Automated System for Structural Break Detection in DNA Sequences

Manual (v1.0)

Background

ASSBD stands for Automatic System for Structural Break Detection. This is an automated system that will identify the DNA sequences breaks. Breaks includes space, comma and any other symbols. This also helps to determine the mutation, mismatches, length of the common segments. Consecutively, we can find out the any repeated segment by providing length of the data set. It will help to find out any desired length repeated segment.

Availability and implementation

Download latest version of ASSBD.jar from

https://github.com/abmmki/assbd

Contact: khademul@du.ac.bd**,** sarwar.saubdcoxbazar@gmail.com

Installation

The program needs to have Java Development Kit (J.D.K) and JRE 1.7 (or latest version) preinstalled. OS requirement is Windows 7. This program also tested in Linux (Ubuntu) OS.

The basic computer requirement is Pentium 4 and any update versions will do. Basic primary requirement is 1GB and secondary memory (Hard Disk) is 40BG.

Chapter one: Sequence Alignment

Sequence Alignment can be performing in both local and global concept. At first we have to double click on the system named ASSBD to open **(Figure 1).**

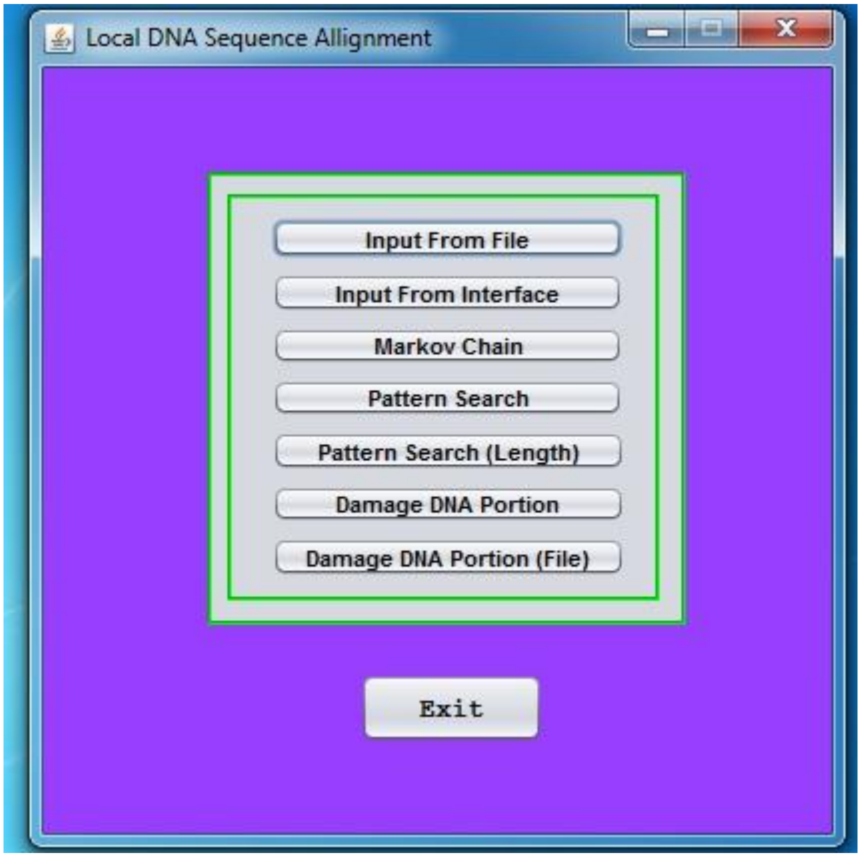


Figure 1: ASSBD Complete Interface

The process is give below in step by step process

1. Click on Input from File and new interface (Figure 2) will visible. Here we see **File name** that is data file. The **Browse** button enables us to select the file from certain position of computer. Two position buttons exist as **First position** that is from where we want to start the alignment and **Last position** indicates the value of the sequences up to length we want to align.

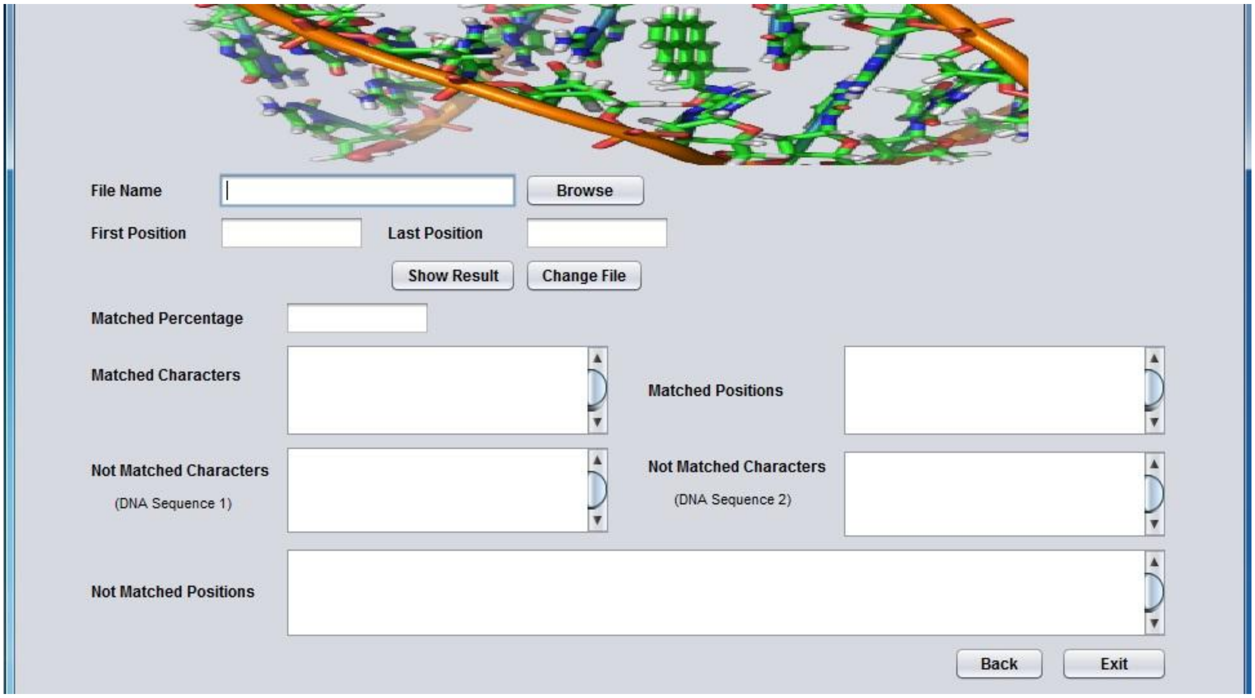
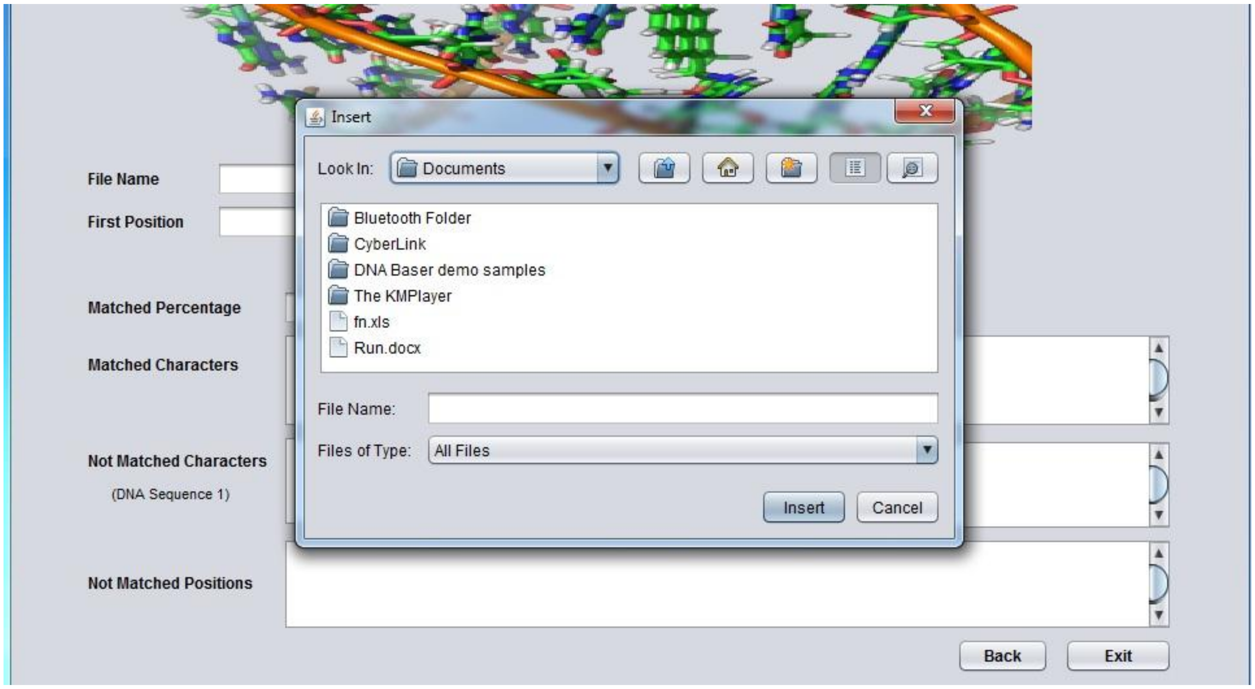


