
Manual of BioAider V1.423

**A richly featured desktop platform libraries for
analysis of bioinformatics data**

Written by Zhou ZJ

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

Version 1.423 || May 25, 2022

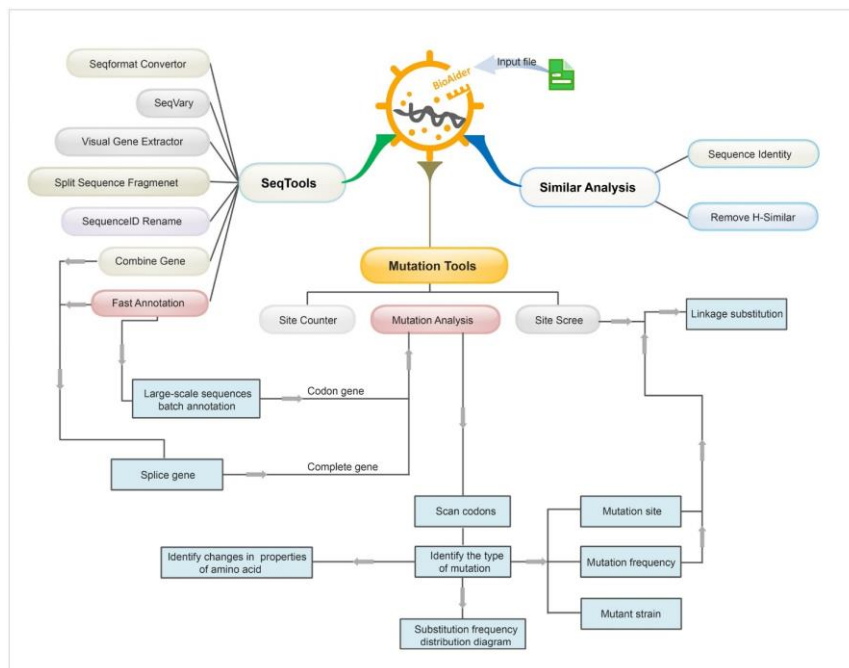
Content

1. Introduction.....	3
2. Download and install	3
3. Functions.....	4
3.1. SeqTools	4
3.1.1. Seqformat Convertor	4
3.1.2. SeqVary	4
3.1.3. SequenceID Rename	5
3.1.4. Split Sequence Fragment	6
3.1.5. Combine Gene (Tandem Gene)	6
3.1.6. Visual Gene Extractor	7
3.1.7. Fast Annotation	10
3.1.8. Vrial *.gb file parser	10
3.1.9. Correction ambiguous bases	11
3.2. Similar Analysis	12
3.2.1. Sequence Identity Matrix	12
3.2.2 Remove High-Similar Sequence	13
3.2.3 Delete Low-Similar Sequence	14
3.2.4 Repeat Fragment Search	14
3.3. Align tools.....	14
3.3.1. Mafft	15
3.3.2. Muscle	15
3.3.3. Clustal-Omeg.....	16
3.4. Mutation Tools	17
3.4.1. Mutation Analysis	17
3.4.2. Site Counter.....	21
3.4.3. Site Scree	22
3.5. Drawing module	22
3.5.1. Lollipop chart of gene mutation	22
3.5.1. Commonly used statistics	24
4. Test Datas	25
5. Ciation.....	25

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.

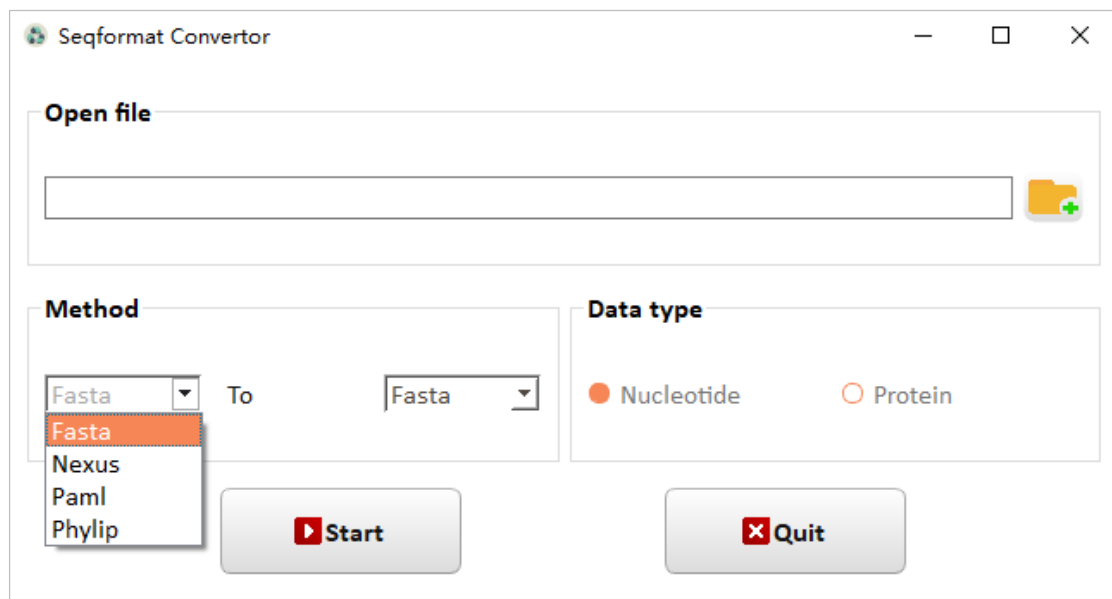
BioAider will be in the long-term update, this document briefly introduces some of its current commonly functions. In V1.423, we've beautified the interface again and added a variety of interface themes to make BioAider more interesting.

