Project in Bioinformatics (236524)

Sanger Aligner

Name: Eden Nagar

ID: 312589815

E-mail: edenagar@campus.technion.ac.il

Supervisors: Yahav Festinger, prof. Ayelet Lamm

Biology faculty, Technion

November 2022

Table of Contents

Abstract

The Sanger Analyzer is a native application designed to help researchers quickly process and analyze Sanger sequencing data. Sanger sequencing is a widely used method for small-scale DNA strand sequencing, but aligning the resulting sequences with a reference can be time-consuming and challenging, especially when done manually.

The Sanger Analyzer was developed to address this need, providing an easy-to-use tool for aligning a user's Sanger sequence with a reference sequence. It can process AB1 and user text files, generate relevant data such as chromatogram values and samples, and use an aligner algorithm based on the Needleman-Wunch method to find the best alignment. The app also includes an interpreter algorithm that translates AB1 files into nucleotide sequences, and a framework library that serves as a communication link between the user interface and the alignment algorithm.

Overall, the Sanger Analyzer is a valuable resource for researchers working with Sanger sequencing data, helping them quickly and accurately align and analyze their sequences without the need for tedious, repetitive manual alignment.

Tool Capabilities

Our app is designed to help users easily convert binary Sanger files to the actual DNA sequence and perform DNA alignment. Here is a closer look at the specific capabilities of our tool:

- Convert binary Sanger files to DNA sequence: Our app makes it easy to convert Sanger files to the actual DNA sequence, allowing users to easily view and analyze the data contained in the file. In addition, the app provides data regarding each nucleotide, including its location in the sequence and height that means record dye fluorescence, which can be very useful for understanding the characteristics of the DNA sequence and identifying any patterns or trends that may be present.

- DNA alignment: Our app has the ability to perform DNA alignment, finding the best match between two DNA sequences. This is a valuable tool for comparing and analyzing different DNA sequences, and the app's ability to provide insight into the alignment, including the closest nucleotide that isn't showing but has a certain height, is likely to be very useful for understanding the details of the alignment and identifying any potential issues or discrepancies.

We believe that these capabilities make our app a valuable tool for anyone working with DNA sequences and alignment data, and we hope you find it to be a useful resource.

New Version Updates

The new version updates of the aligner app introduce a number of significant changes compared to the old version. One of the most significant changes is the way that the app processes and analyzes the data from the Sanger machine.

In the old version of the app, the data from the Sanger machine was output in the form of a PDF file, which was then analyzed using image processing techniques. This approach had a number of limitations, including the potential for errors to be introduced during the image processing step. In contrast, the new version of the app translates the binary data that is output by the Sanger machine and represents the nucleotide locations as numerical derivatives zero and local maxima. This allows the app to analyze the data more accurately and efficiently, without the need for image processing, this can be useful for helping researchers to understand the data and identify patterns or trends in the data.