Figure 2: Input from file interface

2. Click on Browse button. New interface (**Figure 3**) will open that will enable data set from any part of the computer.



Look In: Documents

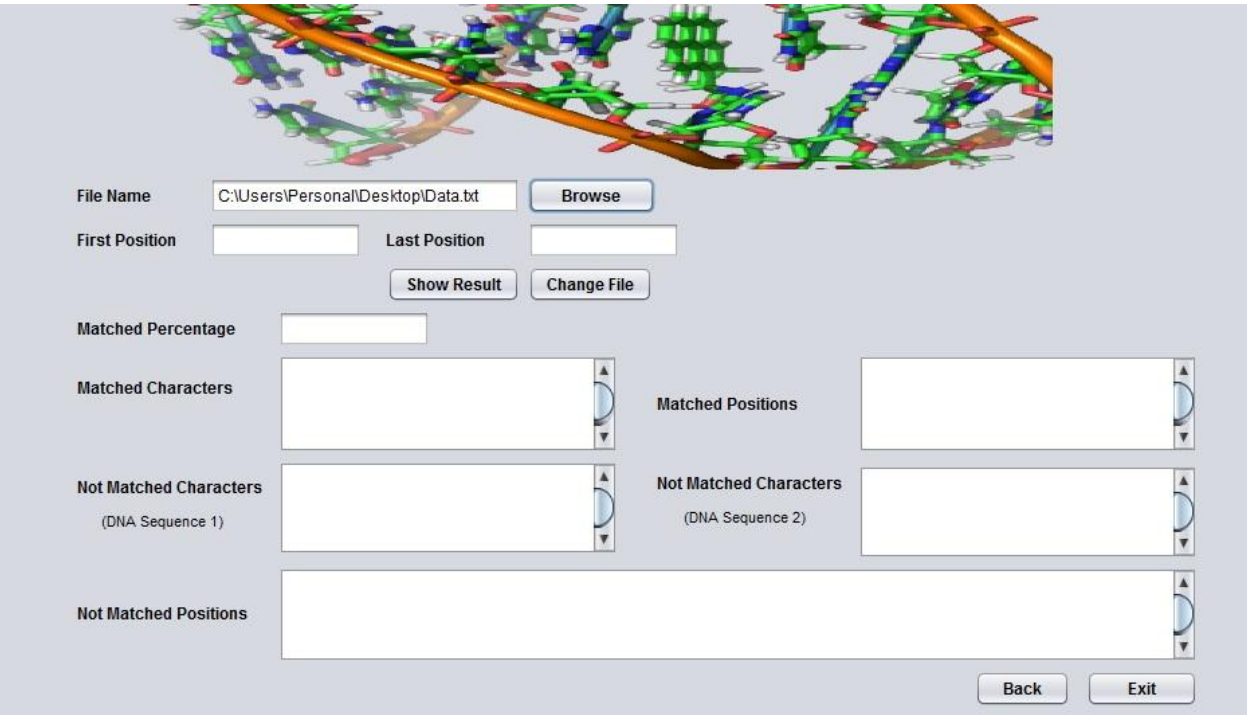
Bluetooth Folder
CyberLink
DNA Baser demo samples
The KMPlayer
fn.xls
Run.docx

File Name:

Files of Type: All Files

Figure 3: Browse Interface for data file selection

3. An example using the data set from desktop has performed (**Figure 4 and Figure 5**).



File Name C:\Users\Personal\Desktop\Data.txt

First Position

Last Position

Matched Percentage

Matched Characters

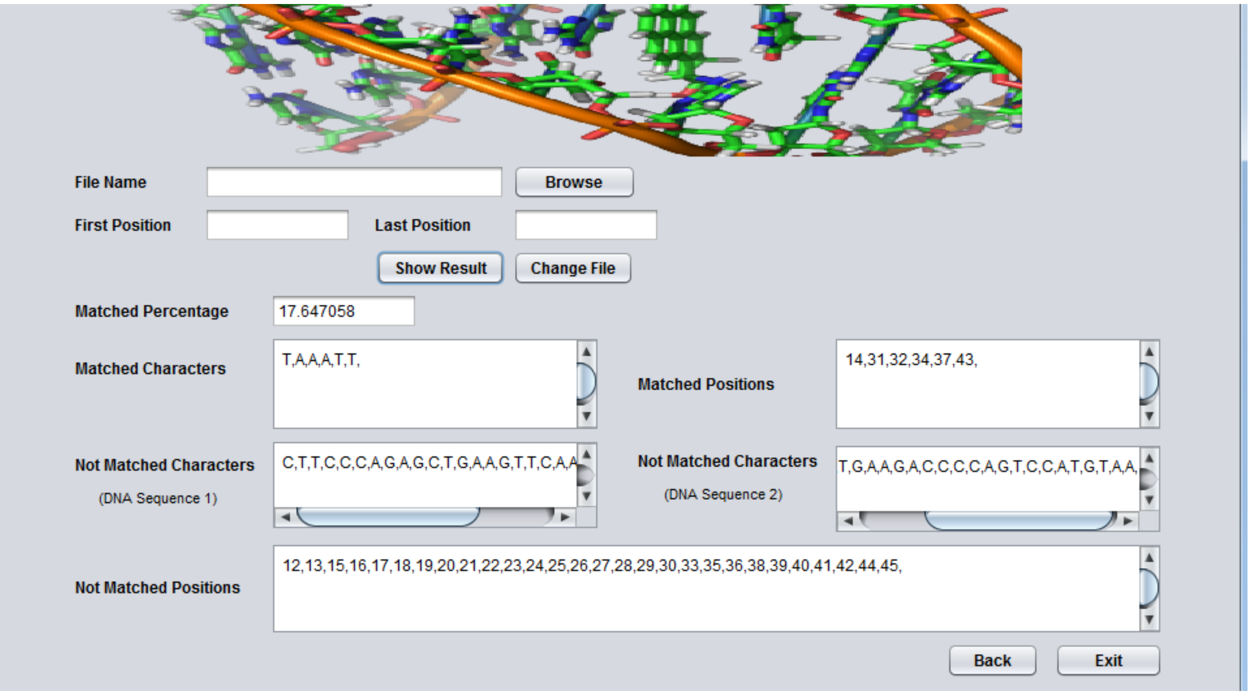
Matched Positions

Not Matched Characters (DNA Sequence 1)

Not Matched Characters (DNA Sequence 2)

Not Matched Positions

Figure 4: A data set file has selected from desktop, named Data.txt



File Name

First Position

Last Position

Matched Percentage 17.647058

Matched Characters T,A,A,A,T,T,

Matched Positions 14,31,32,34,37,43,

Not Matched Characters (DNA Sequence 1) C,T,T,C,C,C,A,G,A,G,C,T,G,A,A,G,T,T,C,A,A

Not Matched Characters (DNA Sequence 2) T,G,A,A,G,A,C,C,C,C,A,G,T,C,C,A,T,G,T,A,A,

Not Matched Positions 12,13,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,33,35,36,38,39,40,41,42,44,45,

Figure 5: Result after data set selection.

4. Same operation as 1 and 2 can be done by using keyboard input by clicking at input from keyboard button (**Figure 1**). In response of that new window (Figure 6) will open.

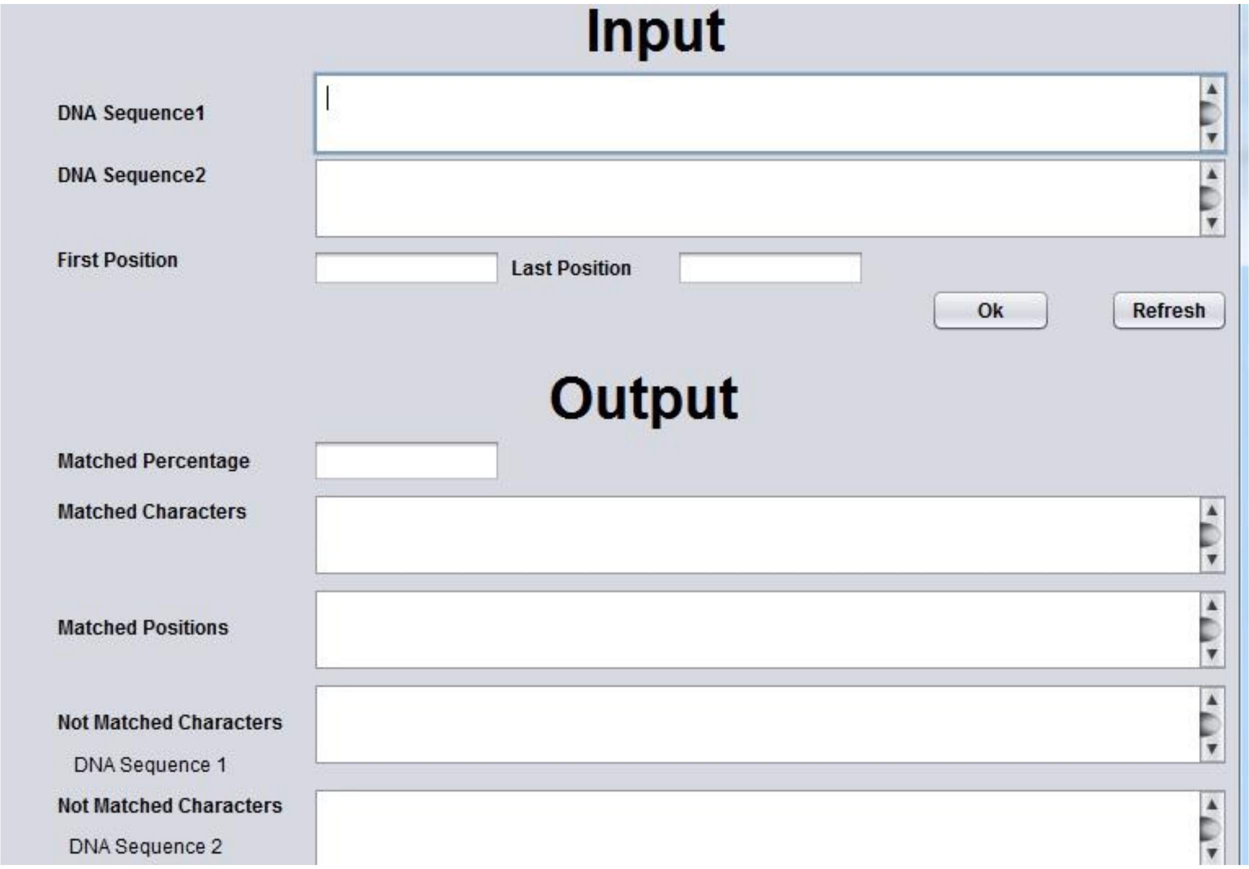
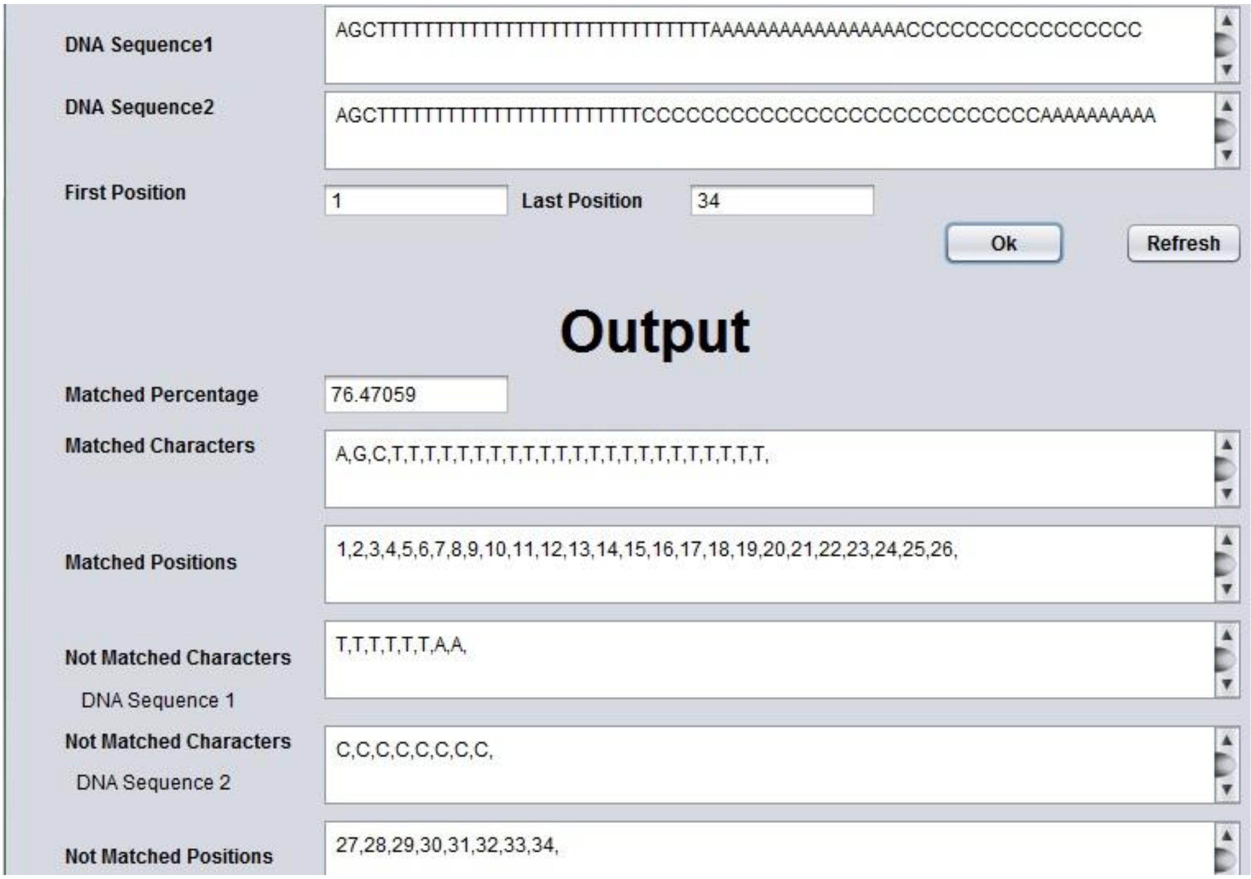