3. Functions

3.1. SeqTools

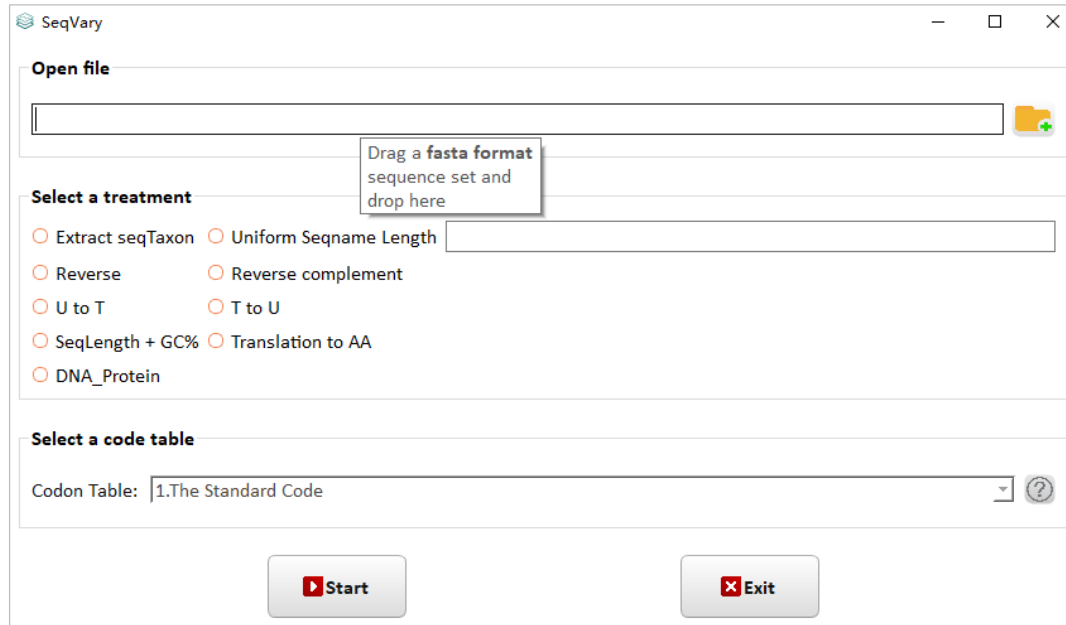
3.1.1. Seqformat Convertor

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.

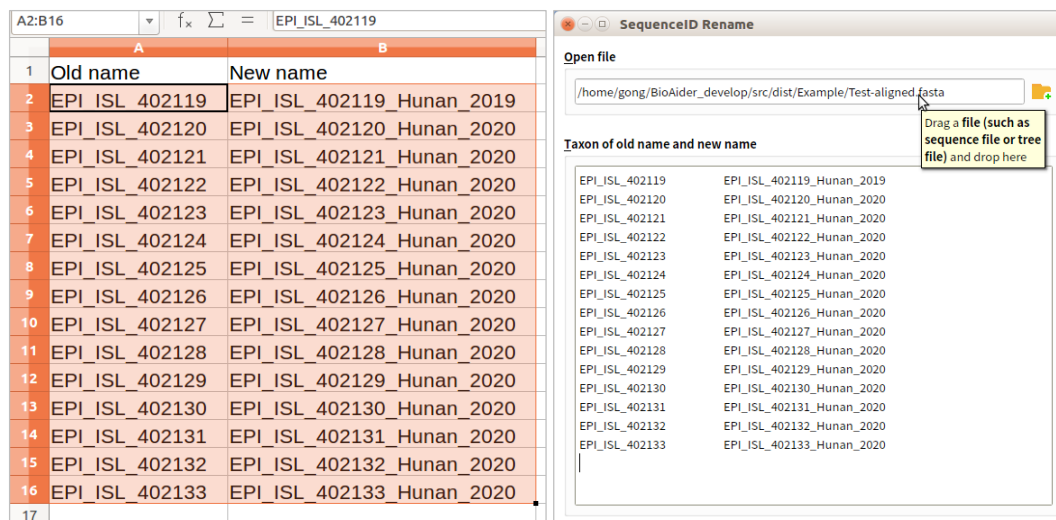


Note: the **"DNA Protein"** option requires the gene sequences data to be aligned based on codons.

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 Fragment

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

Mode of Split

☒ Different range ☐ Equal range

Taxon with seqID and regional

EPI_ISL_402119	1	600
EPI_ISL_402120	3	700
EPI_ISL_402121	46	788
EPI_ISL_402122	7	888
EPI_ISL_402123	9	333
EPI_ISL_402124	5	888
EPI_ISL_402125	33	777
EPI_ISL_402126	23	679
EPI_ISL_402127	33	767
EPI_ISL_402128	33	767
EPI_ISL_402129	55	890
EPI_ISL_402130	33	900
EPI_ISL_402131	44	678
EPI_ISL_402132	38	876
EPI_ISL_402133	56	890

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

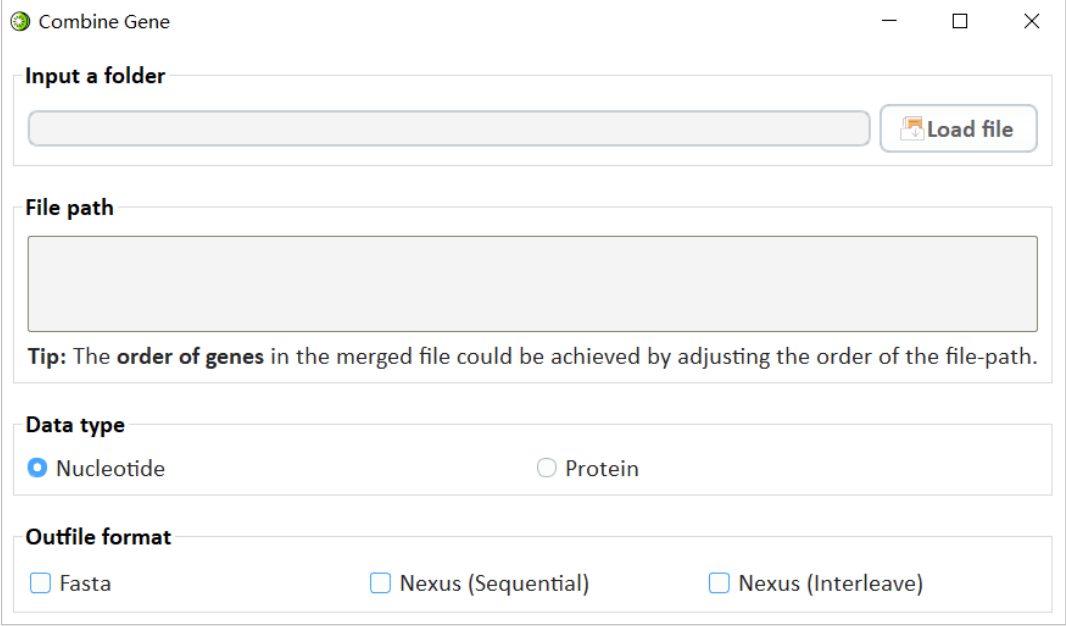
Run and Process

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 ("-").



