FasParser Manual

(2017-3-28::Version 1.1.0)

Sun Yan-Bo

Kunming Institute of Zoology

Chinese Academy of Science

Kunming, Yunnan, China

INTRODUCTION

With the development of sequencing technology in recent times, a great number of molecular sequences (DNA and RNA) have been generated. Molecular analyses based on these sequences have become one of the most important measures for assessing their potential biological significance. The increase in the amount of available sequence data has made its manipulation tricky, especially for those without programming experience. Hence, it has now become necessary to develop one or more user-friendly software to perform such analyses in a **batch mode**, like sequence extraction and filtration, sequence translation, and file format conversion.

Herein, we provide a new program package named '**FasParser**' for manipulating sequence files. It is designed with a user-friendly GUI and also batch processing modes, which allows users to handle multiple sequence files in a simple way. Presently, the package involves 8 main programs/functions (Figure 1) *viz:* (1) counting and viewing the differences between two sequences at both DNA and codon levels, (2) identifying the overlapped columns between two alignments (of a same gene), (3) sorting sequences according to ID, sequence length, or ID list provided by user, (4) concatenating sequences for a particular set of samples from multiple sequence files, (5) batch translating DNA files to protein ones, (6) constructing alignments with different formats, (7) extracting and filtering sequences according to ID or sequence length, and (8) get ORF seqs for cDNA.

1. INSTALLATION

The '**FasParser'** has been developed into a standalone **Windows System Application** (compiled and tested on Windows 7/10). It can run on most Windows systems with no dependence of other programs.

Download the **setup program** (i.e. '***FasParserX.X_setup.exe***') from https://github.com/Sun-Yanbo/FasParser to your disk, and then double click it to install the whole package. Normally, it is well using the default installation parameters (by clicking the Next button to end, **Figure 1**). After successful installation, you would get a screen of the Home page of this package (**Figure 2**).