Figure 6: Input data set using Keyboard

5. An example using keyboard data set **(Figure 7**) is narrated.



DNA Sequence1 AGCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAAAAAAAAAAAAAAACCCCCCCCCCCCCCCC

DNA Sequence2 AGCTTTTTTTTTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCCCCCCCCCCCCAAAAAAAAA

First Position 1 Last Position 34

Output

Matched Percentage 76.47059

Matched Characters A,G,C,T,T,T,T,T,T,T,T,T,T,T,T,T,T,T,T,T,T,T,T,T,T,T,

Matched Positions 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,

Not Matched Characters DNA Sequence 1 T,T,T,T,T,T,A,A,

Not Matched Characters DNA Sequence 2 C,C,C,C,C,C,C,C,

Not Matched Positions 27,28,29,30,31,32,33,34,

Figure 7: An impact of Keyboard

6. Consequently, the pattern search is also an important part of ASSBD. When we click on the pattern search button (**Figure 8**) and we get the outcome (**Figure 9**). This enables the users to find specific pattern from given data set. That is, which pattern is repeatedly exist, will show the outcome. For example, given data set (**Figure 9**) has illustrated that **AGA** pattern is 21% of the total data set and its frequency is 6, i.e. there are six times **AGA** pattern.

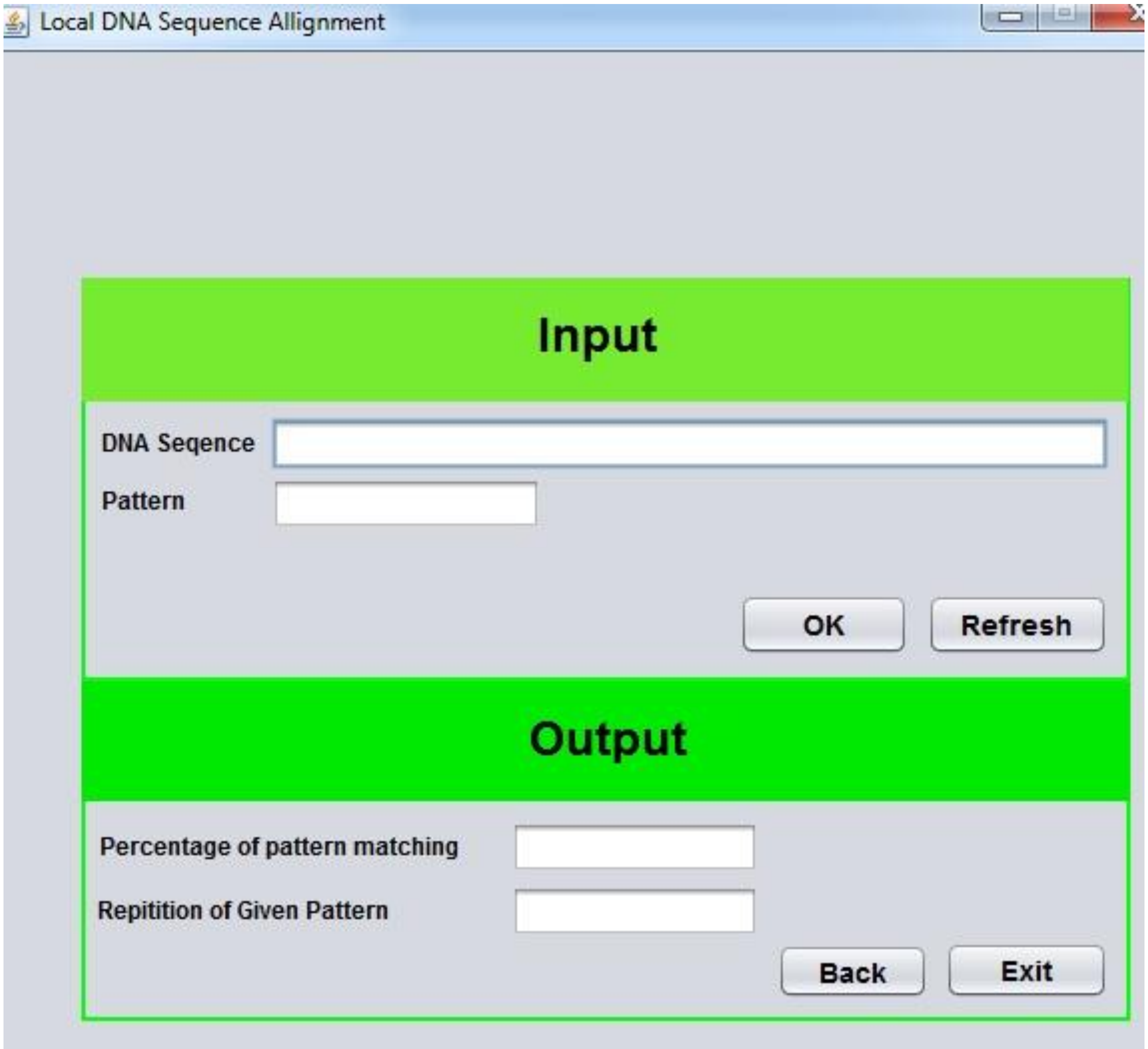
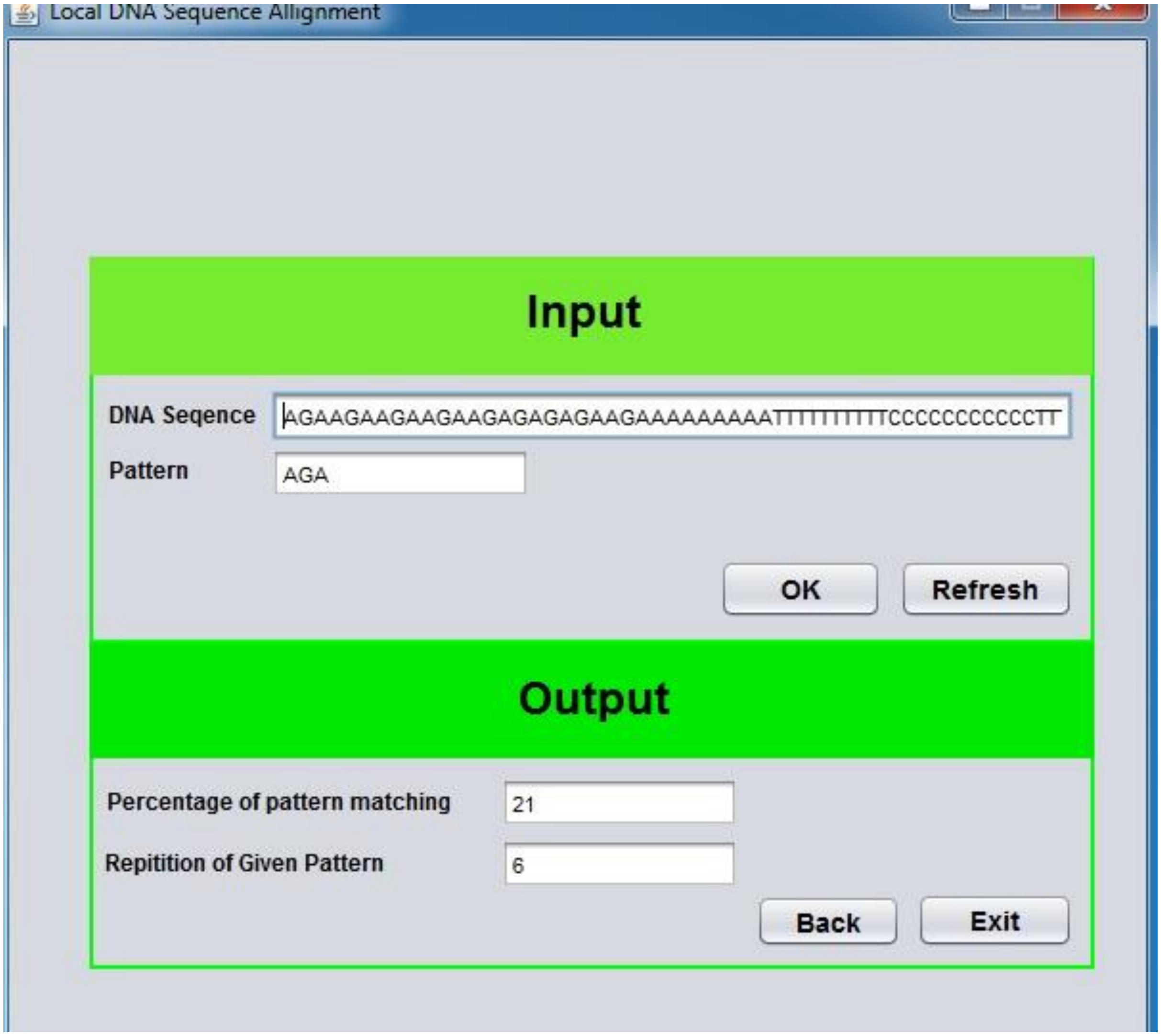

Local DNA Seqence Allignment
Input
DNA Seqence
Pattern
OK
Refresh
Output
Percentage of pattern matching
Repitition of Given Pattern
Back
Exit

Figure 8: Outline for Pattern Search



DNA Seqence AGAAGAAGAAGAAGAGAGAGAAGAAAAAAAAAATTTTTTTTTTCCCCCCCCCCCTT

Pattern AGA

Percentage of pattern matching 21

Repitition of Given Pattern 6

Figure 9: The resultant on pattern search

7. Here, we also find the repeated pattern by providing the length of the pattern (**Figure 10**). The outcome for a certain data set, we have noticed a four length pattern **AAGG** is resulted (**Figure 11**).

