear	"15-APR-2005"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
4	AX154950	AX154950.1	28588	DNA	linear	"22-JUN-2001"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
5	AY518894	AY518894.1	27555	RNA	linear	"21-APR-2004"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
6	AY567487	AY567487.2	27553	ss-RNA	linear	"22-JUN-2004"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
7	AY994055	AY994055.1	29355	RNA	linear	"02-DEC-2008"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
8	CQ870486	CQ870486.1	28588	DNA	linear	"14-SEP-2004"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
9	CS124012	CS124012.1	27553	DNA	linear	"21-JUL-2005"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
10	DJ009246	DJ009246.1	27553	DNA	linear	"04-OCT-2007"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
11	DQ010921	DQ010921.1	29147	RNA	linear	"21-JUL-2005"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales
12	DQ201447	DQ201447.1	28548	RNA	linear	"26-JUL-2016"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales

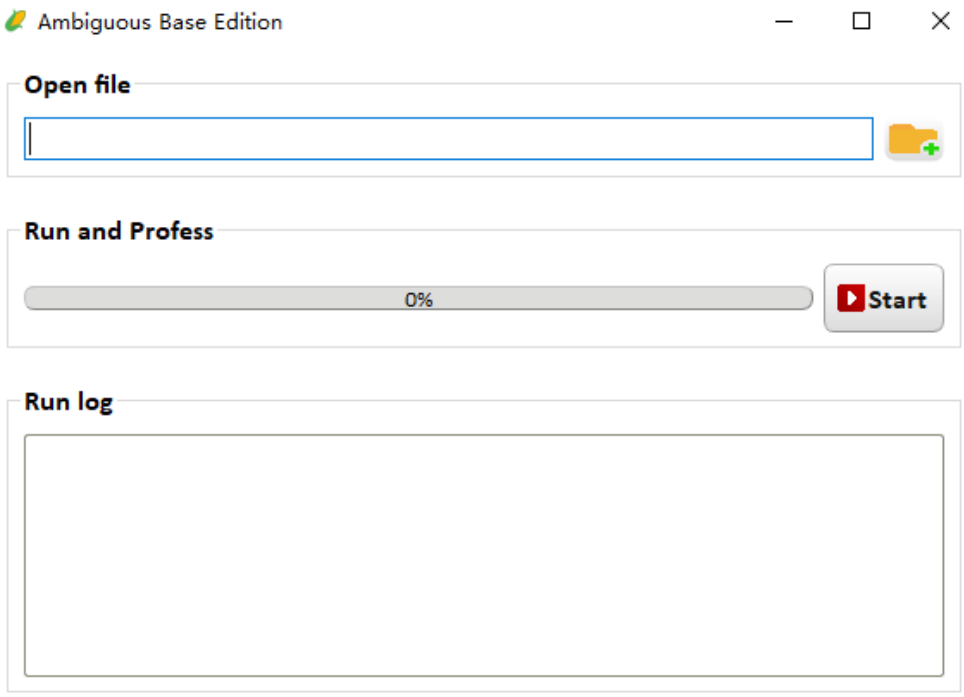
Additional information, such as **host, date and location of sampling, and even published literature**, can be quickly obtained:



Strain	Host	Collection region	Collection date	Title
229E	Unknown	Unknown	"Unknown"	Infectious RNA transcribed in vitro from a cDNA copy of the human cor
CV777	Unknown	Unknown	"Unknown"	Sequence determination of the nucleocapsid protein gene of the porcine ej
Purdue	pig	USA:Indiana	"Unknown"	Engineering the largest RNA virus genome as an infectious bacterial artific
Unknown	Unknown	Unknown	"Unknown"	Artificial chromosome constructs containing nucleic acid sequences capab
Unknown	Unknown	Netherlands	"Unknown"	A previously undescribed coronavirus associated with respiratory disease
Amsterdam I	Unknown	Unknown	"Unknown"	Identification of a new human coronavirus
Unknown	Unknown	USA	"Unknown"	Direct Submission
Unknown	Unknown	Unknown	"Unknown"	Artificial chromosome constructs containing nucleic acid sequences capab
Unknown	Unknown	Unknown	"Unknown"	Coronavirus, nucleic acid, protein, and methods for the generation of vacc
Unknown	Unknown	Unknown	"Unknown"	Coronavirus, nucleic acid, protein, and methods for the generation of vacc
FIPV 79-1146	cat	Unknown	"Unknown"	Genomic RNA sequence of Feline coronavirus strain FIPV WSU-79/1146
TS	pig	Unknown	"Unknown"	Cloning and Structural Characterization Analysis on Spike Glycoprotein

3.1.9. Correction ambiguous bases

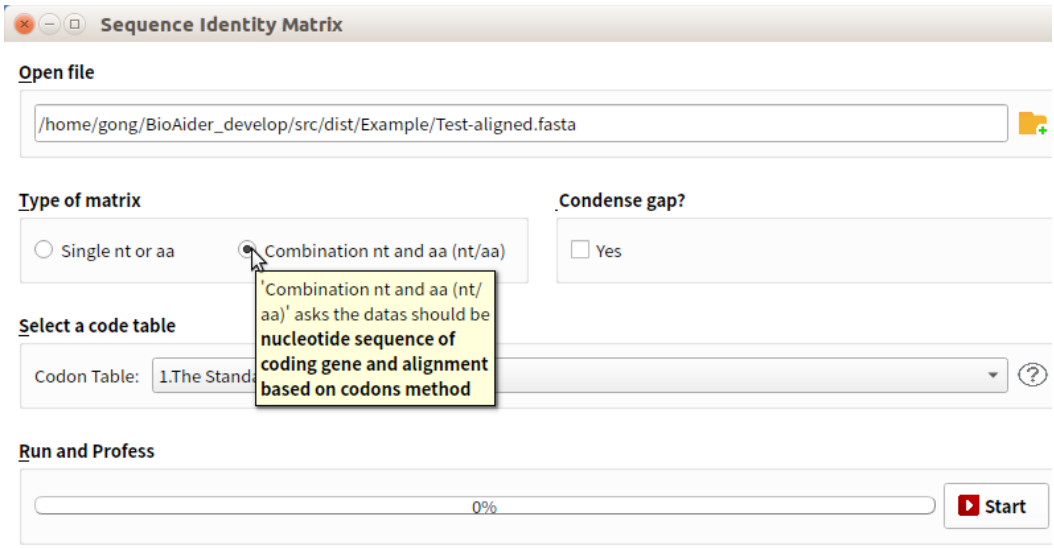
In some scenarios based on multiple sequence analysis, ambiguous bases may cause an impact, such as PAML-based selection pressure analysis. For multiple sequences aligned by codon method, BioAider could correct ambiguous bases:



3.2. *Similar Analysis*

3.2.1. *Sequence Identity Matrix*

By inputting the aligned sequence data in *.fasta format, and a pairwise sequence identity matrix can be generated. This function contains two different modes: identity matrix for single nucleotide or amino acid (***"Single nt or aa"***), identity matrix for combination nucleotide and amino acid (***"Combination nt and aa"***).