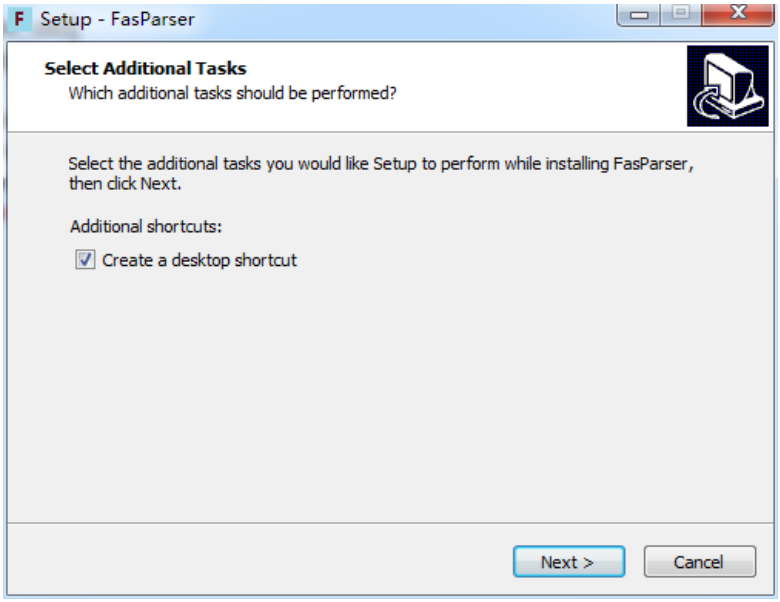


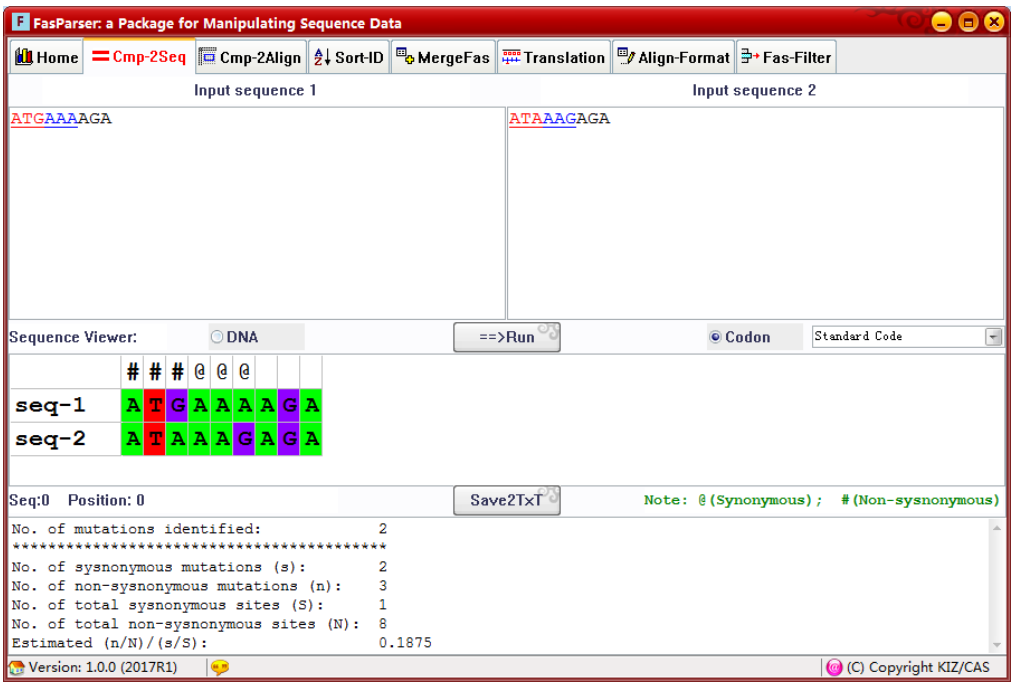
Figure 1. Installation of the FasParser package.

Figure 2. The Home page of the FasParser package.

2. PACKAGE USAGE

1) Sequence comparison and mutation identification (Cmp-2Seq)

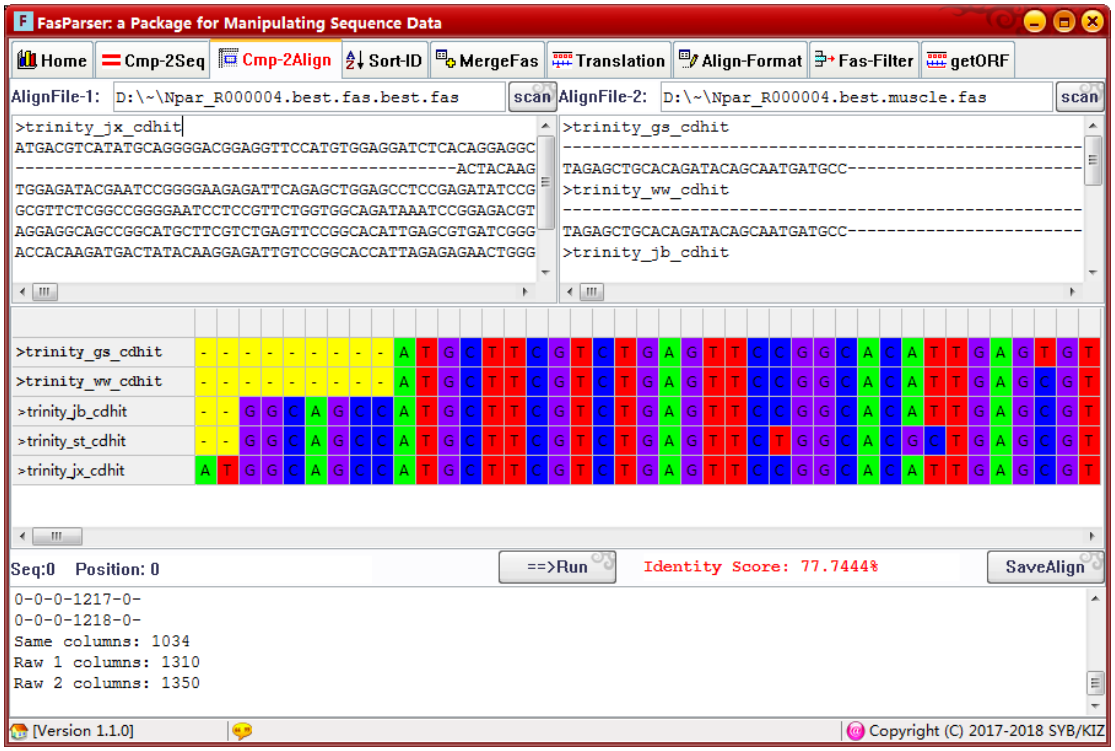
The program "Cmp-2Seq" in the FasParser package was designed to count and view differences between two DNA sequences at both DNA and codon levels. Under the codon level, the program can also estimate the total number of synonymous (S) and non-synonymous (N) sites for the first sequence and then calculate the number of synonymous and non-synonymous substitutions between the two sequences. To do that, users just need to **put two sequences into the two above textboxes** and **click the Run button**. The program could then provide a view the differences between the two sequences in colors and also the summary of identified mutations or substitutions in below region.