Combine Gene

Input a folder

Load file

File path

Tip: The **order of genes** in the merged file could be achieved by adjusting the order of the file-path.

Data type

☒ Nucleotide ☐ Protein

Outfile format

☒ Fasta ☐ Nexus (Sequential) ☐ Nexus (Interleave)

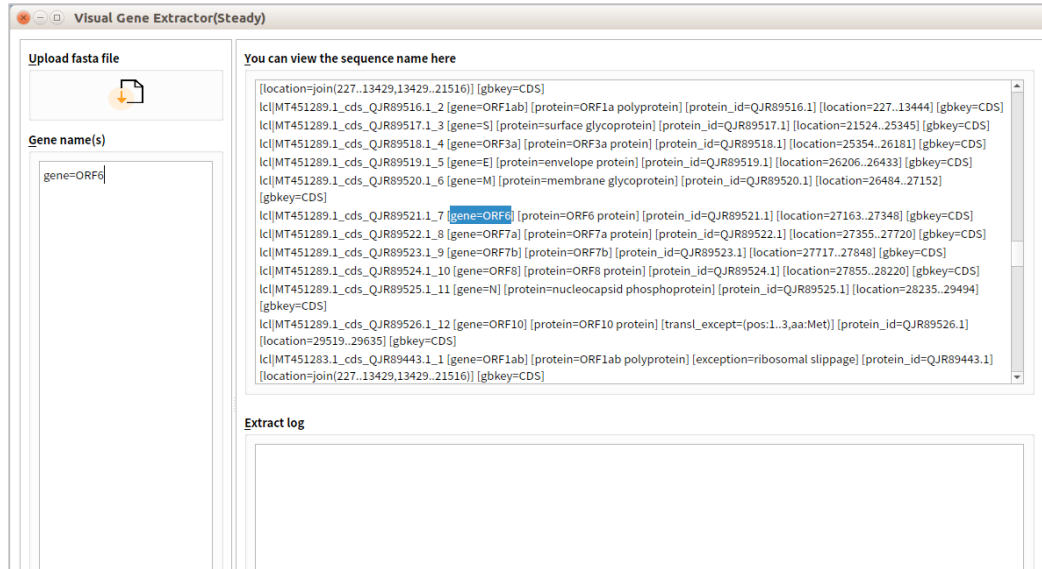
Note, users can **modify the order of genes** in tandemly sequence by adjusting the sort of inputfile path in the textbox. **Of note**, all the sequences which are used for combined should be fasta format.

3.1.6. Visual Gene Extractor

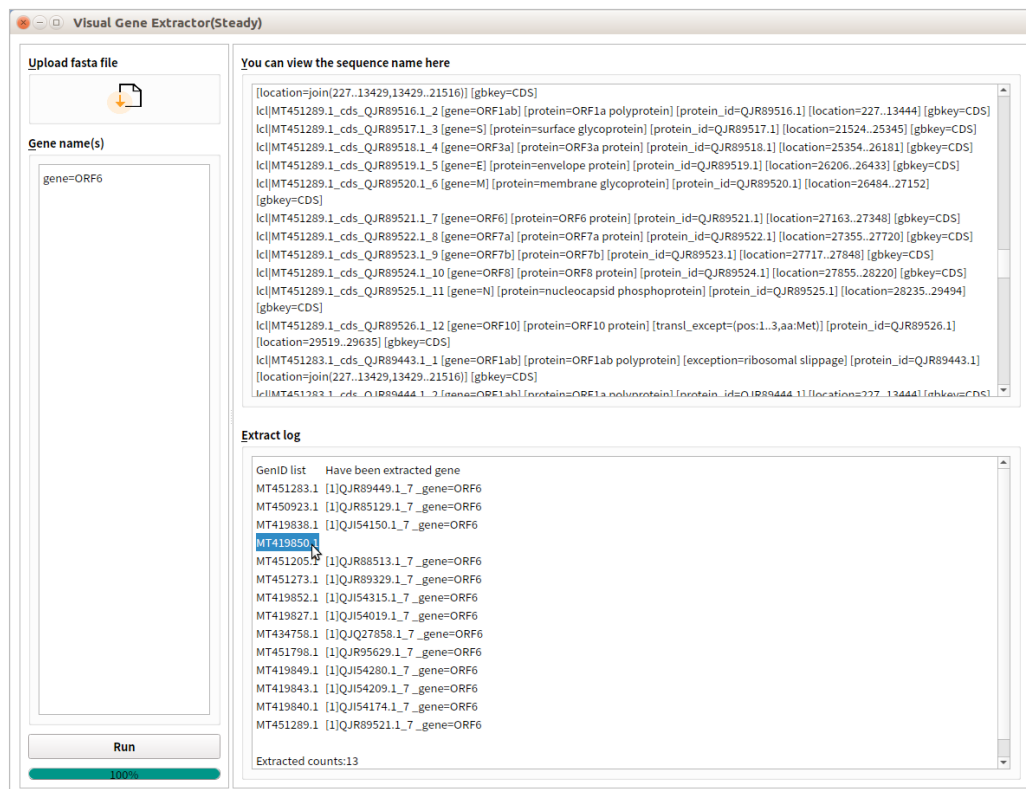
This function is used to extract the sequences included specified gene from mixed coding gene sequence set, especially when these sequences data are downloaded from NCBI database. Given that the same gene may have different manifestations in different studies, the textbox of "Gene name" could enter multiple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



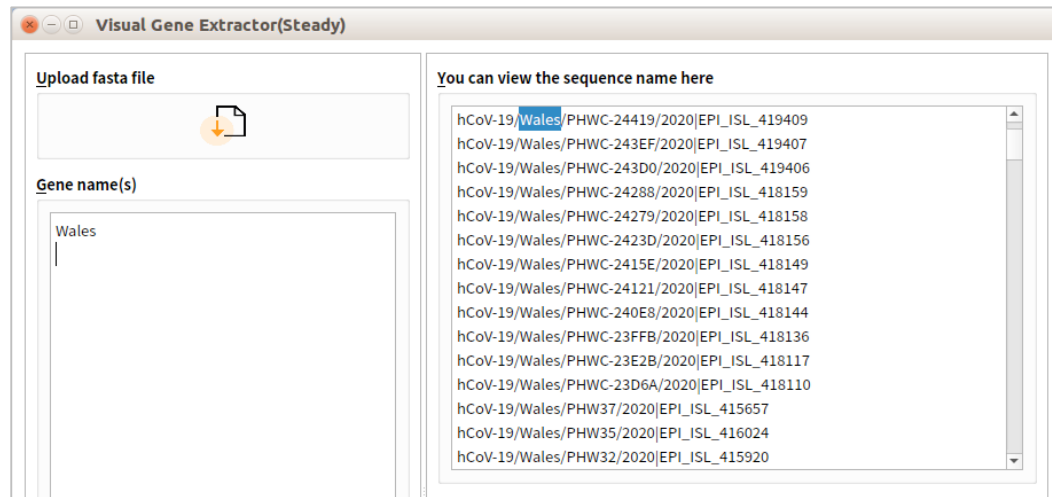
After uploading the sequence to BioAiders as above, then we extract ORF6 gene sequence of SARS-CoV-2. Input a string **containing at least the gene name** to textbox of **"Gene name(s)"**, then click button of **"Run"**, run log as follows:



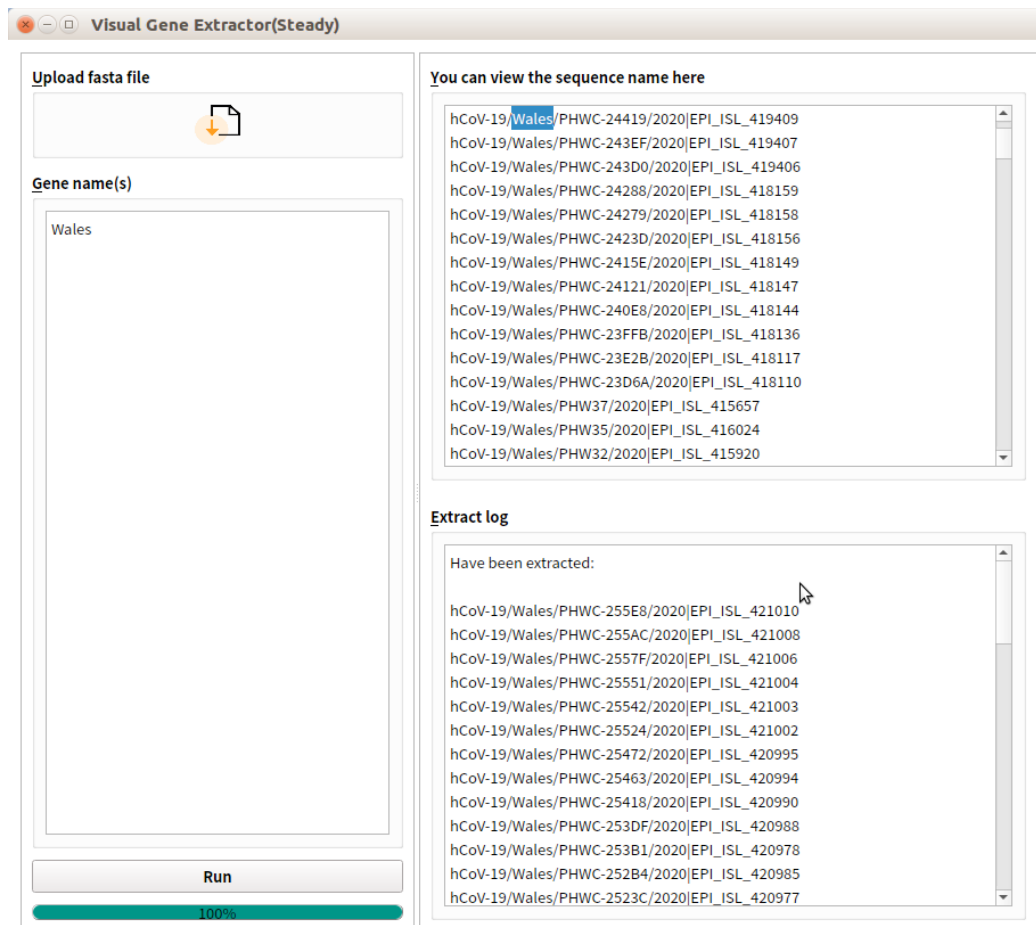
If some gene fragments are not extracted (as shown by the arrow), the possible reason is that the strain does not contain this gene fragment or the gene owns other names in

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.

Information of gene

```
#Example
ORF1ab,AAAAGGGGCCCTT,AAAAACCTTTCCC
S,TTTAAAGGGCCTTC,GGGGTATTTATATCTTAA
ORF3a,AAGCCCTCTGCTCGTCG,TTGTGAGTGTGGA
|
```

Please paste a table with the names, start and end character string of gene of first seq in data set, separated by ",".

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:

Viral *.gb file parser

Input *.gb file

G:/alpha-CoVs.gb

Preview

Preview of strain information

	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	DQ701147	DQ701147.1	28548	RNA	linear	"06-SEP-2016"	Viruses	Riboviria	Orthornavirae	Pisuviricota	Pisoniviricetes	Nidovirales

Taxonomy

Classification extraction? Biological

Run and Preprocess

0%

Start

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

Preview of strain information

Strain	Host	Collection region	Collection date	Title
229E	Unknown	Unknown	"Unknown"	Infectious RNA transcribed in vitro from a cDNA copy of the human coronavirus
CV777	Unknown	Unknown	"Unknown"	Sequence determination of the nucleocapsid protein gene of the porcine coronavirus
Purdue	pig	USA:Indiana	"Unknown"	Engineering the largest RNA virus genome as an infectious bacterial artificial
Unknown	Unknown	Unknown	"Unknown"	Artificial chromosome constructs containing nucleic acid sequences capable
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 capable
Unknown	Unknown	Unknown	"Unknown"	Coronavirus, nucleic acid, protein, and methods for the generation of vaccine
Unknown	Unknown	Unknown	"Unknown"	Coronavirus, nucleic acid, protein, and methods for the generation of vaccine
FIPV 79-1146	cat	Unknown	"Unknown"	Genomic RNA sequence of Feline coronavirus strain FIPV WSU-79/1146
TS	bat	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:

Ambiguous Base Edition

Open file

Run and Profess

0% **Start**

Run log

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").

Sequence Identity Matrix

Open file

/home/gong/BioAider_develop/src/dist/Example/Test-aligned.fasta

Type of matrix

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

Condense gap?

☐ Yes

Select a code table

Codon Table: 1.The Stand... ?

