

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>

Released under GNU public license version 3 (GPL v3).

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**).

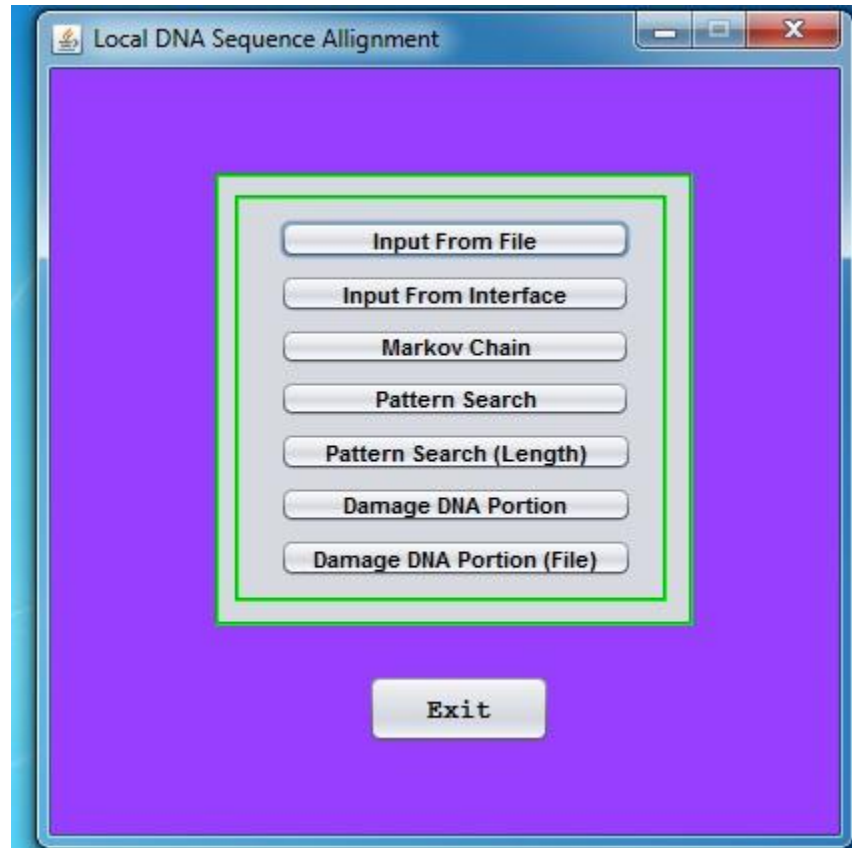


Figure 1: ASSBD Complete Interface

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.