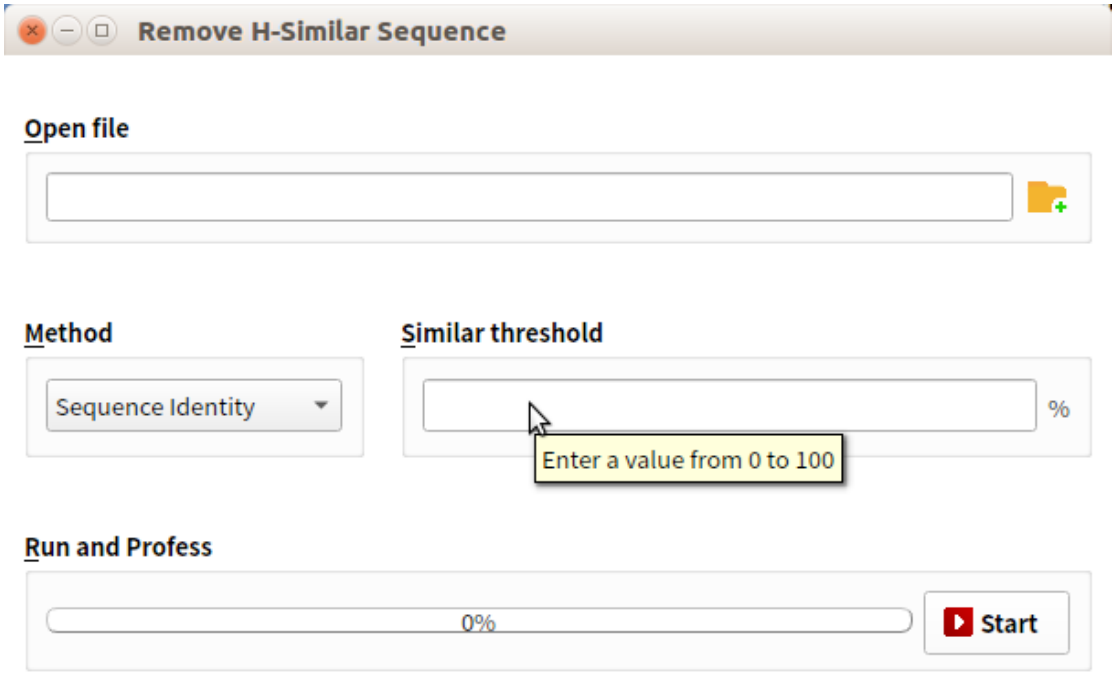


It should be noted that if the ***"Combination nt and aa"*** is selected, the inputed sequences should be aligned based on codon method in advance. In order to better fit

the variation characteristics, BioAider provides the ***"Condense gap"*** function. If the option was selected, the program will treat every three consecutive inserted or deleted bases as one.

3.2.2 *Remove High-Similar Sequence*

This function could remove highly similar sequences and keep one by specifing the threshold of similarity (***"Similar threshold"***). BioAider provides 6 different methods for calculating the similarity of sequences.

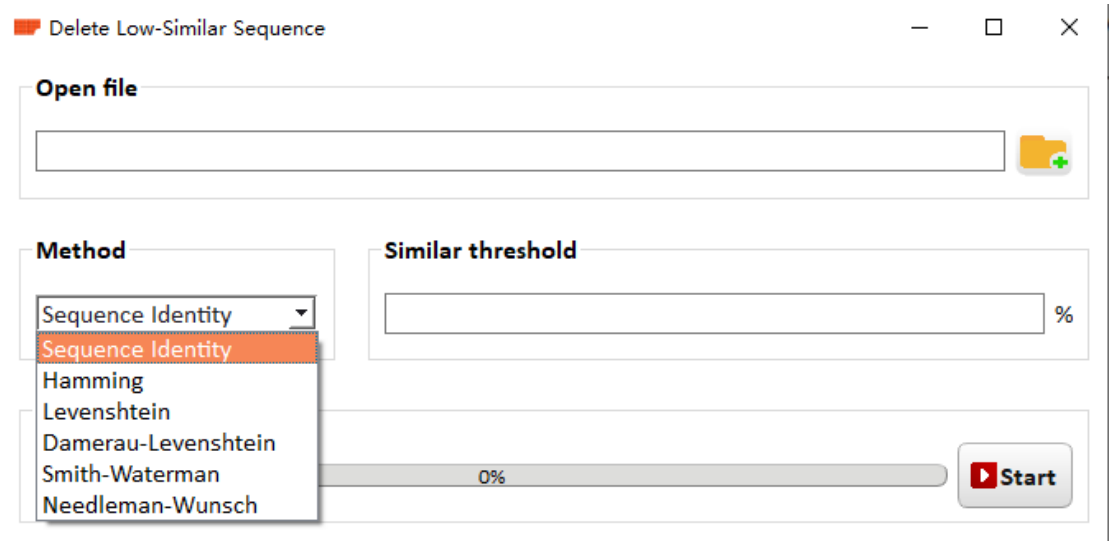


It should be noted that the ***"Sequence Identity"*** and ***"Hamming"*** methods require the input sequences data are aligned, and we suggest that the sequences datasets for remaining 4 methods better not be pre-aligned, because these algorithm own alignment function. If "***Similar threshold***" is set to 100, the function of deduplication will be turned on. **Note,** if the "***Similar threshold***" is set to 100, no matter what algorithm is selected, it is the same because the program adopts another efficient processing mechanism.

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.

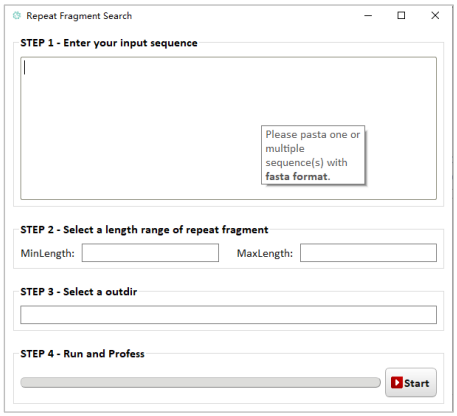
3.2.3 *Delete Low-Similar Sequence*

Specify a threshold (***"Similar threshold"***) to keep only one sequence with a similarity below a certain value.



3.2.4 *Repeat Fragment Search*

This function searches the sequence for repeating domains by specifying the length range of the repeating segment. You could enter multiple sequences (nucleotides or amino acids) for query at the same time, and the result examples are as follows:



	A	B	C
1	Qure sequence: EPI_ISL_420977		
2	**String**	**Position**	**Percentage**
3	TTTGTT	6,82,235	0.021818182
4	TTGTTT	7,236,256	0.021818182
5	TTTATG	10,323,419,788	0.029090909
6	TTATGA	11,420,789	0.021818182
7	GAATCT	17,580	0.014545455
8	TCTTCA	20,493,689	0.021818182
9	CTTCAC	21,618	0.014545455
10	TTCACA	22,737	0.014545455
11	CACAAT	24,741	0.014545455
12	CAATTG	26,497,652	0.021818182
13	TGGAAC	30,779	0.014545455
14	TGTAAC	36,261,486	0.021818182

3.3. *Mutation Tools*

3.3.1. *Mutation Analysis*

This function could be used for analysis of the **mutations characteristicson on large numbers of sequenced strains**. The sequence datas for analysis needs to be aligned in

advance, and they could be nucleotides, proteins（amino acid） sequences or simply coding gene fragments. For nucleotides and proteins sequences, BioAider could summarizes all the mutation sites with corresponding frequency and strains.