Run and Profess

0% **Start**

Combination nt and aa (nt/aa) asks the datas should be nucleotide sequence of coding gene and alignment based on codons method

It should be noted that if the "Combination nt and aa" is selected, the inputted 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 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 **"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	Query 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. Align tools

Multiple-Sequence-Alignment (MSA) is the most common analysis in sequence processing, most classic MSA software runs as a command symbol. It is very inconvenient for non-bioinformatics analysts. **BioAider packed three MSA software**

(Mafft, Muscle and Clustal-Omega) in the graphical interface, **and provided translation-alignment additionally based on multiple sets of codon tables.**

3.3.1. Mafft

Mafft is a very popular MSA software with higher comparison accuracy, and its comparison speed is also relatively good. Some common parameters were encapsulated into the graphical interface in BioAider, and other parameters also could be added flexibly. More detailed information about Mafft could be got from <https://mafft.cbrc.jp/alignment/software/>.

Open file

Tip: Drag a **fasta format** sequence and drop here

Parameter

Datas type: ☒ Nucleotide ☐ Codon ☐ Protein

Codon Table: 1.The Standard Code

Output format: 3. Fasta format / Sorted

Align strategy: 1. --auto

Thread: 1

☐ Others parameter

Run and Profess

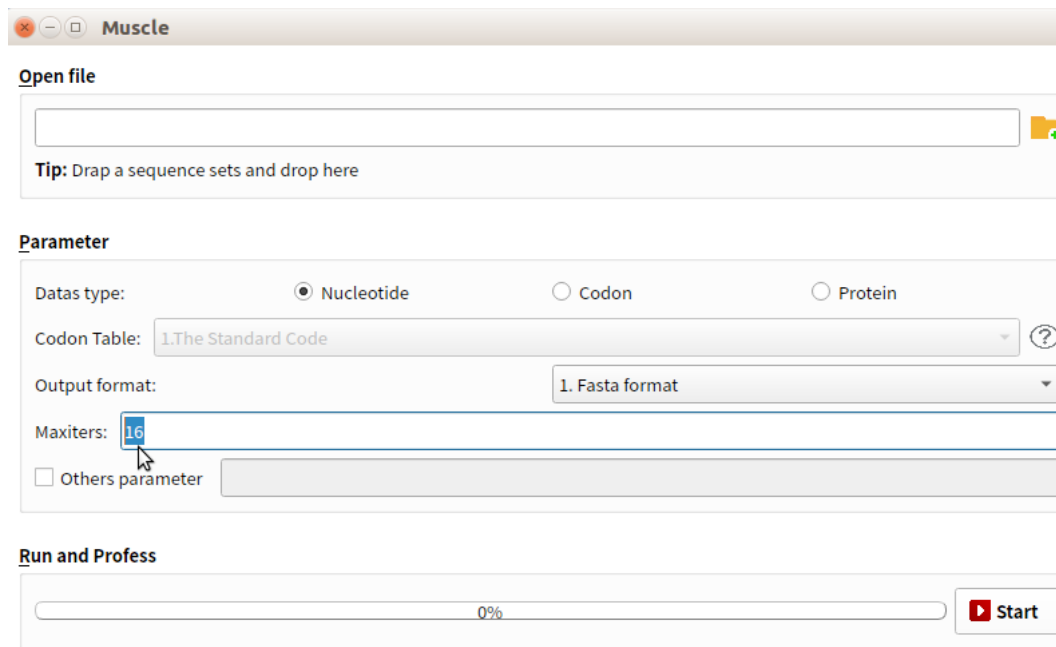
0% Start

3.3.2. 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 please reference

<http://petrov.stanford.edu/software/src/muscle3.6/muscle3.6.html>.



Muscle

Open file

Tip: Drag a sequence sets and drop here

Parameter

Datas type: ☒ Nucleotide ☐ Codon ☐ Protein

Codon Table: 1.The Standard Code

Output format: 1. Fasta format

Maxiters: 16

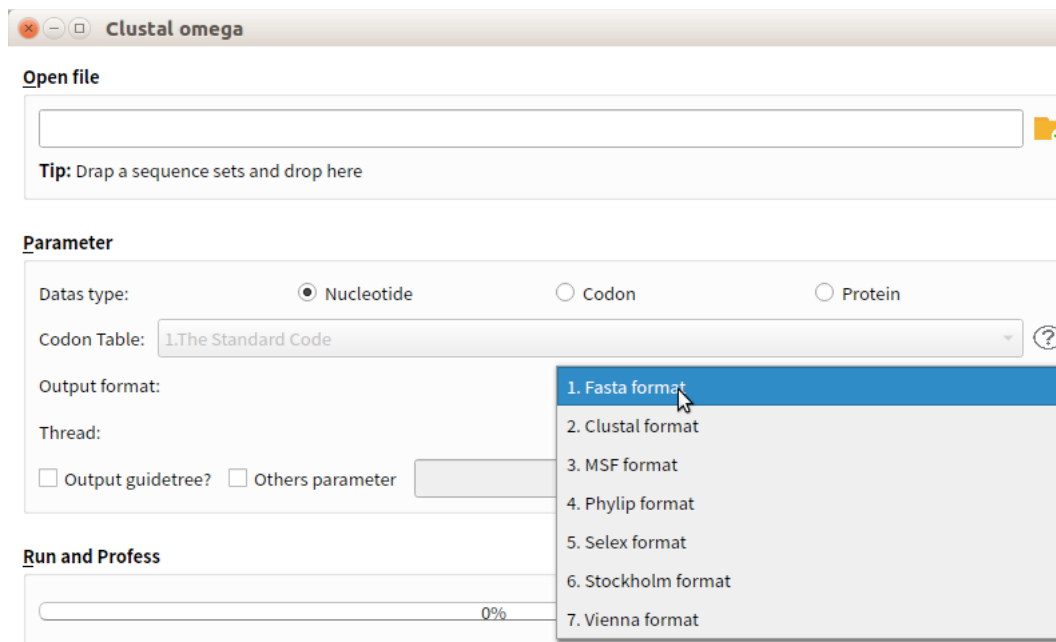
☐ Others parameter

Run and Profess

0% **Start**

3.3.3. Clustal-Omeg

As a relatively classic MSA software, *Clustal* has 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 reference <http://www.clustal.org/omega/>.