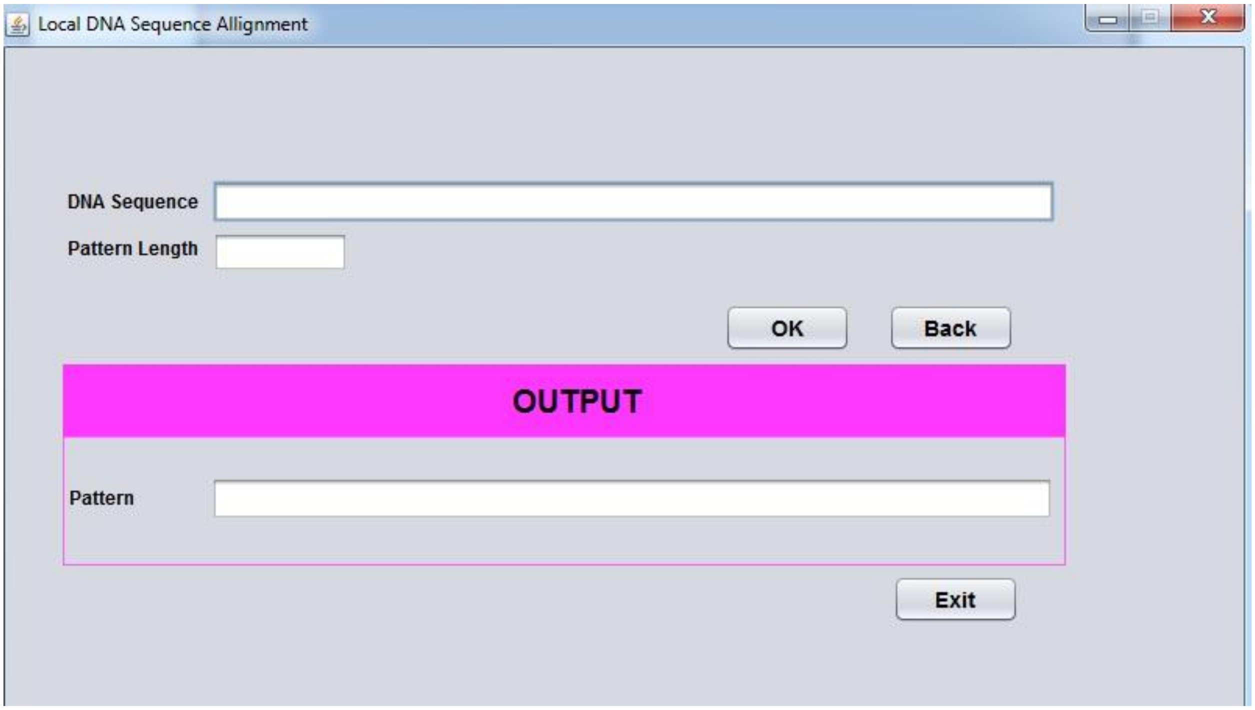
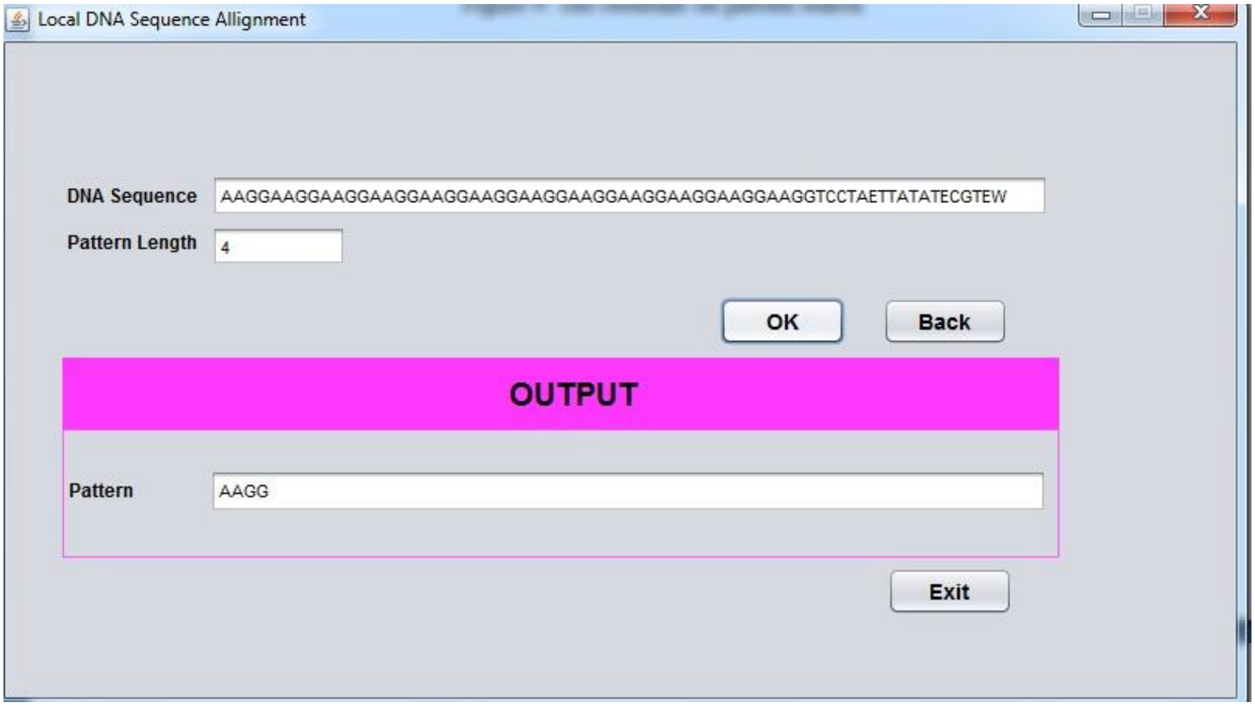


Figure 10: Pattern search according to the length



DNA Sequence AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGTCCTAETTATATECGTEW

Pattern Length 4

Pattern AAGG

Figure 11: The certain patter finding based on length.

