FasParser Manual

(2017-3-7::Version 1.0.2)

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 seven 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, and (7) extracting and filtering sequences according to ID or sequence length.

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. '***FasParser1.0_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, **Figure 1**). And, after finishing this, you would get a screen of the Home page of this package (**Figure 2**).

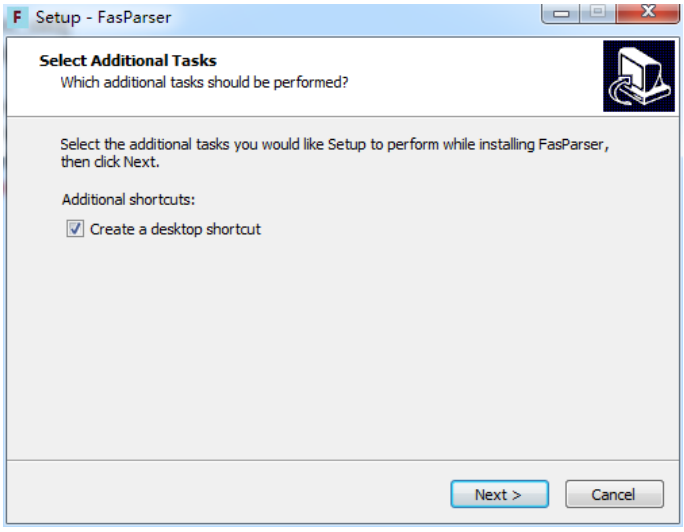


Figure 1. Installation of the FasParser package.

Figure 2. The Home page of the FasParser package.

2. Program 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 the two sequences into the two textboxes** and **click the Run button**. The program could then provide a view the differences between the two sequences and also the identified mutations or substitutions in the below (**Figure 3**).

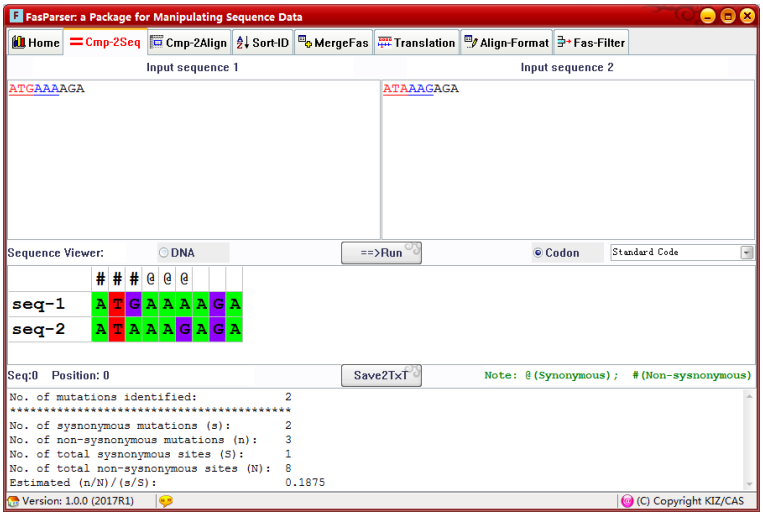


Figure 3. Overview of the Cmp-2Seq program.

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, **Figure 4**).

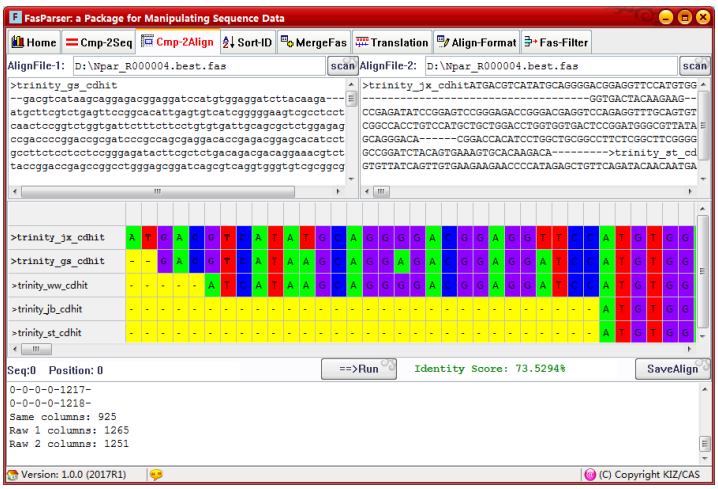


Figure 4. Overview of the Cmp-2Align program.

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 first continuous string of the raw ID. 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.

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 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. If the raw FASTA files are not aligned before, the final concatenated FASTA file should be aligned first before conducting some other analyses.

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

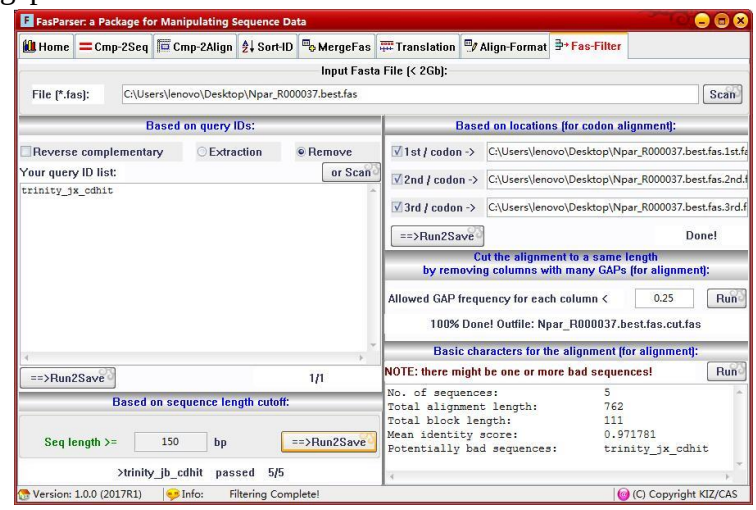
the FASTA files to translate.

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, so users can run them without installing them.

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 ('-'). 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.



3. Citation

If you use this package in your studies, please cite this paper:

Sun Yan-Bo 2017 FasParser: a package for manipulating sequence data, Zoological Research

Sun Yan-Bo