Clustal omega

Open file

Tip: Drag a sequence sets and drop here

Parameter

Datas type: ☒ Nucleotide ☐ Codon ☐ Protein

Codon Table: 1.The Standard Code

Output format:

- 1. Fasta format
- 2. Clustal format
- 3. MSF format
- 4. Phylip format
- 5. Selex format
- 6. Stockholm format
- 7. Vienna format

Thread:

☐ Output guidetree? ☐ Others parameter

Run and Profess

0%

3.4. *Mutation Tools*

3.4.1. *Mutation Analysis*

This function could be used for analysis of the **mutations characteristics 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 frequency 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*":

Mutation Analysis

Open file

C:/Users/j/Desktop/BioAider_v1.423_win-20220321/Example/SARS-CoV-2_ORF3a_aligned.fas

Tip: The first sequence in the data set will be used as the reference sequence for mutation analysis

Choose an analysis method

☒ Single Mutation
☐ Linked (Multiply) Mutation

Parameter

Dats type: ☐ Nucleotide ☒ Codon ☐ Protein

Codon Table: 1.The Standard Code

Output substitution frequency distribution? Yes
Delete both Sys-Nonsys nt sites? Yes

Groups of substitution frequency

Run and profess

100%

Start

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/VIDC-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
4	8	1	Nonsynonymous	TTT to CTC (F to L)	No	1
5	8	3	Synonymous	TTT to TTT	No	1
6	13	1	Nonsynonymous	GTA to TTA (V to L)	No	8
7	14	2	Nonsynonymous	ACT to ATT (I to I)	No	7
8	15	3	Nonsynonymous	TTG to TTT (L to F)	No	1
9	19	3	Synonymous	GAA to GAG	No	2
10	26	1	Nonsynonymous	TCA to CCA (S to P)	No	1
11	27	3	Synonymous	GAT to GAC	No	3
12	28	1	Synonymous	TTT to TTC	No	3
13	31	1	Nonsynonymous	GCT to ACT (A to T)	Yes	2
14	34	2	Nonsynonymous	ACG to ATG (T to M)	Yes	1
15	34	3	Synonymous	ACG to ACT	No	1
16	36	3	Synonymous	CCG to CCT	No	15
17	36	3	Synonymous	CCG to CCA	No	2
18	43	3	Synonymous	TTC to TTT	No	1
19	46	1	Nonsynonymous	CTT to TTT (L to F)	No	2
20	49	1	Nonsynonymous	GGC to TGC (G to C)	No	1
21	49	3	Synonymous	GCC to GGT	No	2
22	50	2	Nonsynonymous	GTT to GCT (V to A)	No	1
23	54	1	Nonsynonymous	GCT to TCT (A to S)	No	1
24	54	3	Synonymous	GCT to GCG	No	1
25	54	2	Nonsynonymous	GCT to GTT (A to V)	No	3
26	56	2	Nonsynonymous	TTT to TGT (F to C)	Yes	1
27	57	3	Nonsynonymous	CAG to GAT (Q to H)	Yes	578
28	59	2	Nonsynonymous	GCT to GAT (A to D)	Yes	2
29	60	3	Synonymous	TCC to TCT	No	4
30	60	2	Nonsynonymous	TCC to TTC (S to F)	Yes	1
31	61	3	Nonsynonymous	AAA to AAC (K to N)	Yes	2
32	65	3	Synonymous	CTC to CTT	No	1
33	67	3	Nonsynonymous	AAG to AAT (K to N)	Yes	5
34	67	3	Nonsynonymous	AAG to AAC (K to N)	Yes	2
35	68	1	Synonymous	AGA to CGA	No	1
36	74	1	Nonsynonymous	TCC to GCC (S to A)	Yes	1
37	75	1	Nonsynonymous	AAG to GAG (K to E)	Yes	2
38	76	1	Nonsynonymous	GGT to AGT (G to S)	No	1
39	78	1	Nonsynonymous	CAC to TAC (H to Y)	Yes	1
40	84	3	Synonymous	CTG to CTT	No	1
41	85	1	Synonymous	TTG to CTG	No	1
42	87	2	Nonsynonymous	TTT to TAT (F to Y)	Yes	1
43	88	2	Nonsynonymous	GTA to GCA (V to A)	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
		codon_site8	codon_site13	codon_site14	codon_site15	codon_site19	codon_site26
1							
2	Synonymous	1	0	0	0	2	0
3	Nonsynonymous	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
	Codon	Base index	Nucleotide site	Type	Substitution frequency
1					
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:

	A	B	C
1	Group1	0	1
2	Group2	1	5
3	Group3	5	10
4	Group4	10	15
5	Group5	15	20
6	Group6	20	3040
7			
8			

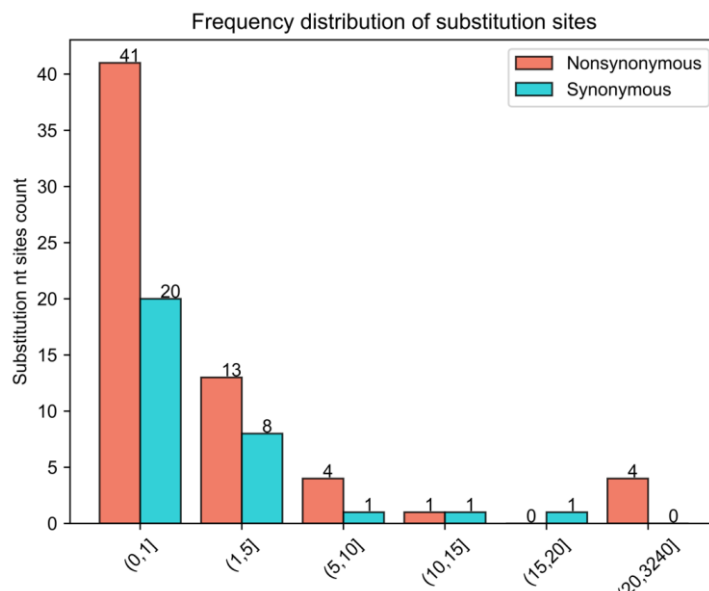
Each group of substitution frequency contains start value and end value which are separated by tab symbol. **Note, the start value** of each group is not included in the range

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

Then copy them to the textedit box of BioAider,

0	1
1	5
5	10
10	15
15	20
20	3240

You could also know 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.