Chapter 2: Structural Break

Structural Break is the changes in DNA sequences as any kind of symbols or space among sequences.

AAAA-GCT C-CCCTTTTAT-A-CTGATGCTG-A

Here, there are structural break at four points.

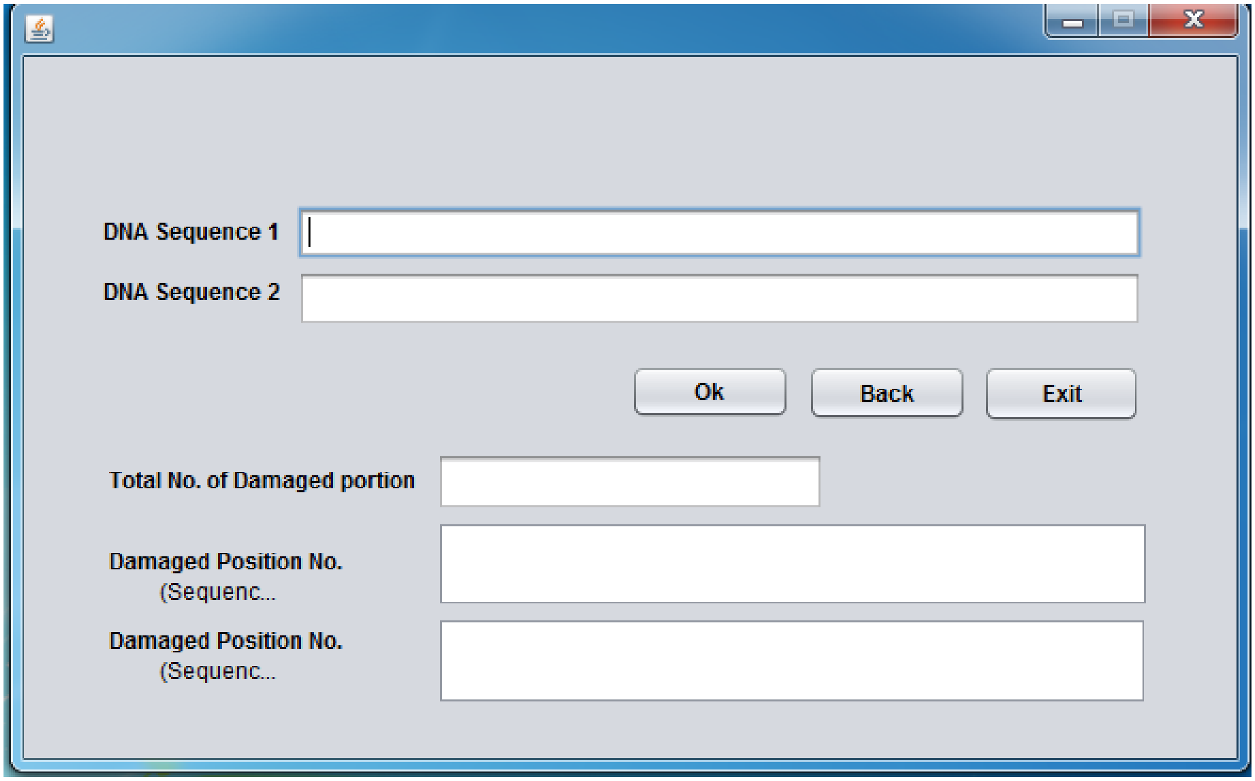
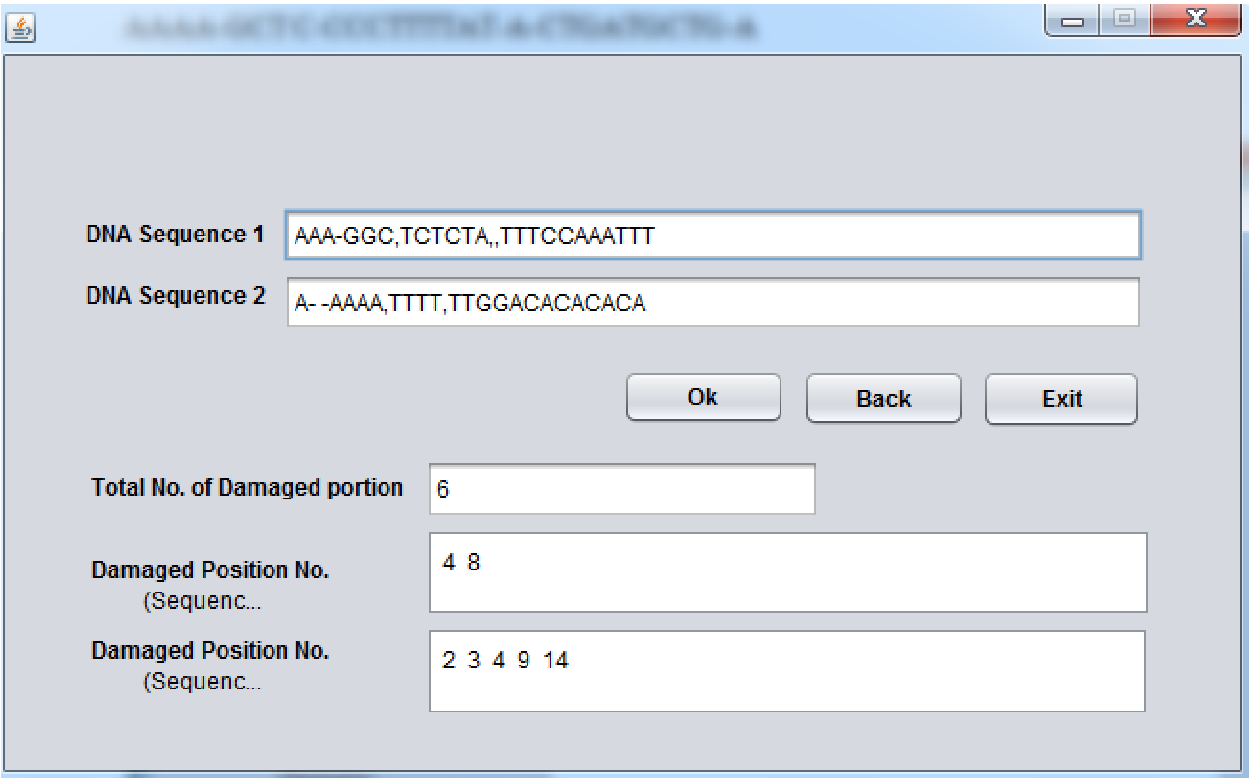