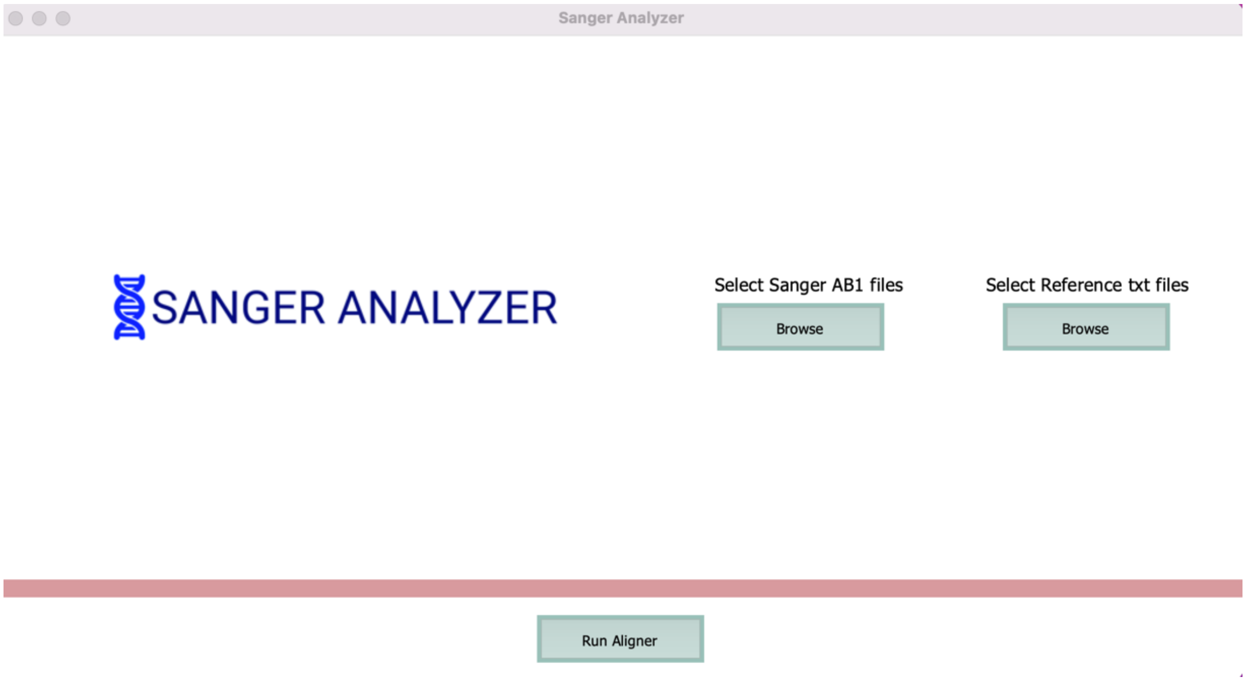
The old version of our app was a useful tool for converting Sanger PDF files to the actual DNA sequence, but it didn't have the capability to perform DNA alignment. This limited its usefulness for users who needed to compare and analyze different DNA sequences.

The new version of our app addresses this limitation by introducing a powerful DNA alignment algorithm. This allows users to easily find the best match between two DNA sequences, and provides insight into the alignment, including the closest nucleotide that isn't showing but has a certain height. This additional capability makes our app a much more powerful tool for anyone working with DNA sequences and alignment data.

One other notable improvement in the new version of our app is the user interface (UI). We have redesigned the UI to make it more intuitive and user-friendly, making it easier for users to navigate and use the app effectively.