seq name							
A	B	C	D	E	F	G	H
Reference sequence: hCoV-19/Wuhan/VI-01/2019/EPI_ISL_402119							
Site (codon)	Index of nt	seq name	mutation type	mutation	changed in properties of aa?	changed type of properties	
1	9	[1337]hCoV-19/Netherlands/NA_30/2020/EPI_ISL_415487	Nonsynonymous	TTC to CTC (F to L)	No	No	
5	8	[1985]hCoV-19/France/Pollinay_1733/2020/EPI_ISL_416745	Synonymous	TTC to TTT	No	No	
6	13	[474]hCoV-19/USA/WA-5119/2020/EPI_ISL_417172	Unknown	GTA to NTA	No	No	
7	13	[562]hCoV-19/USA/UT-00023/2020/EPI_ISL_418965	Nonsynonymous	GTA to TTA (V to L)	No	No	
8	13	[1799]hCoV-19/Iceland/127/2020/EPI_ISL_417727	Nonsynonymous	GTA to TTA (V to L)	No	No	
9	13	[2241]hCoV-19/England/SHF-C04AC/2020/EPI_ISL_420193	Nonsynonymous	GTA to TTA (V to L)	No	No	
10	13	[2298]hCoV-19/England/20142061804/2020/EPI_ISL_420772	Nonsynonymous	GTA to TTA (V to L)	No	No	
11	13	[2331]hCoV-19/England/20126007102/2020/EPI_ISL_418751	Nonsynonymous	GTA to TTA (V to L)	No	No	
12	13	[2332]hCoV-19/England/20126006802/2020/EPI_ISL_418749	Nonsynonymous	GTA to TTA (V to L)	No	No	
13	13	[2376]hCoV-19/Australia/VIC303/2020/EPI_ISL_419994	Nonsynonymous	GTA to TTA (V to L)	No	No	
14	13	[3006]hCoV-19/Australia/VIC275/2020/EPI_ISL_419968	Nonsynonymous	GTA to TTA (V to L)	No	No	
15	14	[2088]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	[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	[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	[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	[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	[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	[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	[2716]hCoV-19/Belgium/SR-0319112/2020/EPI_ISL_420369	Nonsynonymous	TTG to TTT (L to F)	No	No	
23	19	[1535]hCoV-19/Italy/T16222/2020/EPI_ISL_420583	Unknown	GAA to GAW	No	No	
24	19	[2956]hCoV-19/Australia/VIC324/2020/EPI_ISL_420015	Synonymous	GAA to GAG	No	No	
25	19	[3142]hCoV-19/Australia/VIC135/2020/EPI_ISL_419828	Synonymous	GAA to GAG	No	No	
26	20	[2252]hCoV-19/England/SHF-C036D/2020/EPI_ISL_420206	Unknown	ATC to ATN	No	No	
27	26	[2552]hCoV-19/Canada/BCC_8896915/2020/EPI_ISL_418854	Nonsynonymous	TCA to CCA (S to P)	Yes	(polar,none-charge) to (non-polar,none-charge)	
28	27	[1381]hCoV-19/Malaysia/188407/2020/EPI_ISL_417918	Synonymous	GAT to GAC	No	No	
29	27	[3102]hCoV-19/Australia/VIC178/2020/EPI_ISL_419875	Synonymous	GAT to GAC	No	No	
30	27	[3133]hCoV-19/Australia/VIC144/2020/EPI_ISL_419841	Synonymous	GAT to GAC	No	No	
31	28	[178]hCoV-19/USA/WI-13/2020/EPI_ISL_417516	Synonymous	TTT to TTC	No	No	
32	31	[1818]hCoV-19/Hungary/mib1/2020/EPI_ISL_416426	Nonsynonymous	GCT to ACT (A to T)	Yes	(non-polar,none-charge) to (polar,none-charge)	
33	31	[2301]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	[76]hCoV-19/Wales/PHWC-2415E/2020/EPI_ISL_418149	Nonsynonymous	ACG to ATG (T to M)	Yes	(polar,none-charge) to (polar,none-charge)	
35	34	[300]hCoV-19/USA/WA-UW215/2020/EPI_ISL_417364	Synonymous	CCG to ACT	No	No	
36	36	[116]hCoV-19/USA/WI-77/2020/EPI_ISL_421335	Synonymous	CCG to CCT	No	No	
37	36	[120]hCoV-19/USA/WI-73/2020/EPI_ISL_421331	Synonymous	CCG to CCT	No	No	
38	36	[122]hCoV-19/USA/WI-71/2020/EPI_ISL_421329	Synonymous	CCG to CCT	No	No	
39	36	[123]hCoV-19/USA/WI-70/2020/EPI_ISL_421328	Synonymous	CCG to CCT	No	No	
40	36	[124]hCoV-19/USA/WI-69/2020/EPI_ISL_421327	Synonymous	CCG to CCT	No	No	
41	36	[126]hCoV-19/USA/WI-67/2020/EPI_ISL_421325	Synonymous	CCG to CCT	No	No	
42	36	[127]hCoV-19/USA/WI-66/2020/EPI_ISL_421324	Synonymous	CCG to CCT	No	No	
43	36	[132]hCoV-19/USA/WI-61/2020/EPI_ISL_421319	Synonymous	CCG to CCT	No	No	

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*".

3.4.2. 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:

Open file

Data type

☒ Nucleotide ☐ Protein

Progress

0%

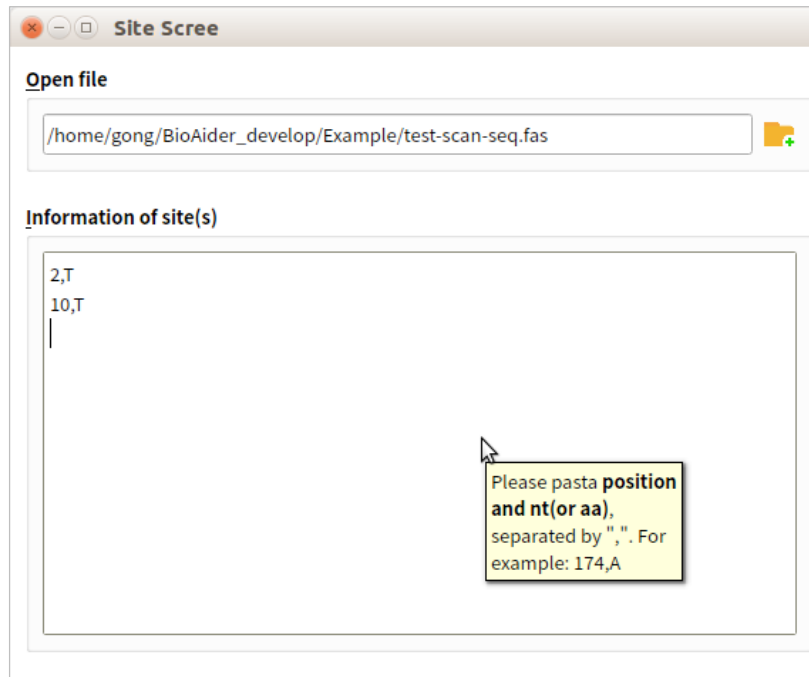
Start

Click run

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

3.4.3. Site Scree

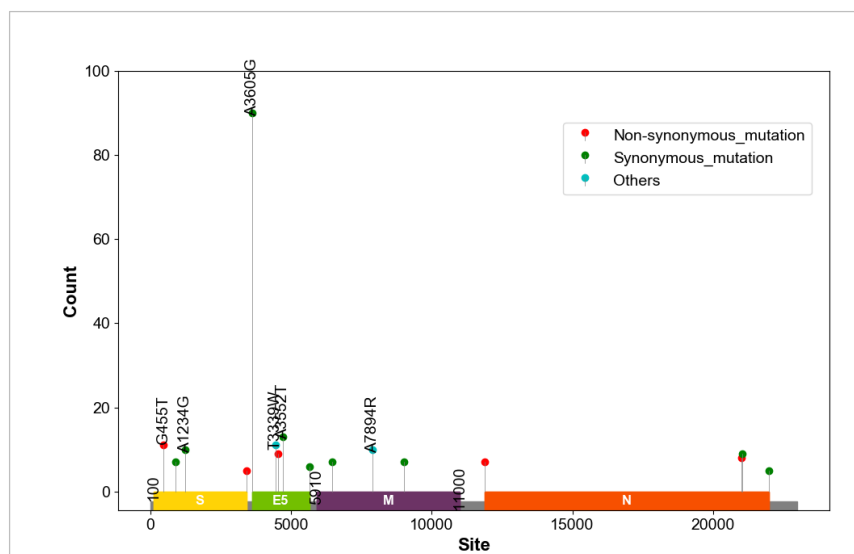
This function is used to extract the sequences with corresponding base (or amino acid) in *specified one or more* site(s). It is very useful for studying whether there is linkage inheritance among different gene sites.



3.5. Drawing module

3.5.1. 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
1	100	S	Start	11	100	#FFD306
2	455	S	Non-synonymous_mutation	11	G455T	#FFD306
3	900	S	Synonymous_mutation	7	C900A	#FFD306
4	1234	S	Synonymous_mutation	10	A1234G	#FFD306
5	3400	S	Non-synonymous_mutation	5	G3400A	#FFD306
6	3605	E5	Synonymous_mutation	90	A3605G	#73BF00
7	4439	E5	Others	11	T3339W	#73BF00
8	4540	E5	Non-synonymous_mutation	9	G3440C	#73BF00
9	4700	E5	Synonymous_mutation	13	A3552T	#73BF00
10	5653	E5	Synonymous_mutation	6	C5653G	#73BF00
11	5910	M	Start	7	5910	#6C3365
12	6444	M	Synonymous_mutation	7	G6444C	#6C3365
13	7894	M	Others	10	A7894R	#6C3365
14	9004	M	Synonymous_mutation	7	G9004C	#6C3365
15	11000	M	End	2	11000	#6C3365
16	11894	N	Non-synonymous_mutation	7	G11894A	#F75000
17	21004	N	Non-synonymous_mutation	8	G21004T	#F75000
18	21984	N	Synonymous_mutation	5	A21984G	#F75000
19	21029	N	Synonymous_mutation	9	C21029T	#F75000
20						

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:

Sunny super Lollipop

Examples data

Axis
 X_max: ; X_step: ; Y_max: ; Tick font size: ; Title size:

Legend
 Horizontal deviation: ; Vertical deviation: ; Font size:

Gene block
 High: ; Font size:

Lollipop
 Rotation: ; Line width: ; Label size:
☒ Show mutation sites? Threshold: Circle distance:

Start

3.5.1. Commonly used statistics

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

Point & Histogram & Violin

INPUT FILE

Examples data

SPECIFY VARIABLE(COLUMN NAME IN MATRIX)

X-Variable **Group** ; Y-Variable **Value** ; Kinds **Sample**

TYPE OF CHART

☐ Scatter ☐ Bar chart ☐ Boxplot ☒ Violinplot

SCATTER

☒ Jitter ☐ No Jitter ☐ Swarm

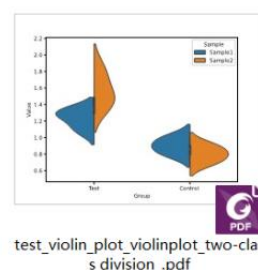
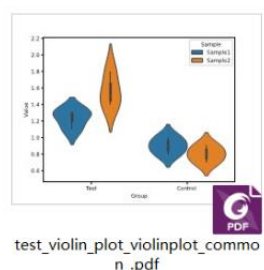
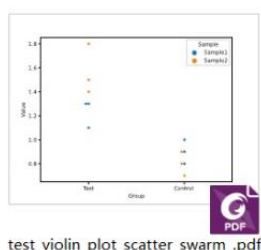
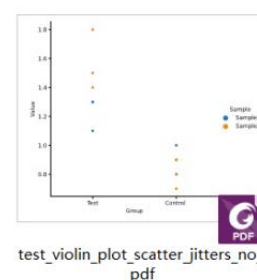
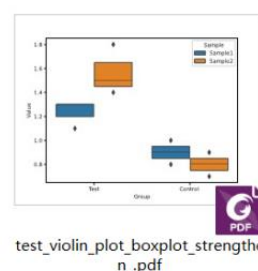
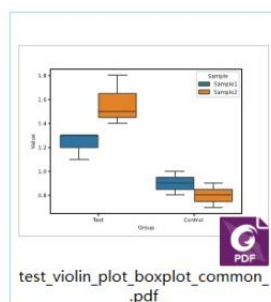
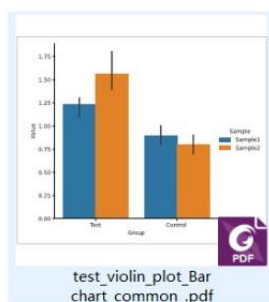
BOXPLOT

☒ Common ☐ Strengthen

VIOLINPLOT

☒ Common ☐ Two-class division

PLOT



4. Test Datas

Examples and test are available

at: <https://github.com/ZhijianZhou01/BioAider/tree/master/Example>.

5. Ciation

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](#)