DNA Sequence 1
DNA Sequence 2
Ok
Back
Exit
Total No. of Damaged portion
Damaged Position No. (Sequenc...
Damaged Position No. (Sequenc...

Figure 12: Opening face for Structural Break



DNA Sequence 1 AAA-GGC,TCTCTA,,TTTCCAAATTT

DNA Sequence 2 A- -AAAA,TTTT,TTGGACACACACA

Total No. of Damaged portion 6

Damaged Position No. (Sequenc... 4 8

Damaged Position No. (Sequenc... 2 3 4 9 14

Figure 13: Outcome for Given Data set

It is possible to get input by making browsing data set from computer. This development (Figure 14) helps easy browsing for analysis.

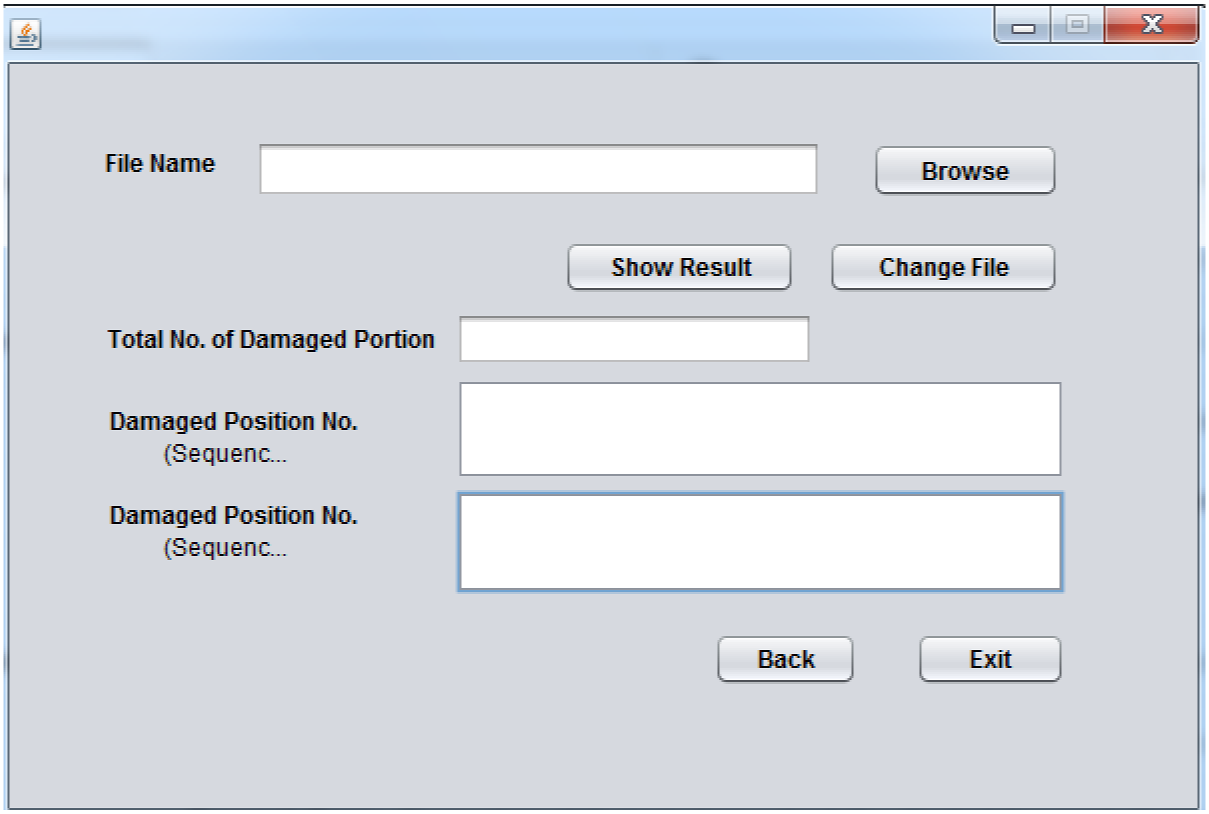
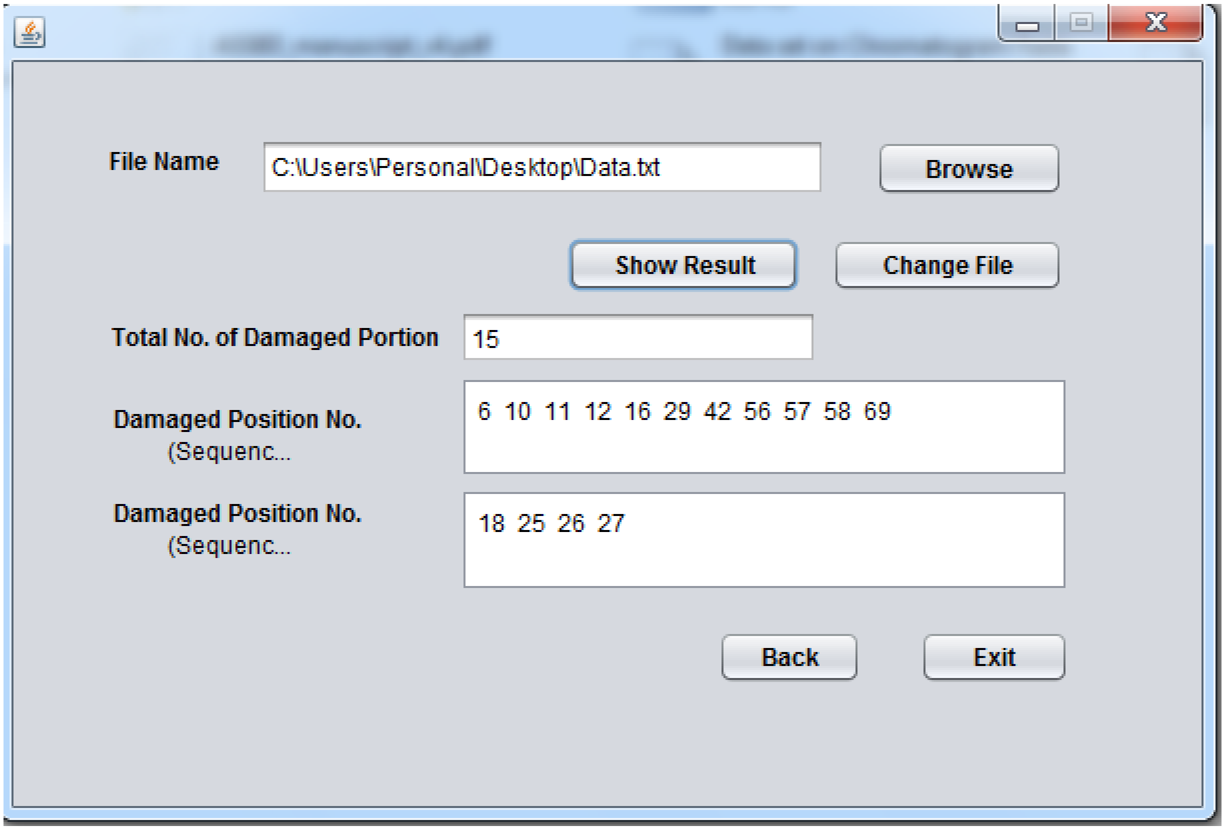