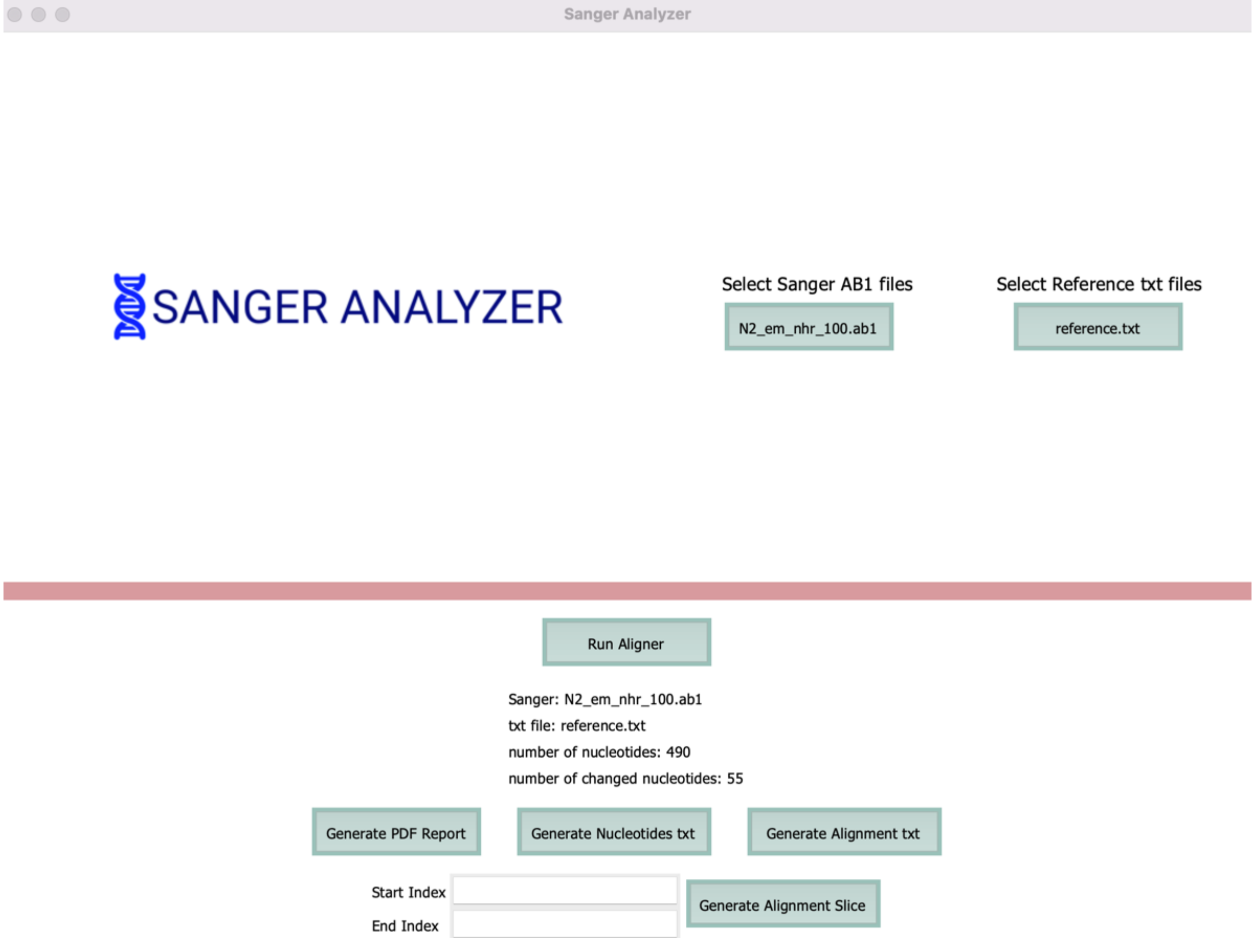
Overall, the new version of our app is a significant improvement over the old version, and we believe it will be a valuable resource for anyone working with DNA sequences and alignment data.

The User Interface



A Simple user interface (UI) was developed. The user needs to select an AB1 file and a text reference file and Run the Aligner. This can be done by clicking “Browse”, and then “Run Aligner”.

Then, the bottom section will appear, reviling all the data features:



Sanger: N2_em_nhr_100.ab1
txt file: reference.txt
number of nucleotides: 490
number of changed nucleotides: 55

Now, instead of "Browse" appears the name of the chosen file.