2) Alignment comparison and overlapped columns identification (Cmp-2Align)

This program "Cmp-2Align" was designed to compare different alignments of a same gene that might be generated by different aligners. It is well known that there is almost no current method correctly aligns the entire sequence, and different aligners always correctly align different regions. One simple method is to identify the overlapped regions between different aligner-generated alignments, which might be useful for some other analyses, like phylogenetic reference and/or positive selection detection. This function needs users to **input two alignments through the above**

scan buttons and **click the Run** to view their overlaps. The **SaveAlign button** could save the overlapped regions into an alignment file (in FASTA format).



3) Sequence Sorting (Sort-ID)

This program will allow users to sort their FASTA files according to either the **ID names**, **sequence lengths**, or **a provided list of IDs**. Part of this function (with the ID list provided by users) is much similar to the extraction analysis in "Fas-Filter" section. Please note that **the ID is recognized from the raw ID as the first continuous string before symbols space, ".", and "|" (Do not include these symbols)**. For example, the raw ID in a FASTA file is:

'>Uma_R000001.2 locus=scaffold79:384179:406202:-'

You can use the ID 'Uma_R000001.2' to search its sequence, sometime the the ID 'Uma_R000001' is also ok if there is only targeted ID. If the provided IDs cannot be recognized, there will be no sequence reported. Another example:

'>gi|947195581|ref|XP_006139827.2| PREDICTED…'

You can use 'gi|947195581' to search the target sequence.

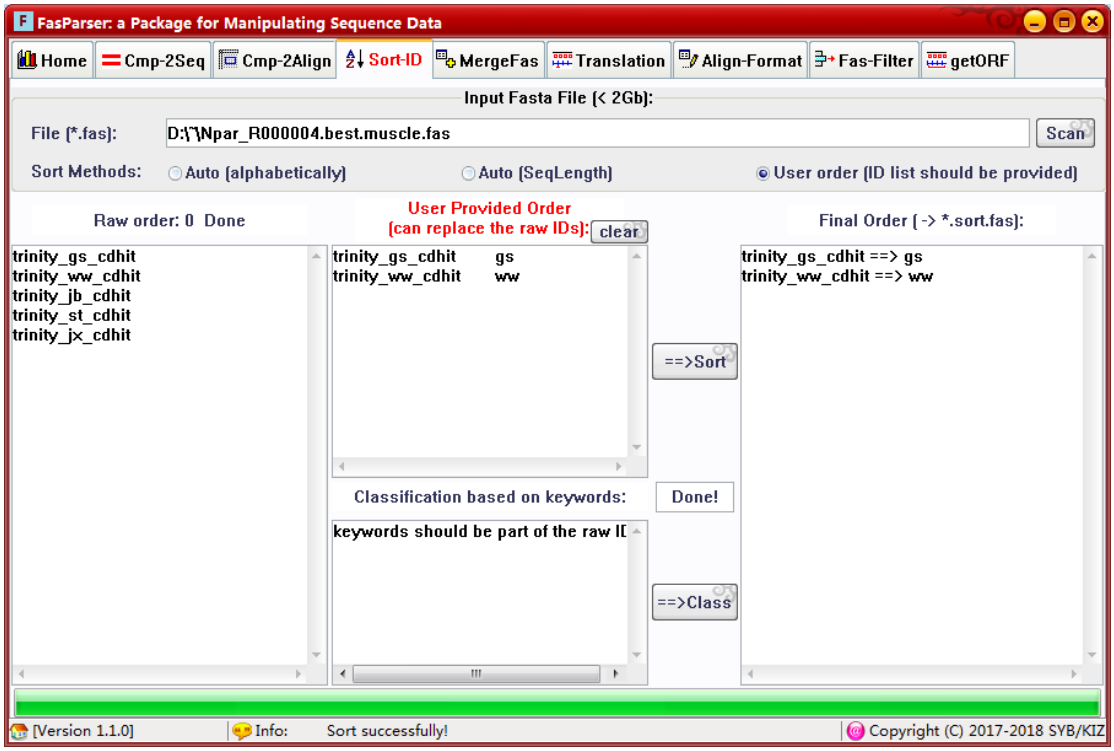
With the provided ID list, the program can also rename the raw sequence IDs. To achieve that, the ID list should have two columns, the first column refers to the raw ID, and the second column refers to the new ID, the separator between the two columns can be spaces or Tab(s), as below:

Uma_R000001.2 R000001.2

```
Uma_R000003.2
Uma_R000002.2    R000002
...
```

Under the above query IDs, program will use `R000001.2` to replace `Uma_R000001.2`. If there is only one column, the raw IDs will not be changed.

In addition, this section also provides a classification function, which means that you can use one or more keywords to extract sequences from a raw FASTA file and save them into separate FASTA files. The keywords can be the genus name, species name, or some other words. **Please NOTE that all the keywords should be part of the raw IDs.**



4) Sequence concatenation (MergeFas)

This program is used to concatenate sequences of the same IDs from multiple FASTA files. It is much useful in phylogenetic or other analyses, especially when users generated multiple loci sequences for a particular set of samples, and want to derive a "super" sequence by concatenating all the loci sequences together for each sample.

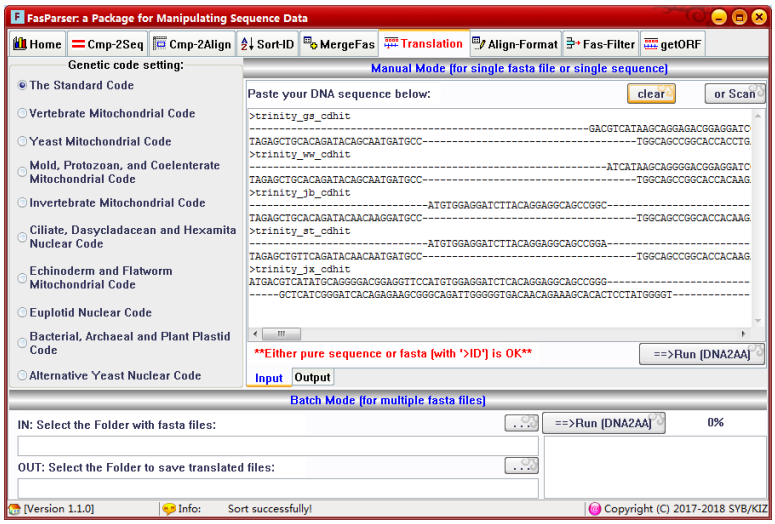
This program can be run in a batch mode, for which user should **define a folder** which contains all the FASTA files to merge. If the raw FASTA files are not aligned before, the final concatenated FASTA file should be aligned first before conducting some other analyses. (We recommend user to construct alignment before for each file, and each file have a same ID list)

