Bash script to compute sequence fragment sizes in a FASTA file.

This is a Bash program called **seq_size_freq_bash.sh** that computes the size of DNA sequence fragments in a FASTA file and prints the sequence ID and fragment size in base pairs in an output file. It is designed to be run from the Linux Ubuntu command line. Put the program in the same folder as the input file. Then navigate to the folder in question, provide the command to make the script executable, and run it from the command line. See details on how to do this below. The program requires only an input FASTA file with the sequences. The program can be created in any text editor and saved with an .sh extension.

```
#!/bin/bash

# Input and output files
input_file="1all.fasta"
output_file="seq_size_freq"

# Initialize variables
seq_id=""
seq=""

# Create or clear the output file
> "$output_file"

# Process the fasta file
awk '/^>/ {
    if (seq_id != "") {
        print seq_id, length(seq) >> "'$output_file'";
    }
    seq_id=$0;
    seq="";
  }
  /^[^>]/ {
    seq=seq $0;
  }
  END {
    if (seq_id != "") {
        print seq_id, length(seq) >> "'$output_file'";
    }
  }' "$input_file"

echo "Sequence sizes have been saved to $output_file"
```

To run it, make it executable by navigating to the folder that it is in and running the following command from the command line: `chmod +x seq_size_freq_bash.sh`

Then type: `./seq_size_freq_bash.sh`

After it runs, you should see the output file called seq_size_freq, which is a text file with two columns, one for the sequence ID and one for the fragment size.

Created by Windsor Aguirre using ChatGPT
Created: Sep/24/24