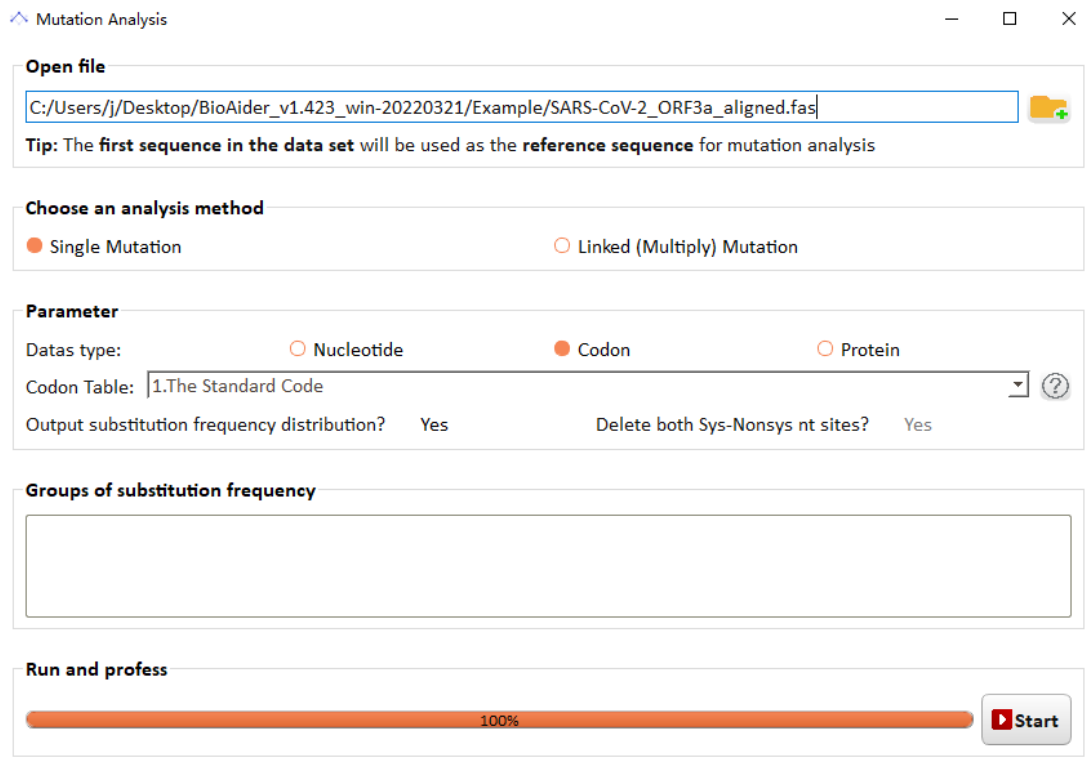
Of course, if the datas is codon gene, BioAider provides multiple sets of different codon tables for users, and could scan each condon sites in aligned sequence datasets, and identifies the type of mutation, including synonymous, non-synonymous, insertions and deletions and early termination. Finally, BioAider will automatically summarize and output the relevant analysis results.

Note: The codon gene sequences for mutations analysis have to be aligned by translation-alignment methon in advance, It is worth mentioning that BioAider packed three multiple-sequence-alignment software (mafft, muscle and clsutal-omega) in the graphical interface, and provided translation-alignment additionally.

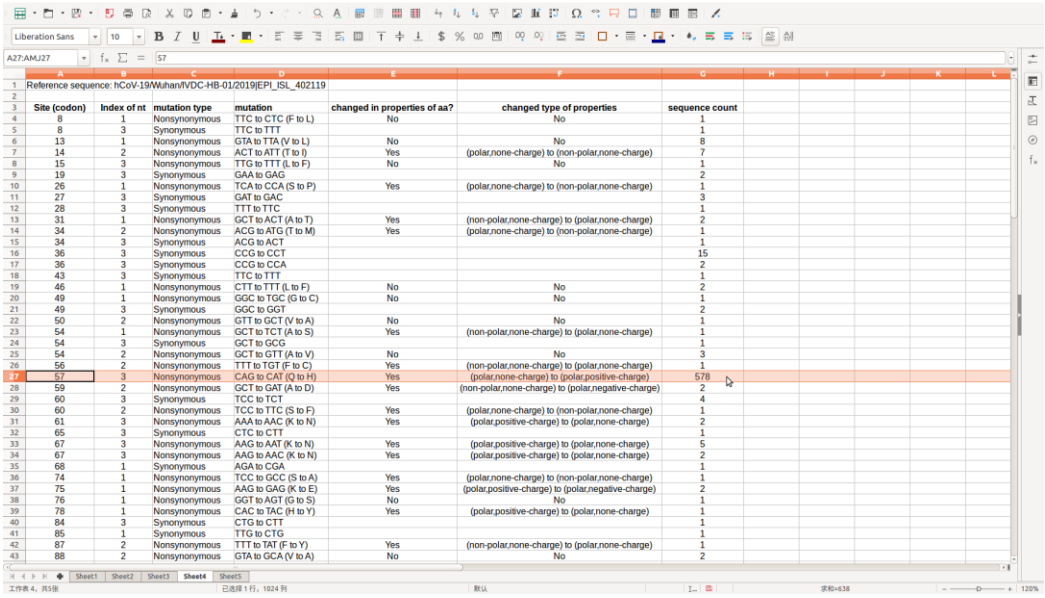
Whether it’s nucleotides or amino acids or coding genes, BioAider could plot the frequency distribution graph for mutation sites through specifing groups of substitution frequencey in custom.

Eaxmple of mutations analysis for aligned SARS-CoV-2 ORF3a gene sequences（an aligned coding gene sequence）.

First, Drag the sequence to be analyzed to the input box, and select "Codon" single button in ***"Datas type"***:



After the run is over, these analysis result could be found in the directory where the source file is located, you could scan the ***_mutation site summary*** file then know the overall variation and mutation hotspots.



Reference sequence: hCoV-19/Wuhan/IVDC-HB-01/2019|EPI_ISL_402119

Site (codon)	Index of nt	mutation type	mutation	changed in properties of aa?	changed type of properties	sequence count
8	1	Nonsynonymous	TTC to CTC (F to L)	No	No	1
8	3	Synonymous	TTC to TTT			1
13	1	Nonsynonymous	GTA to TTA (V to L)	No	No	8
14	2	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)	7
15	3	Nonsynonymous	TTG to TTT (L to F)	No	No	1
19	3	Synonymous	GAA to GAG			2
26	1	Nonsynonymous	TCA to CCA (S to P)	Yes	(polar,none-charge) to (non-polar,none-charge)	1
27	3	Synonymous	GAT to GAC			3
28	3	Synonymous	TTT to TTC			1
31	1	Nonsynonymous	GCT to ACT (A to T)	Yes	(non-polar,none-charge) to (polar,none-charge)	2
34	2	Nonsynonymous	ACG to ATG (T to M)	Yes	(polar,none-charge) to (non-polar,none-charge)	1
34	3	Synonymous	ACG to ACT			1
36	3	Synonymous	CCG to CCT			15
36	3	Synonymous	CCG to CCA			2
43	3	Synonymous	TTC to TTT			1
46	1	Nonsynonymous	CTT to TTT (L to F)	No	No	2
49	1	Nonsynonymous	GGC to TGC (G to C)	No	No	1
49	3	Synonymous	GGC to GGT			2
50	2	Nonsynonymous	GTT to GCT (V to A)	No	No	1
54	1	Nonsynonymous	GCT to TCT (A to S)	Yes	(non-polar,none-charge) to (polar,none-charge)	1
54	3	Synonymous	GCT to GCG			1
54	2	Nonsynonymous	GCT to GTT (A to V)	No	No	3
56	2	Nonsynonymous	TTT to TGT (F to C)	Yes	(non-polar,none-charge) to (polar,none-charge)	1
57	3	Nonsynonymous	CAG to CAT (Q to H)	Yes	(polar,none-charge) to (polar,positive-charge)	578
59	2	Nonsynonymous	GCT to GAT (A to D)	Yes	(non-polar,none-charge) to (polar,negative-charge)	2
60	3	Synonymous	TCC to TCT			4
60	2	Nonsynonymous	TCC to TTC (S to F)	Yes	(polar,none-charge) to (non-polar,none-charge)	1
61	3	Nonsynonymous	AAA to AAC (K to N)	Yes	(polar,positive-charge) to (polar,none-charge)	2
65	3	Synonymous	CTC to CTT			1
67	3	Nonsynonymous	AAG to AAT (K to N)	Yes	(polar,positive-charge) to (polar,none-charge)	5
67	3	Nonsynonymous	AAG to AAC (K to N)	Yes	(polar,positive-charge) to (polar,none-charge)	2
68	1	Synonymous	AGA to CGA			1
74	1	Nonsynonymous	TCC to GCC (S to A)	Yes	(polar,none-charge) to (non-polar,none-charge)	1
75	1	Nonsynonymous	AAG to GAG (K to E)	Yes	(polar,positive-charge) to (polar,negative-charge)	2
76	1	Nonsynonymous	GGT to AGT (G to S)	No	No	1
78	1	Nonsynonymous	CAC to TAC (H to Y)	Yes	(polar,positive-charge) to (polar,none-charge)	1
84	3	Synonymous	CTG to CTT			1
85	1	Synonymous	TTG to CTG			1
87	2	Nonsynonymous	TTT to TAT (F to Y)	Yes	(non-polar,none-charge) to (polar,none-charge)	1
88	2	Nonsynonymous	GTA to GCA (V to A)	No	No	2