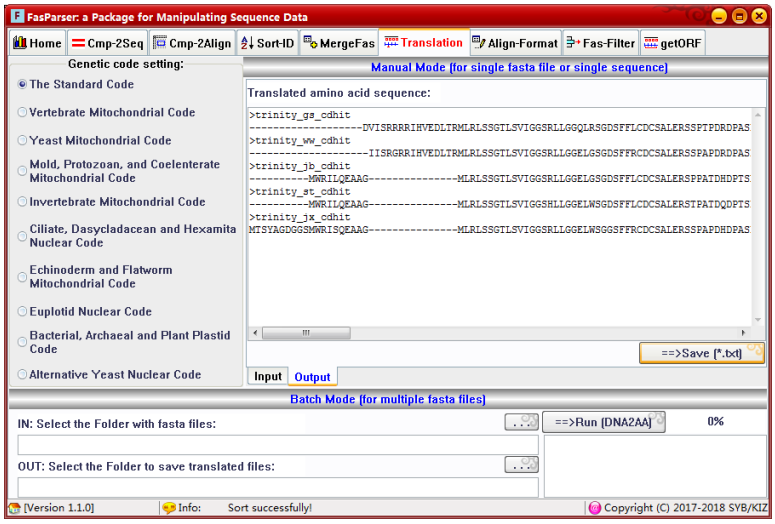
There is also **a manual mode** to merge two FASTA files. If the first file has been defined, all the second files (you can change the second file if the previous one has been merged) would be merged recursively to the first file until you click the Save button.

5) Translation from DNA to protein (Translation)

This program is used to translate the DNA sequences to protein sequences. There are also two modes available. If users have many FASTA files, you can use **the batch mode** to conduct such analyses, in which, you should **define a folder** which contains the FASTA files to translate.

In the manual mode, you can input a pure DNA sequence or a FASTA file, the program can automate recognize them.

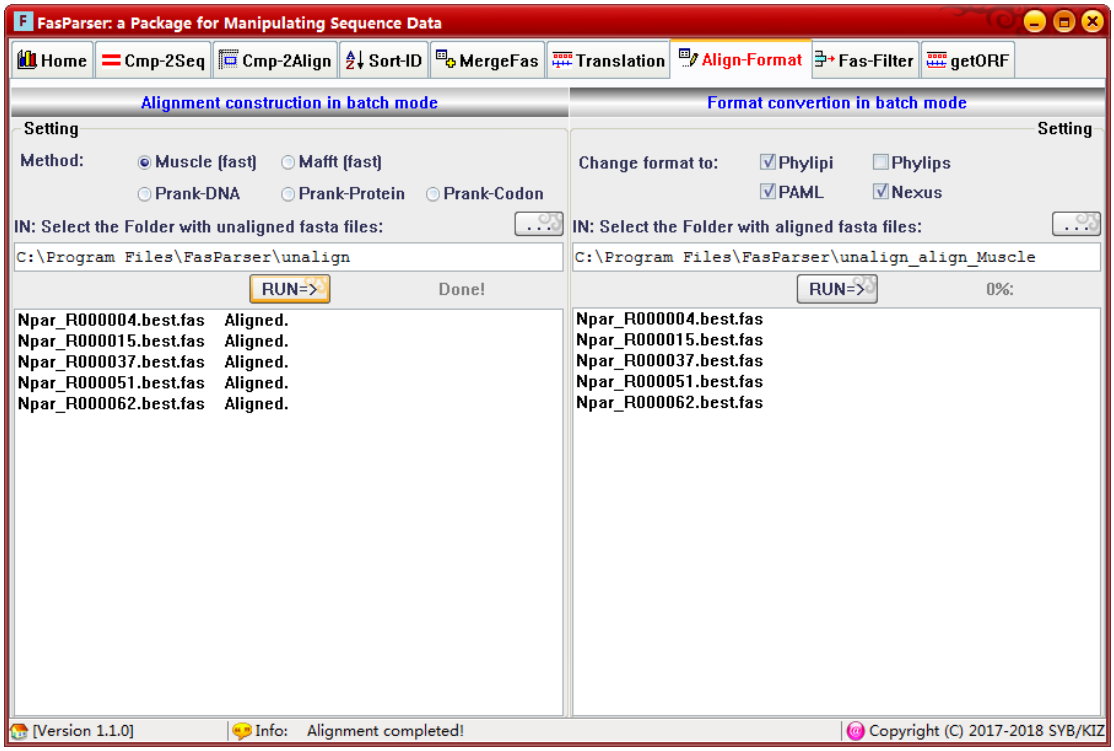




6) Alignment construction and format conversion (Align-Format)

This program is designed to **construct the alignment for multiple FASTA files** and **convert the alignment format to others**. There is only a batch mode available for this program. There are 3 aligners (MUSCLE, MAFFT, and PRANK) have been integrated into the FasParser package for use; users can run them without installing them.