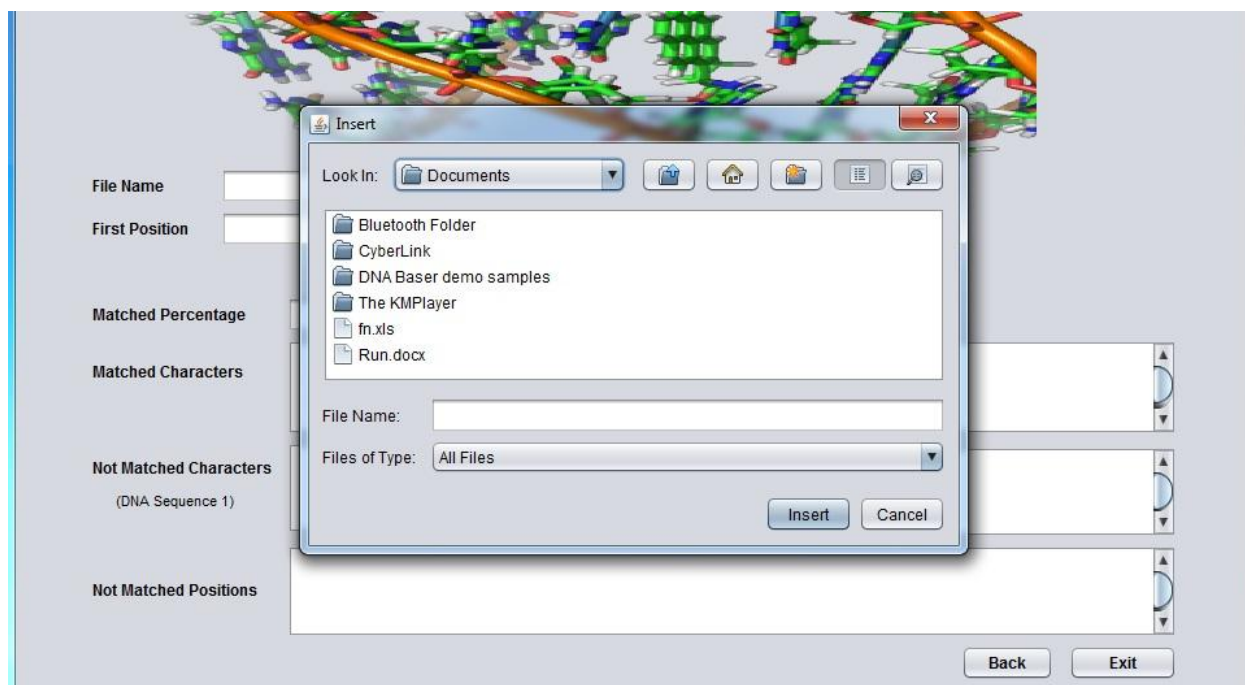



Figure 3: Browse Interface for data file selection

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



File Name

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

Matched Characters

Matched Positions

Not Matched Characters (DNA Sequence 1)

Not Matched Characters (DNA Sequence 2)

Not Matched Positions

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.

The image shows a web application interface with two main sections: 'Input' and 'Output'.

Input Section:

- DNA Sequence1:** A text input field with a vertical scrollbar.
- DNA Sequence2:** A text input field with a vertical scrollbar.
- First Position:** A small text input field.
- Last Position:** A small text input field.
- Buttons:** 'Ok' and 'Refresh' buttons.

Output Section:

- Matched Percentage:** A small text input field.
- Matched Characters:** A text input field with a vertical scrollbar.
- Matched Positions:** A text input field with a vertical scrollbar.
- Not Matched Characters:** A text input field with a vertical scrollbar.
- DNA Sequence 1:** A text input field with a vertical scrollbar.
- DNA Sequence 2:** A text input field with a vertical scrollbar.

Figure 6: Input data set using Keyboard

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

DNA Sequence1	AGCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAAAAAACC	
DNA Sequence2	AGCTTTTTTTTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCCCCCCAAAAAA	
First Position	<input type="text" value="1"/>	Last Position <input type="text" value="34"/>
<div>Ok Refresh</div>		
<h2>Output</h2>		
Matched Percentage	<input type="text" value="76.47059"/>	
Matched Characters	A,G,C,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	T,T,T,T,T,A,A,	
DNA Sequence 1		
Not Matched Characters	C,C,C,C,C,C,C,C,	
DNA Sequence 2		
Not Matched Positions	27,28,29,30,31,32,33,34,	

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 Sequence Allignment

Input

DNA Sequence

Pattern

Output

Percentage of pattern matching

Repitition of Given Pattern

Figure 8: Outline for Pattern Search

The screenshot shows a software window titled "Local DNA Sequence Alignment". It contains two main sections: "Input" and "Output", both highlighted with green headers.

Input Section:

- DNA Sequence:** A text box containing the sequence "AGAAGAAGAAGAAGAGAGAGAAGAAAAAAAAATTTTTTTTCCCCCCCCCCTT".
- Pattern:** A text box containing the pattern "AGA".
- Buttons:** "OK" and "Refresh" buttons are located to the right of the input fields.

Output Section:

- Percentage of pattern matching:** A text box displaying the value "21".
- Repetition of Given Pattern:** A text box displaying the value "6".
- Buttons:** "Back" and "Exit" buttons are located to the right of the output fields.

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**).

The screenshot shows a window titled "Local DNA Sequence Alignment". It contains two input fields: "DNA Sequence" and "Pattern Length", both of which are empty. Below these fields are two buttons: "OK" and "Back". A pink rectangular box labeled "OUTPUT" is positioned below the buttons. Inside this box, there is a "Pattern" label and an empty text field. At the bottom right of the window is an "Exit" button.