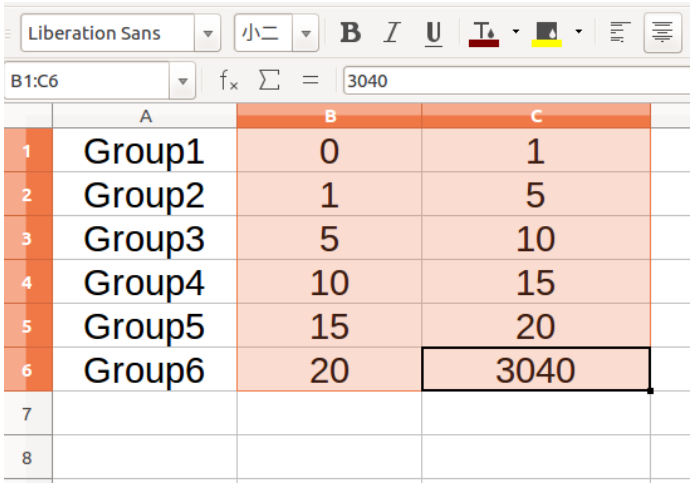
Codon-wise statistics on synonymous and non-synonymous substitutions are also provided in ***"Statistics in codons"*** directory：

	A	B	C	D	E	F	G
1		codon_site8	codon_site13	codon_site14	codon_site15	codon_site19	codon_site26
2	Synonymous	1	0	0	0	2	0
3	Nonsynonymou	1	8	7	1	0	1
4	Termination	0	0	0	0	0	0
5	Insertion	0	0	0	0	0	0
6	Deletion	0	0	0	0	0	0

Besides, BioAider uniquely provides statistical **synonymous and non-synonymous substitution** nucleotide positions **in "base" units :**

	A	B	C	D	E
1	**Codon**	**Base index**	**Nucleotide site**	**Type**	**Substitution frequency**
2	8	3	24	Synonymous	1
3	19	3	57	Synonymous	2
4	27	3	81	Synonymous	3
5	28	3	84	Synonymous	1
6	34	3	102	Synonymous	1
7	36	3	108	Synonymous	17
8	43	3	129	Synonymous	1
9	49	3	147	Synonymous	2
10	54	3	162	Synonymous	1
11	60	3	180	Synonymous	4

If you also need to plot the distribution of synonymous/non-synonymous substitution bases, you can prepare a grouping table first:

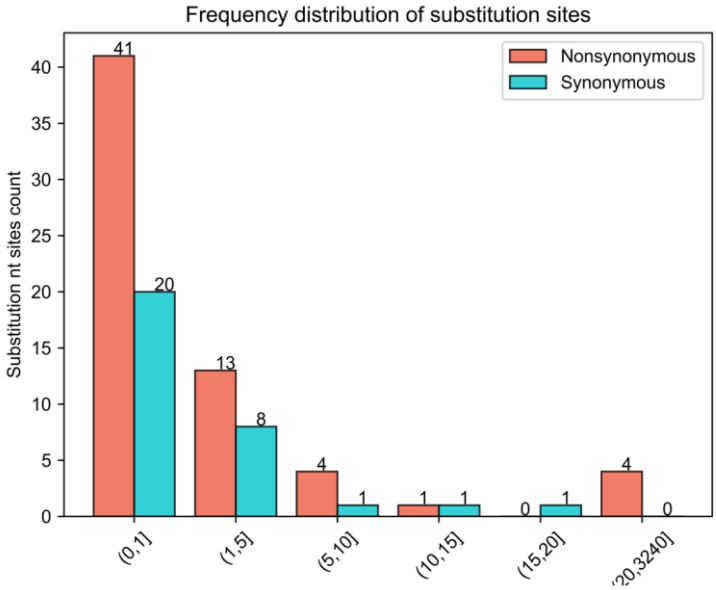


Each group of substitution frequency contains start value and end value which are separated by tab symbol. **Note, *the start value*** of each group is not included in the range

of frequency, **and the frequencies of different groups need to be consecutive integers.**

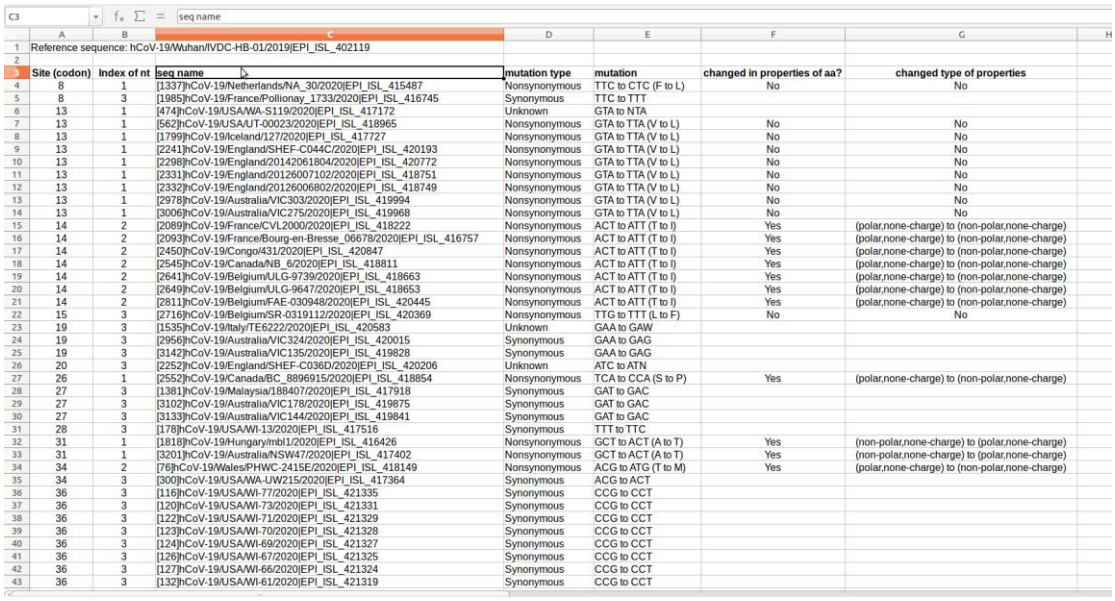
Then copy them to the textedit box of BioAider,

You could also konw the number of mutation nucleotide site under each mutation frequency group through view ****_substitution frequency distribution.png***.



It is not difficult to find that more than half of the mutation sites only appear in a single strain, although there are many mutation sites in ORF3a gene. Of course, BioAider additionally provides vector graphics ***(*_substitution frequency distribution.pdf)***, users can edit them and facilitate publication.

Besides, users could obtain the corresponding mutant strains of these variant sites in the detailed ****_log.txt*** file.