After alignment construction, you can use the format conversion function to convert the FASTA files to other formatted ones. There are 4 different formats available: FASTA, PHYLIP, PAML, and NEXUS. Only a batch mode is available. So, you should **define a folder name** which contains all the aligned FASTA files to perform this analysis.

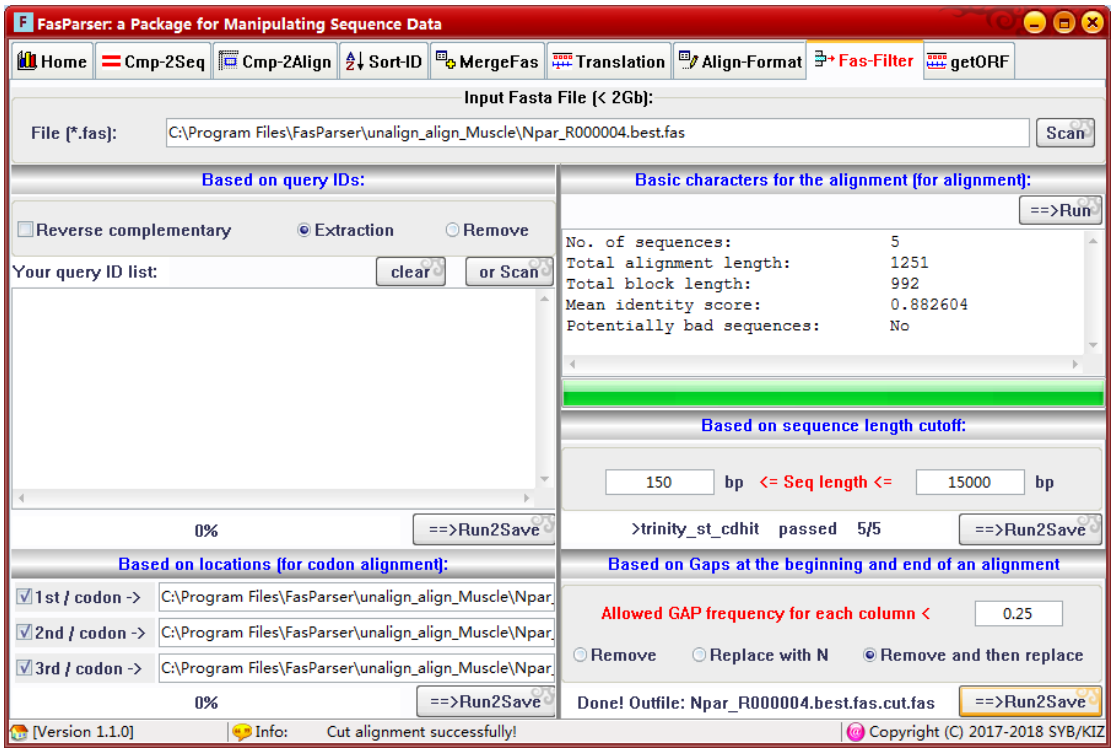


7) Extraction and filtration (Fas-Filter)

This program is designed to perform some **extraction and filtration analyses** within a particular FASTA file. It is a much common sequence file operation. With this program, users can **extract or remove a set of sequences** from the raw FASTA file based on **query IDs** and the positions per codon, and can also **filter the raw FASTA** file according to the **sequence length**.

In this program, we also provide a function to cut the raw alignment by removing the columns with many gaps ('-'). This function can also fill up the missing data (with 'N') at the beginning and end of an alignment, which is useful for some phylogenetic analyses. In addition, the 'FasParser' can also provide a **statistic summary** of a raw alignment, such as showing if there are one or more bad sequences (short) in an alignment and the length of gap-free blocks.

The way to recognized input IDs is same as that in “Sort-ID” section, like the total ID, or strings before symbols “ ”, “.”, and “|”.