File Name
Browse
Show Result
Change File
Total No. of Damaged Portion
Damaged Position No. (Sequenc...
Damaged Position No. (Sequenc...
Back
Exit

Figure 14: Interface for file input for damage selection and outcome shows (Figure 15)



File Name: C:\Users\Personal\Desktop\Data.txt

Total No. of Damaged Portion: 15

Damaged Position No. (Sequenc...: 6 10 11 12 16 29 42 56 57 58 69

Damaged Position No. (Sequenc...: 18 25 26 27

Figure 15: Outcome for file browsing data set. Here we see that total 15 structural breaks occurred and the positions are identified in both sequences

It is also possible to change the data set from the interface (Figure 16). For this user have to click change file (Figure 15)

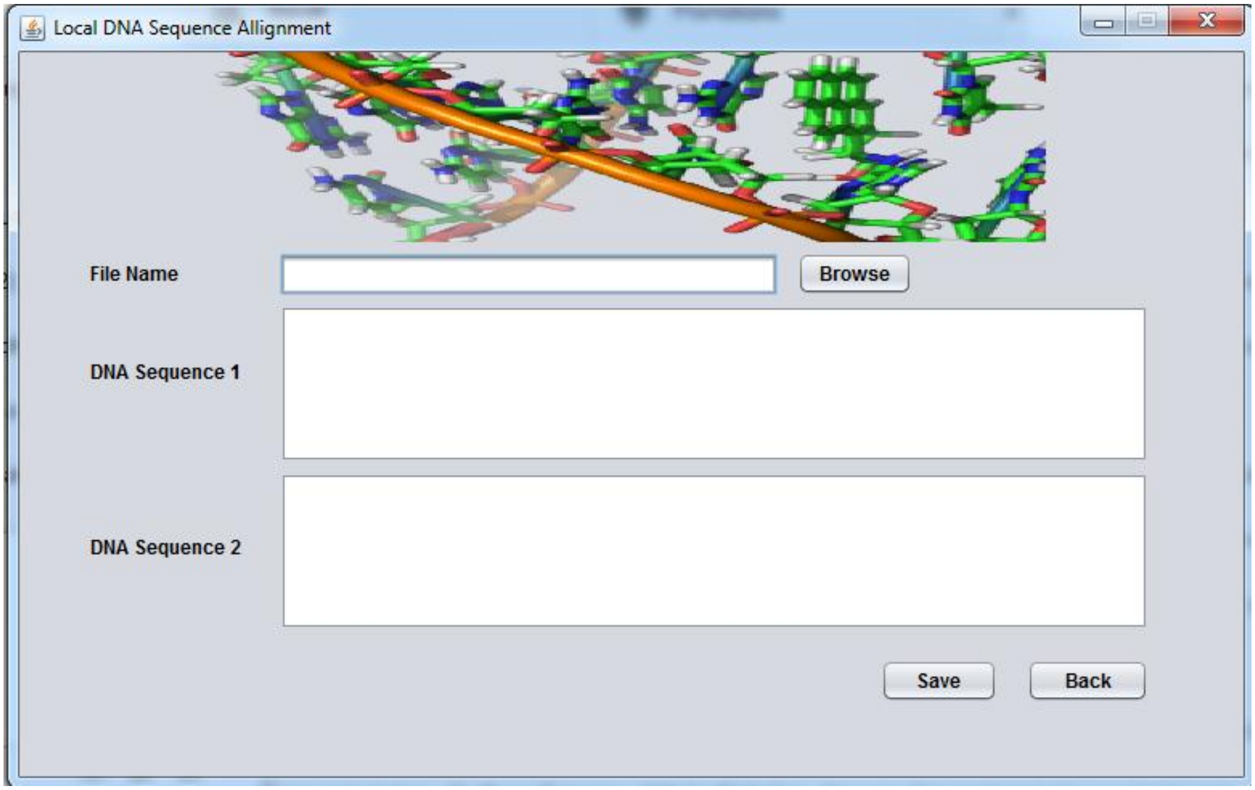


Figure 15: Data set change options.