	A	B	C	D	E	F	G
1	Reference sequence: hCoV-19/Wuhan/IVDC-HB-01/2019|EPI_ISL_402119						
2							
3	Site (codon)	Index of nt	seq name	mutation type	mutation	changed in properties of aa?	changed type of properties
4	8	1	[1337]hCoV-19/Netherlands/NA_30/2020|EPI_ISL_415487	Nonsynonymous	TTC to CTC (F to L)	No	No
5	8	3	[1985]hCoV-19/France/Pollionay_1733/2020|EPI_ISL_416745	Synonymous	TTC to TTT		
6	13	1	[474]hCoV-19/USA/WA-S119/2020|EPI_ISL_417172	Unknown	GTA to NTA		
7	13	1	[562]hCoV-19/USA/UT-00023/2020|EPI_ISL_418965	Nonsynonymous	GTA to TTA (V to L)	No	No
8	13	1	[1799]hCoV-19/Iceland/127/2020|EPI_ISL_417727	Nonsynonymous	GTA to TTA (V to L)	No	No
9	13	1	[2241]hCoV-19/England/SHEF-C044C/2020|EPI_ISL_420193	Nonsynonymous	GTA to TTA (V to L)	No	No
10	13	1	[2298]hCoV-19/England/20142061804/2020|EPI_ISL_420772	Nonsynonymous	GTA to TTA (V to L)	No	No
11	13	1	[2331]hCoV-19/England/20126007102/2020|EPI_ISL_418751	Nonsynonymous	GTA to TTA (V to L)	No	No
12	13	1	[2332]hCoV-19/England/20126006802/2020|EPI_ISL_418749	Nonsynonymous	GTA to TTA (V to L)	No	No
13	13	1	[2978]hCoV-19/Australia/VIC303/2020|EPI_ISL_419994	Nonsynonymous	GTA to TTA (V to L)	No	No
14	13	1	[3006]hCoV-19/Australia/VIC275/2020|EPI_ISL_419968	Nonsynonymous	GTA to TTA (V to L)	No	No
15	14	2	[2089]hCoV-19/France/CVL2000/2020|EPI_ISL_418222	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)
16	14	2	[2093]hCoV-19/France/Bourg-en-Bresse_06678/2020|EPI_ISL_416757	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)
17	14	2	[2450]hCoV-19/Congo/431/2020|EPI_ISL_420847	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)
18	14	2	[2545]hCoV-19/Canada/NB_6/2020|EPI_ISL_418811	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)
19	14	2	[2641]hCoV-19/Belgium/ULG-9739/2020|EPI_ISL_418663	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)
20	14	2	[2649]hCoV-19/Belgium/ULG-9647/2020|EPI_ISL_418653	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)
21	14	2	[2811]hCoV-19/Belgium/FAE-030948/2020|EPI_ISL_420445	Nonsynonymous	ACT to ATT (T to I)	Yes	(polar,none-charge) to (non-polar,none-charge)
22	15	3	[2716]hCoV-19/Belgium/SR-0319112/2020|EPI_ISL_420369	Nonsynonymous	TTG to TTT (L to F)	No	No
23	19	3	[1535]hCoV-19/Italy/TE6222/2020|EPI_ISL_420583	Unknown	GAA to GAW		
24	19	3	[2956]hCoV-19/Australia/VIC324/2020|EPI_ISL_420015	Synonymous	GAA to GAG		
25	19	3	[3142]hCoV-19/Australia/VIC135/2020|EPI_ISL_419828	Synonymous	GAA to GAG		
26	20	3	[2252]hCoV-19/England/SHEF-C036D/2020|EPI_ISL_420206	Unknown	ATC to ATN		
27	26	1	[2552]hCoV-19/Canada/BC_8896915/2020|EPI_ISL_418854	Nonsynonymous	TCA to CCA (S to P)	Yes	(polar,none-charge) to (non-polar,none-charge)
28	27	3	[1381]hCoV-19/Malaysia/188407/2020|EPI_ISL_417918	Synonymous	GAT to GAC		
29	27	3	[3102]hCoV-19/Australia/VIC178/2020|EPI_ISL_419875	Synonymous	GAT to GAC		
30	27	3	[3133]hCoV-19/Australia/VIC144/2020|EPI_ISL_419841	Synonymous	GAT to GAC		
31	28	3	[178]hCoV-19/USA/WI-13/2020|EPI_ISL_417516	Synonymous	TTT to TTC		
32	31	1	[1818]hCoV-19/Hungary/mb1/2020|EPI_ISL_416426	Nonsynonymous	GCT to ACT (A to T)	Yes	(non-polar,none-charge) to (polar,none-charge)
33	31	1	[3201]hCoV-19/Australia/NSW47/2020|EPI_ISL_417402	Nonsynonymous	GCT to ACT (A to T)	Yes	(non-polar,none-charge) to (polar,none-charge)
34	34	2	[76]hCoV-19/Wales/PHWC-2415E/2020|EPI_ISL_418149	Nonsynonymous	ACG to ATG (T to M)	Yes	(polar,none-charge) to (non-polar,none-charge)
35	34	3	[300]hCoV-19/USA/WA-UW215/2020|EPI_ISL_417364	Synonymous	ACG to ACT		
36	36	3	[116]hCoV-19/USA/WI-77/2020|EPI_ISL_421335	Synonymous	CCG to CCT		
37	36	3	[120]hCoV-19/USA/WI-73/2020|EPI_ISL_421331	Synonymous	CCG to CCT		
38	36	3	[122]hCoV-19/USA/WI-71/2020|EPI_ISL_421329	Synonymous	CCG to CCT		
39	36	3	[123]hCoV-19/USA/WI-70/2020|EPI_ISL_421328	Synonymous	CCG to CCT		
40	36	3	[124]hCoV-19/USA/WI-69/2020|EPI_ISL_421327	Synonymous	CCG to CCT		
41	36	3	[126]hCoV-19/USA/WI-67/2020|EPI_ISL_421325	Synonymous	CCG to CCT		
42	36	3	[127]hCoV-19/USA/WI-66/2020|EPI_ISL_421324	Synonymous	CCG to CCT		
43	36	3	[132]hCoV-19/USA/WI-61/2020|EPI_ISL_421319	Synonymous	CCG to CCT		

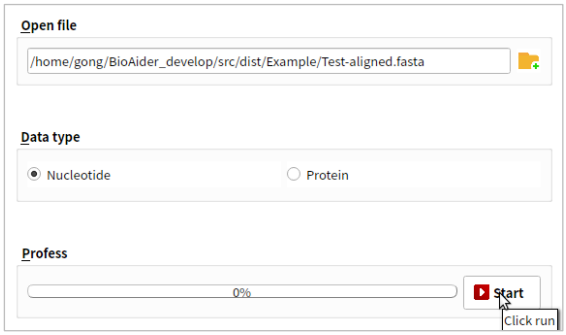
Of note, if these sequences are much divergent, such as from different family enver order and contain a lot of gaps ("-") in the aligned sequence, I usually don't recommend using them for mutation analysis. On the one hand, they would make a lot of calculations, on the other hand, they are inherently highly variable and have no value of analysis.

But if you still want to study their variation, it is recommended to use the following function of ***"Site Counter"***.

3.3.2. *Site Counter*

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.