Figure 10: Pattern search according to the length

The screenshot shows the same "Local DNA Sequence Alignment" window, but now with data entered. The "DNA Sequence" field contains the text "AAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGTCTTAETTATATECGTEW". The "Pattern Length" field contains the number "4". The "OK" and "Back" buttons are still present. The pink "OUTPUT" box now displays the "Pattern" as "AAGG" in its text field. The "Exit" button remains at the bottom right.

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.

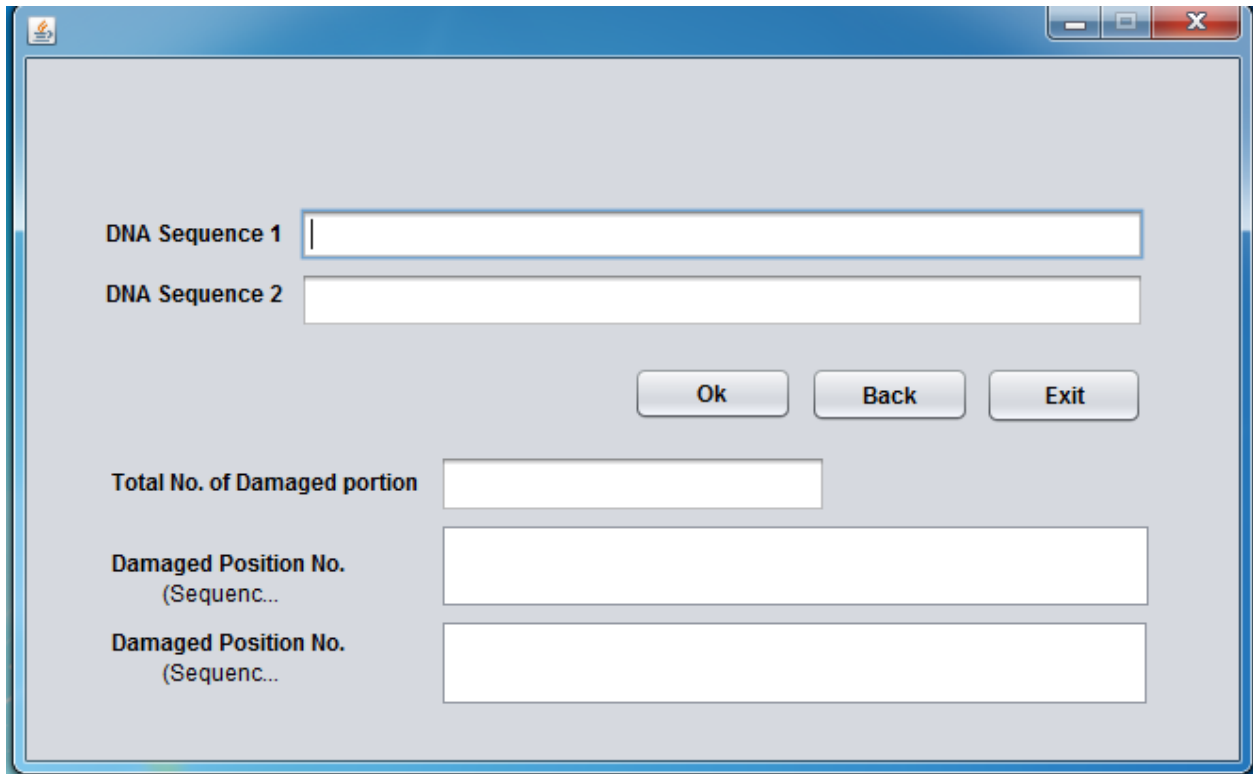


Figure 12: Opening face for Structural Break

The screenshot shows a software window with a title bar containing the text "DNA Sequence Analysis". The window has a light gray background and contains the following elements:

- DNA Sequence 1:** A text input field containing the sequence "AAA-GGC,TCTCTA,,TTTCCAAATTT".
- DNA Sequence 2:** A text input field containing the sequence "A- -AAAA,TTTT,TTGGACACACACA".
- Buttons:** Three buttons labeled "Ok", "Back", and "Exit" are positioned below the DNA sequence input fields.
- Total No. of Damaged portion:** A text input field containing the value "6".
- Damaged Position No. (Sequenc...):** A text input field containing the values "4 8".
- Damaged Position No. (Sequenc...):** A text input field containing the values "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

Total No. of Damaged Portion

Damaged Position No.
(Sequenc...