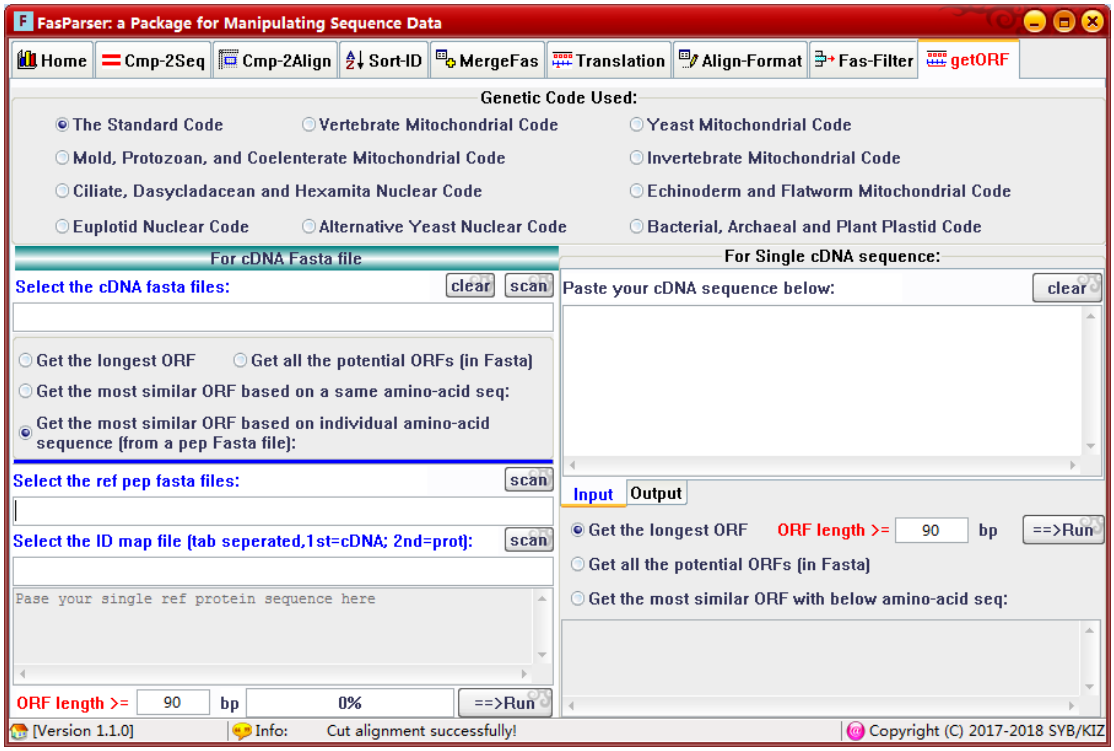
8) Open reading frame identification (getORF)

This program is designed to **identify the ORF for cDNA sequence(s)**. The cDNA sequences can be organized as a FASTA file or single sequence. There are three choice to obtain the ORFs: a) is to get the longest one if there are present more than one potential ORF in the input cDNA sequence; b) is to get all the potential ORF sequences; and c) to get the most similar ORF to one/more protein sequences.

For identifying ORFs for a FASTA file, if all the sequences in the file encode homologous proteins, only **a single reference protein sequence** (paste into the below textbox) can be used to get the homologous ORFs for all the sequences. If the sequences in FASTA file encode different proteins, like all the cDNA sequences of human genome, the **reference protein sequences** should be also save into FASTA file and then taken as another input for this program. Moreover, an **ID map file** should be also provided to tell the program which protein sequences is used as reference seq for which cDNA. For example:

```
XM_014583262    XP_014438748
XM_006166216    XP_006166278
XM_006141264    XP_006141326
...    ...
```

Please note that the 1st and 2nd columns must be IDs for cDNA and protein, respectively.



3. REFERENCES

If you want to cite the FasParser package, please use this reference:

Yan-Bo Sun. FasParser: a package for manipulating sequence data, Zoological Research 38(2): 110-112, 2017

Below is a non-exhaustive list of publications related to the FasParser package and the programs it integrates:

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5: 113.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30: 3059-3066.

Loytynoja A & Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America, 102: 10557-10562.

Nei M & Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution, 3: 418-426.