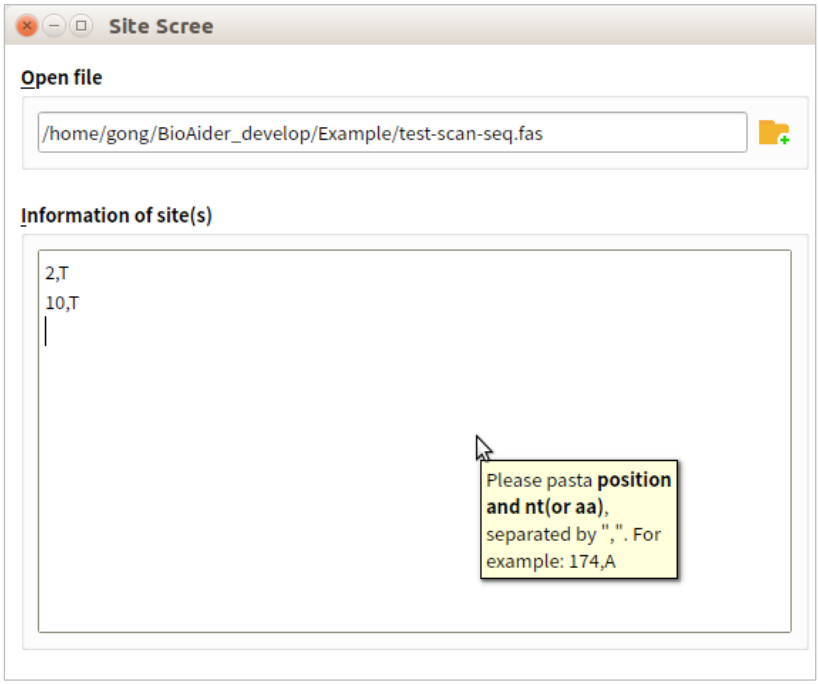
For DNA sequence datasets, the one of results (****_site_count.csv***) was as follows:



A	B	C	D	E	F
	site1	site2	site3	site4	site5
A	29	0	0	0	0
G	0	0	29	0	0
C	0	0	0	0	0
T	0	29	0	29	29
-	0	0	0	0	0

3.3.3. *Site Scree*

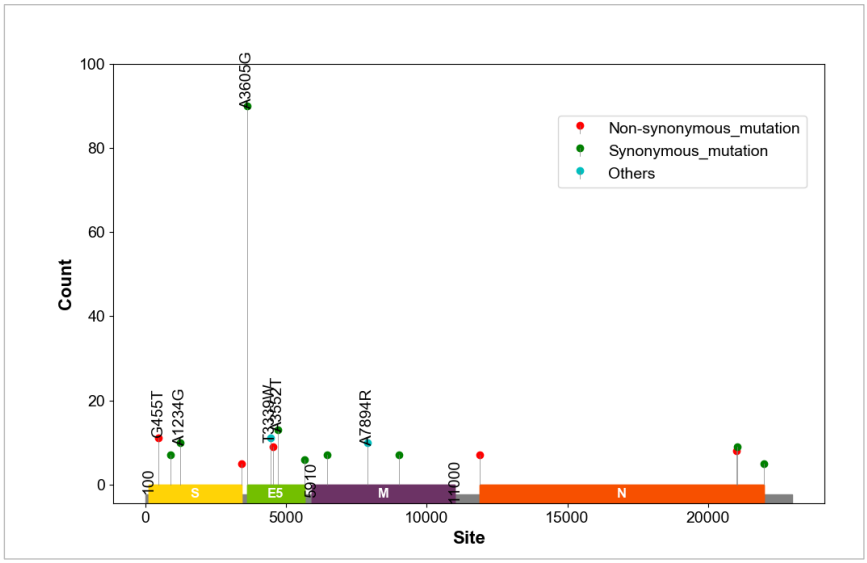
This function is used to extract the sequences with corresponding base (or amino acid) in ***specified one or more*** site(s). It is very useful for studying whether there is linkage inheritance among different gene sites.



3.4. Drawing module

3.4.1. Lollipop chart of gene mutation

Lollipop map is an efficient method to display gene mutation sites and frequencies, they look like the following:

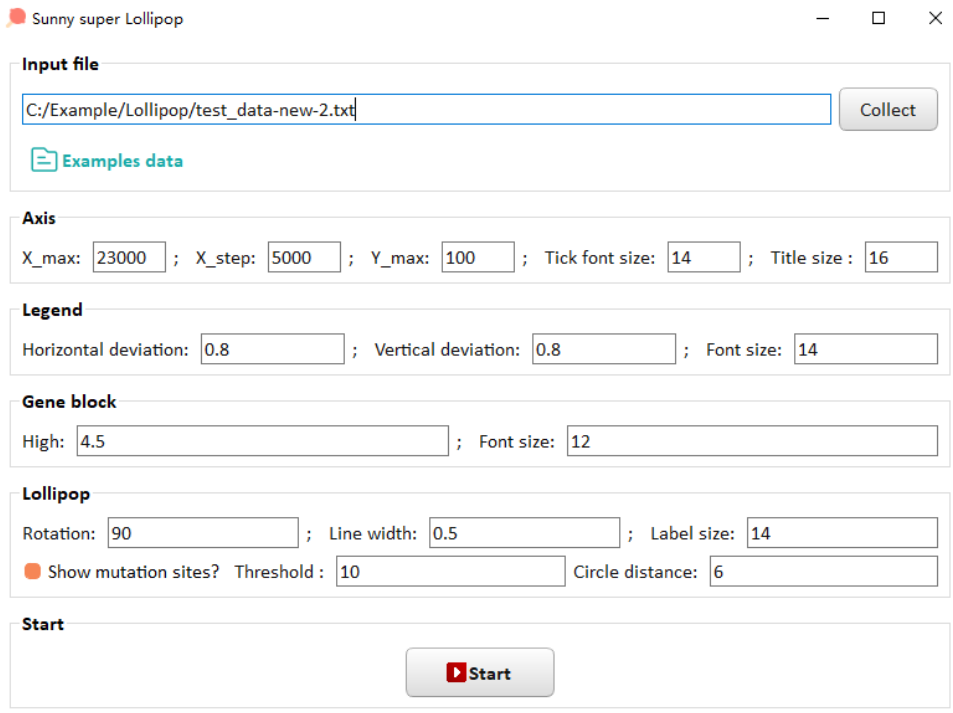


In BioAider, you only need to prepare the corresponding matrix file and simply set the parameters to quickly complete the drawing:

	Site	Gene	Type	Count	Mutation	Color
2	100	S	Start	11	100	#FFD306
3	455	S	Non-synonymous_mutation	11	G455T	#FFD306
4	900	S	Synonymous_mutation	7	C900A	#FFD306
5	1234	S	Synonymous_mutation	10	A1234G	#FFD306
6	3400	S	Non-synonymous_mutation	5	G3400A	#FFD306
7	3605	E5	Synonymous_mutation	90	A3605G	#73BF00
8	4439	E5	Others	11	T3339W	#73BF00
9	4540	E5	Non-synonymous_mutation	9	G3440C	#73BF00
10	4700	E5	Synonymous_mutation	13	A3552T	#73BF00
11	5653	E5	Synonymous_mutation	6	C5653G	#73BF00
12	5910	M	Start	7	5910	#6C3365
13	6444	M	Synonymous_mutation	7	G6444C	#6C3365
14	7894	M	Others	10	A7894R	#6C3365
15	9004	M	Synonymous_mutation	7	G9004C	#6C3365
16	11000	M	End	2	11000	#6C3365
17	11894	N	Non-synonymous_mutation	7	G11894A	#F75000
18	21004	N	Non-synonymous_mutation	8	G21004T	#F75000
19	21984	N	Synonymous_mutation	5	A21984G	#F75000
20	21029	N	Synonymous_mutation	9	C21029T	#F75000