Damaged Position No.
(Sequenc...

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

The screenshot shows a software window with a light blue border and a standard Windows-style title bar. Inside the window, the background is a light gray. At the top left, there is a small icon of a flame. The main content area contains the following elements:

- File Name:** A text box containing the path "C:\Users\Personal\Desktop\Data.txt". To its right is a "Browse" button.
- Show Result:** A button located below the File Name field.
- Change File:** A button located to the right of the Show Result button.
- Total No. of Damaged Portion:** A text box containing the number "15".
- Damaged Position No. (Sequenc...):** A text box containing the sequence "6 10 11 12 16 29 42 56 57 58 69".
- Damaged Position No. (Sequenc...):** A text box containing the sequence "18 25 26 27".
- Back:** A button located at the bottom right of the interface.
- Exit:** A button located to the right of the Back button.

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)

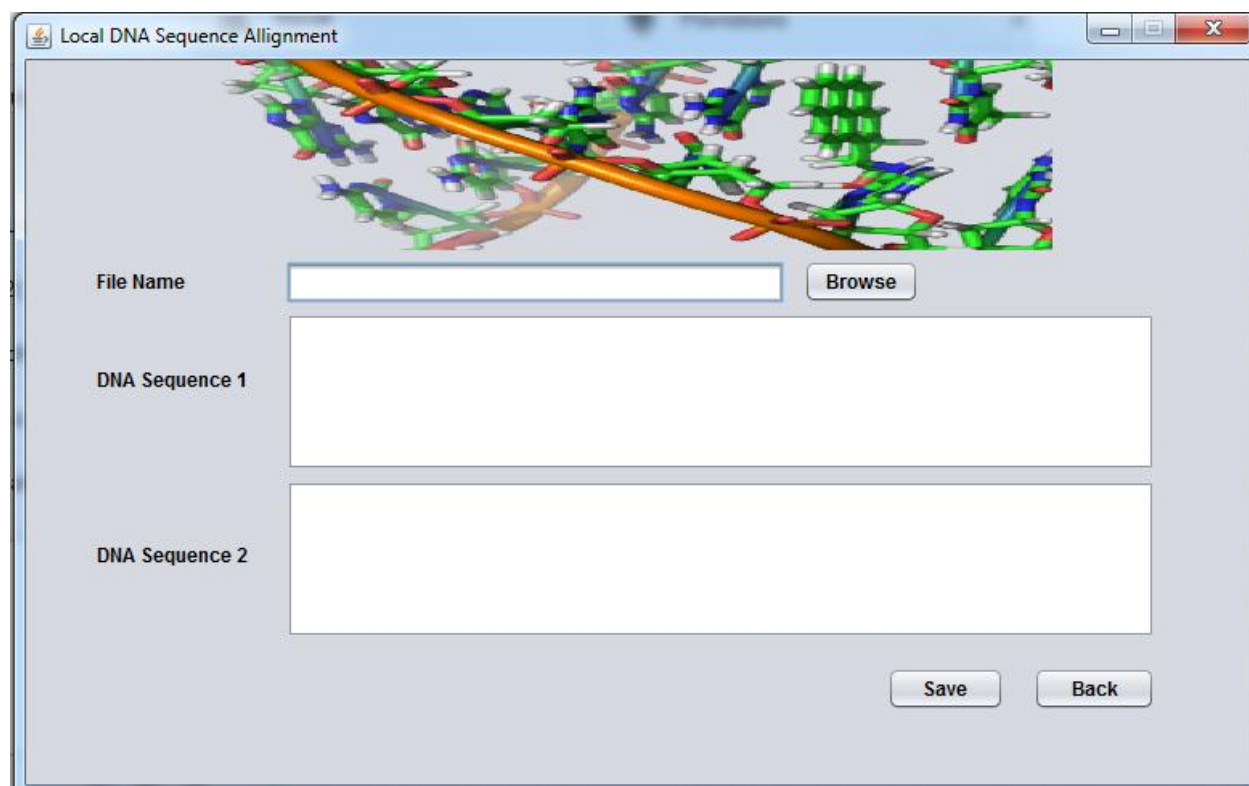


Figure 15: Data set change options.