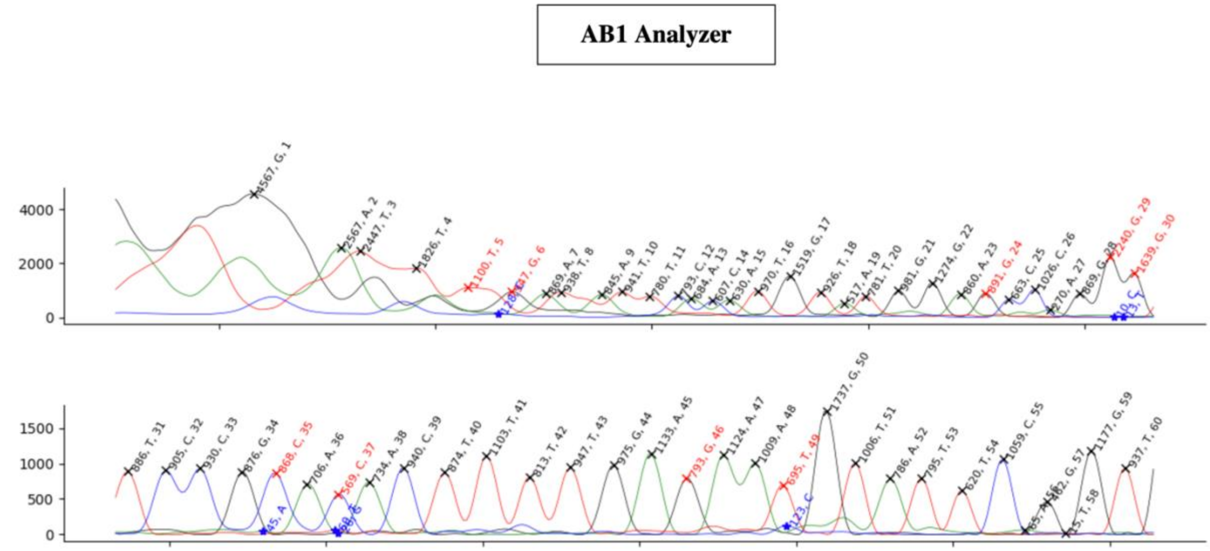
At this window, the user can choose the following options:

1. Generate PDF Report:

Generates a full report of the chromatogram, a relevant sanger sequence, a nucleotide type, the location of every nucleotide with its sequence Index (for instance: the sequence “ATAGCC” the index of nucleotide G is 4), and the value of the chromatogram.

Overall, a PDF and image of nucleotide plot generator is a useful tool for lab researchers who are working with nucleotide data and want to gain a deeper understanding of the data through visual representation.

Another feature is the color of the nucleotides, that hints information about the alignment:



Black- match, refers to the optimum alignment permutation in the alignment where the nucleotide in the reference matches the nucleotide in the other sanger sequence.

Red- mismatch, refers to the optimum alignment permutation in the alignment where the nucleotide in the reference don’t match the nucleotide in the other sanger sequence.

Blue- a peak that is located under a mismatch.

The Y axes represent the volume dye fluorescence of the given dideoxynucleotides.

2. Generate Nucleotides txt:

The nucleotide data text file generator can then output information in a text file format that can be easily imported and analyzed by other tools or software. This can be especially useful for the lab data engineers who are working with large datasets and need a fast and efficient way to extract and process the data nucleotides originated from the AB1 file:

- Location: location of the nucleotide relative to the sample time by the Sanger machine.
- Height: chromatogram mass.
- left_ips: start approximation of the amplitude.
- right_ips: end approximation of the amplitude.
- index: nucleotide index.
- is_changed: if this nucleotide was change after alignment.
- under_peaks: all the peaks that are beneath the current amplitude and is not part of the sequence.
- amplitude_ratio: the ratio between the peak’s height and the highest peak in the sequence.

For example:

```
location=746, height=695.0, left_ips=741.8306451612904,
right_ips=749.4961832061068, nucleotide='T', index=49), is_changed=True,
under_peaks=(746, [Peak(location=747, height=123.0, left_ips=743.1458333333334,
right_ips=751.0263157894736, nucleotide='C', index=None)]),
amplitude_ratio=0.13568918391253418
```

Overall, a nucleotide data text file generator is a valuable tool for the lab data engineers working in the lab, as it allows them to quickly and easily extract and process large amounts of data from DNA sequencing files.

3. Generate Alignment text:

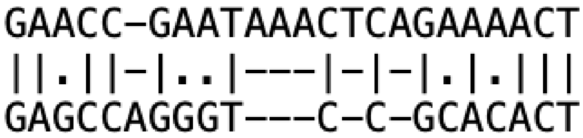
The Generate Alignment text is a tool that allows researchers to generate a text representation of the optimum alignment between the reference and sanger sequences. This can be useful for helping researchers to understand and analyze the alignment in more detail, as it provides a clear and concise summary of the alignment that can be easily interpreted.