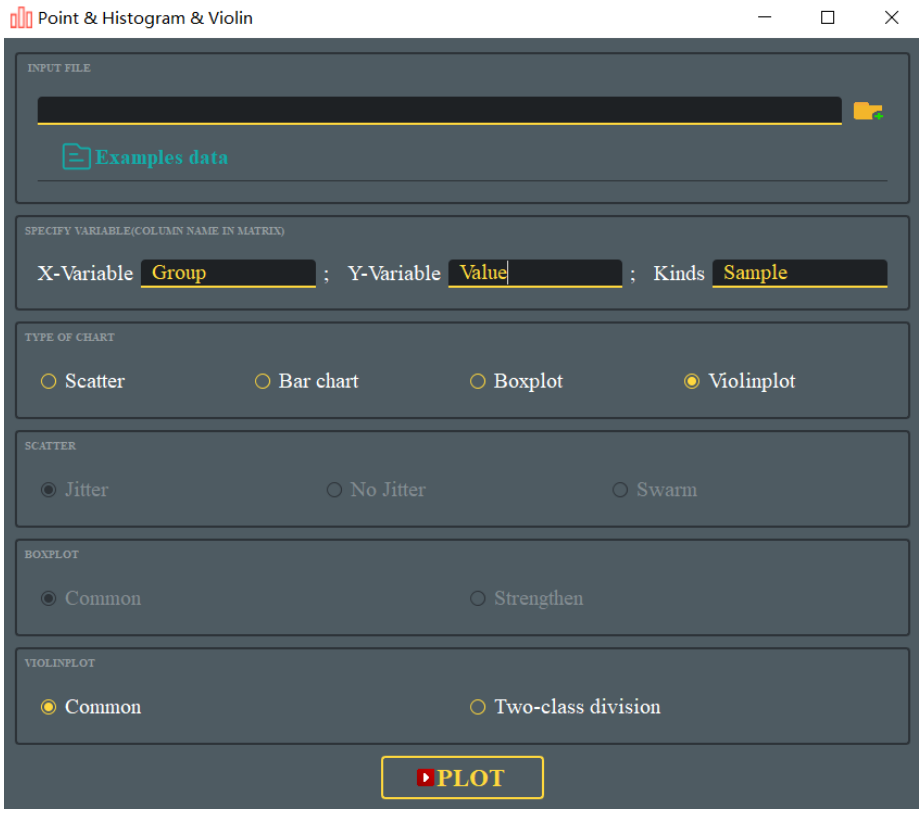
Tip： Note that the data **has only 6 columns**, **and the column names cannot be changed**, and other information can be flexibly configured. **Besides**, lollipops are not drawn at sites marked "Start" or "End" in the "Type" column, but are used to assist in gene scoping.

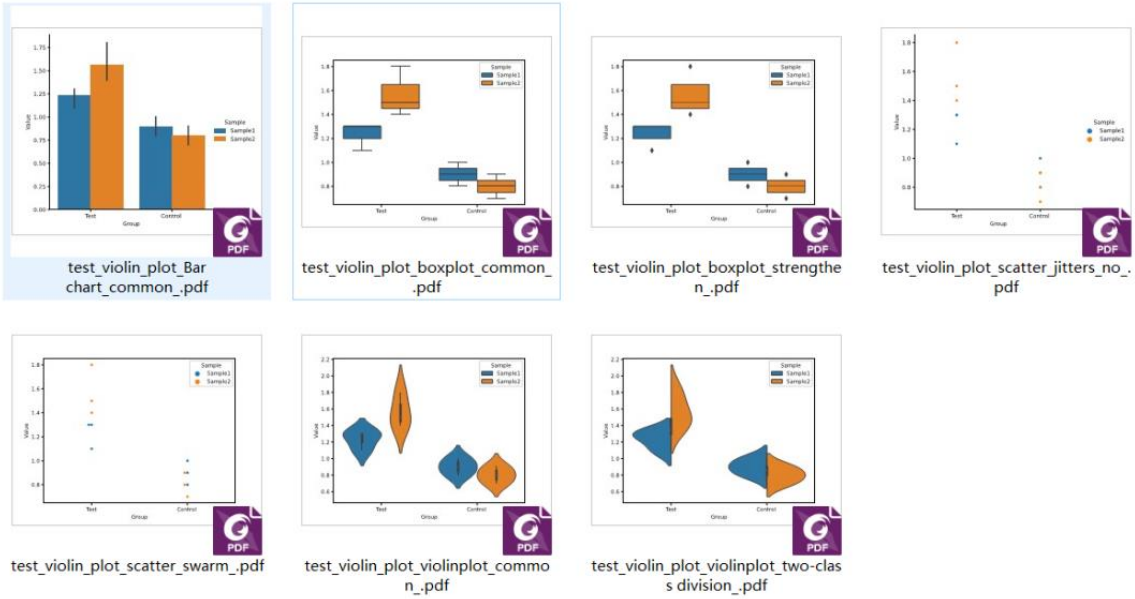
Then submit to BioAider for drawing:



3.4.2. Commonly used statistics

BioAider provides a GUI interface for quickly drawing scatter, box and violin plots based on the seaborn package:





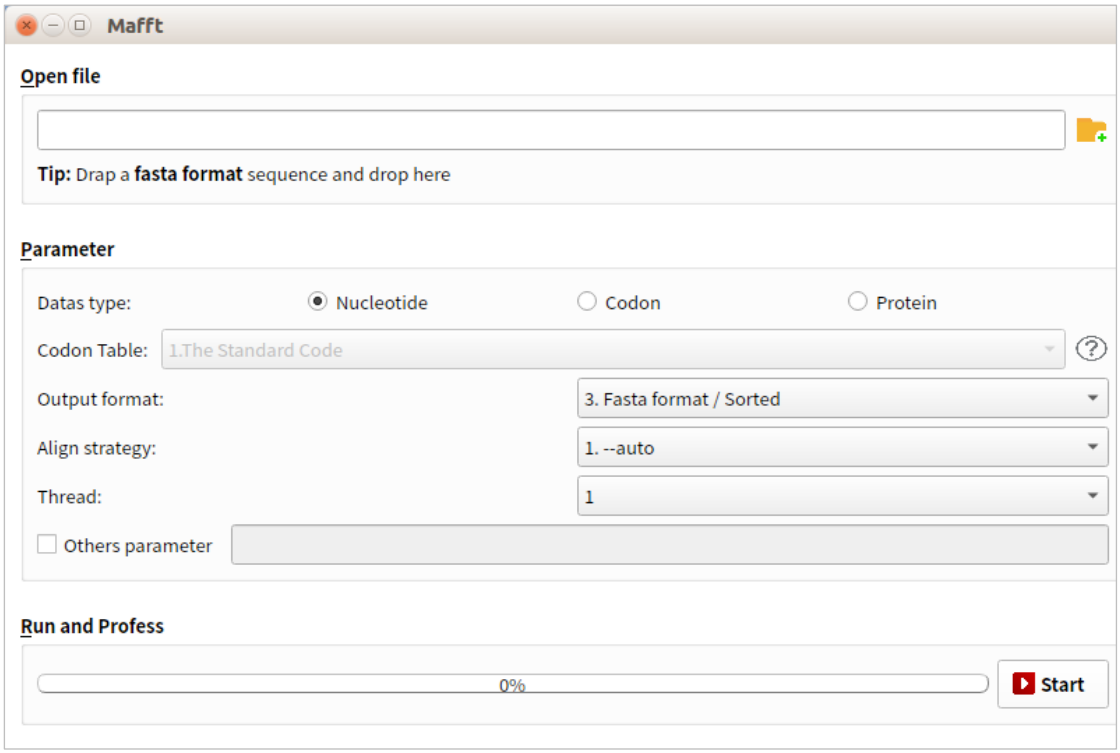
4. Plugins supported

4.1. Alignment tools

Multiple-Sequence-Alignment (MSA) is the most common analysis in sequence processing, most classic MSA software runs as a command symbol. Sometimes it is difficult to use for some biological or clinical expert without bioinformatics skill. In order to facilitate parameter setting and usage, BioAider provides graphical interfaces with main features for three classic MSA software (Mafft, Muscle and Clsutal-Omega) through plugins form, **and provided translation-alignment additionally based on multiple sets of codon tables.**

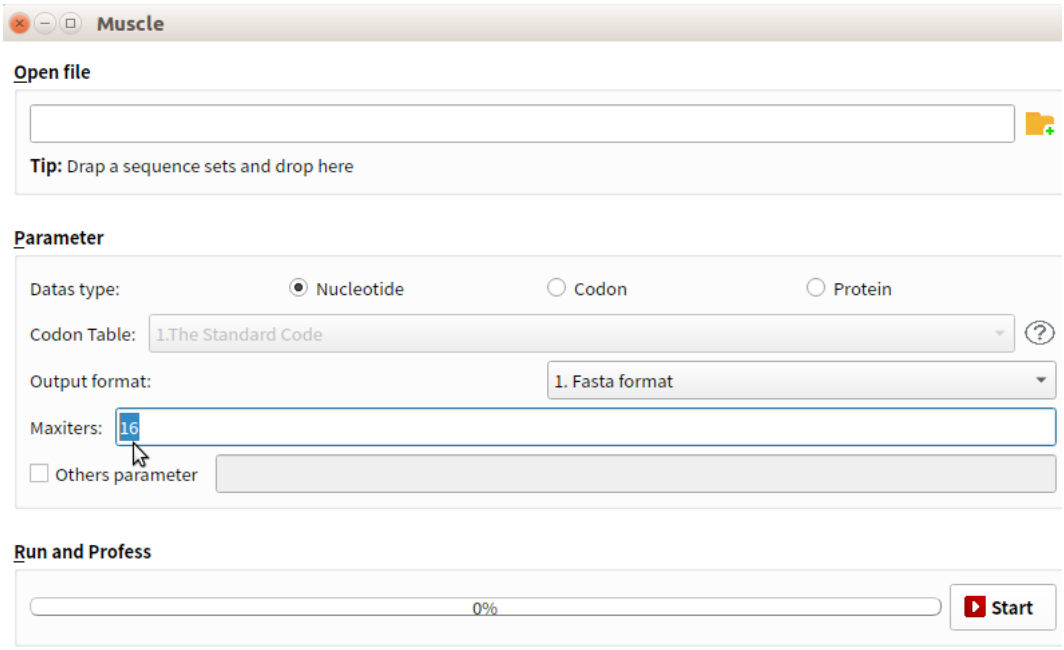
4.1.1. *Mafft*

Mafft is a very popular MSA software with higher comparison accuracy, and its comparison speed is also relatively good. More detailed information about Mafft please see https://mafft.cbrc.jp/alignment/software/. The reference of Mafft7 is at https://academic.oup.com/mbe/article/30/4/772/1073398.



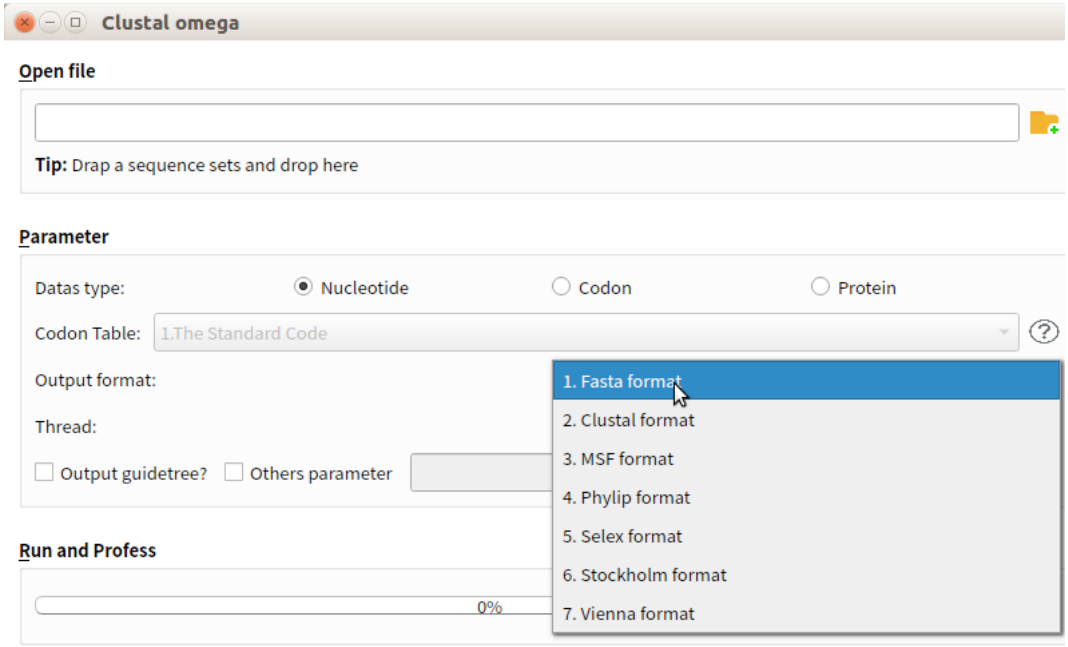
4.1.2. *Muscle*

The comparison rate and accuracy of Muscle are good, according to the instruction manual of Muscle, setting ***"Maxiters"*** to 1 or 2 will significantly speed up the operation. More detailed information about Muscle please see http://petrov.stanford.edu/software/src/muscle3.6/muscle3.6.html. The reference of Muscle is at https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-113.



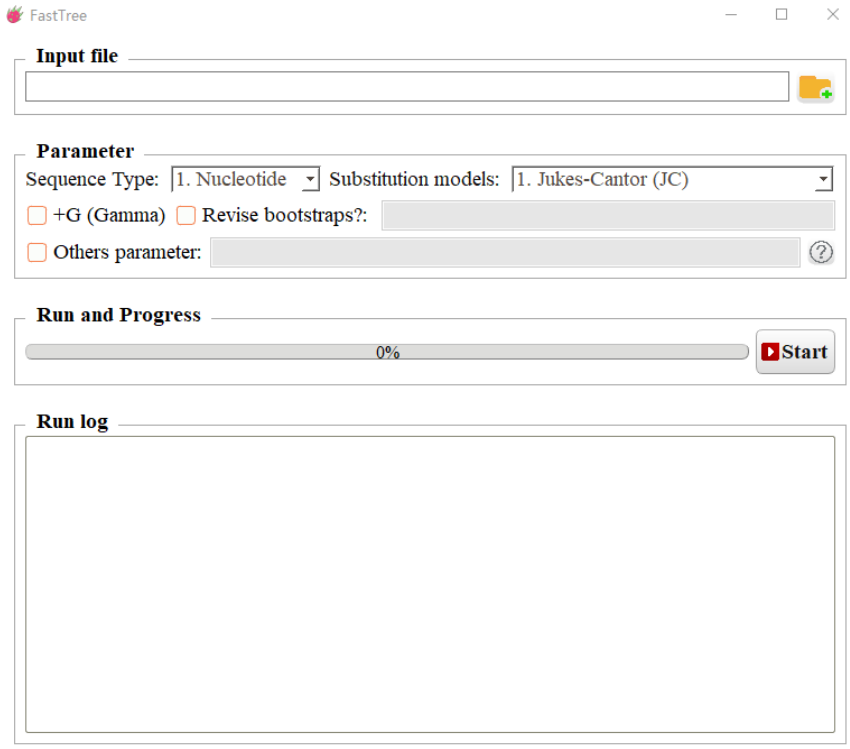
4.1.3. *Clustal-Omega*

As a relatively classic MSA software, *Clustal* owns a broad user base. As the latest addition to the Clustal family, Clustal-Omega offers a significant increase in scalability over previous versions. More detailed information about Clustal-Omega, please see http://www.clustal.org/omega/. The reference of Clustal-Omega is at https://www.embopress.org/doi/full/10.1038/msb.2011.75.



4.2. Phylogenetic-tree tool

FastTree is a super-fast tool for building maximum likelihood tree. BioAider provides graphical interfaces with main features for FastTree software, to facilitate parameter setting and usage. The reference of FastTree is at https://www.embopress.org/doi/full/10.1038/msb.2011.75.



5. Test Datas

Examples and test are available at:
https://github.com/ZhijianZhou01/BioAider/tree/master/Example.

6. Citation

If you wish to cite BioAider in a publication, we suggest the following:

Zhou ZJ, Qiu Y, Pu Y, Huang X, Ge XY*. BioAider: An efficient tool for viral genome analysis and its application in tracing SARS-CoV-2 transmission. Sustain Cities Soc. 2020;63:102466. doi:10.1016/j.scs.2020.102466.

Publication of BioAider is available at Journal of : Sustainable Cities and Society