The first line shows the reference sequence.

The last line shows the AB1 sequence.

The middle line represents the symbols line when "-" stands for a gap, "|" for a match, and "." presents a mismatch.

*Every row contains 100 nucleoides of the alignment.

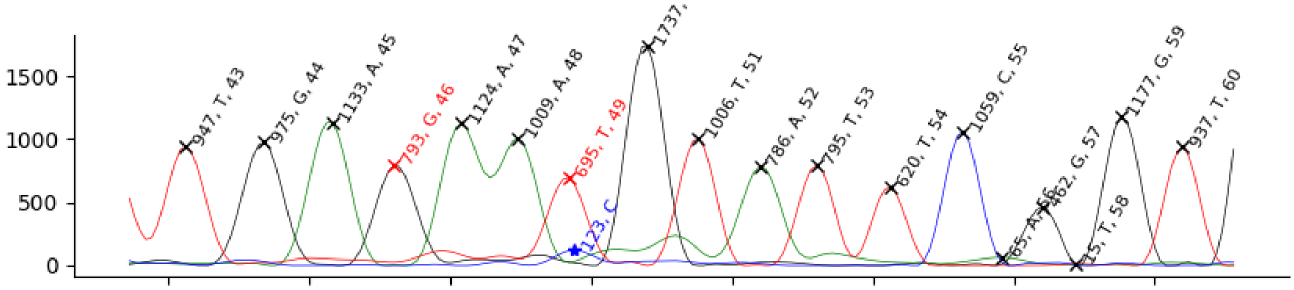


```
GAACC-GAATAAACTCAGAAAACT
||.||-|..|---|-|-|.|.|||
GAGCCAGGGT---C-C-GCACACT
```

4. Generate Alignment Slice:

The user needs to fill a specific nucleotide's start and end indices. Given these indices, the application generates a cropped image of the desired part of the sequence.

This can be useful for helping researchers to understand and analyze a specific region of the alignment in more detail, as it provides a visual representation of the nucleotides in that region.



This allows researchers to easily share the alignment slice with others or include it in reports or publications.

Discussion & Suggestions for Further Development

Although the Sanger Aligner application is enough to work with the Sanger machine and align to a given reference, these are several suggestions for future work in this field:

1. The user interface can be served on a web-based user interface, using technologies such as Django or Flask (no python environment and run). In this way, it can be accessed by any computer and include history data.
2. To add more options of file types such as generating plots of JPEG or SVG.
3. Expand the plotting configuration possibilities, to change the width of the removed text and to change colors.
4. When an AB1 file is given, the application can check on Blast's close relative sequence in the European database and add it as an extra information for the user.
5. The algorithm can be best suited for the task by distributing the weight value, of the match, mismatch and gap values by a normal distribution or a Bayesian distribution.
6. The algorithm can be improved by classifying noisy amplitudes using advance tools such as Machine learning.

Appendix

Requirements:

The framework needs python 3.6 and above.

To start the App:

On Linux/Mac:

1. Open the command line in the project directory
2. Run “./install_linux.sh”.
3. Now from here on to open the app just run “./run_app.sh”.

On windows:

1. Double click “install_windows.bat”.
2. Double click “run_app.sh”.
3. If there is an issue just open the command line and run:
    a. “call install_windows.bat” to install.
    b. “call run_app.bat” every time you want to run the app.

“Aligner” Framework Library

In this project, a framework library named “Aligner” was developed. The Aligner framework is the backbone of the application; it contains all the methods that are needed for the user interface and the algorithms. The framework works like any other standard class, which needs initializing with a path to an AB1 file. The user is grunted with the following methods:

- Align_with – gets a string of the sequence and performs the alignment.
- Number_of_nucleotides – return the number of nucleotides that the AB1 sequence contains.
- Number_of_changed_nucleotides -return the number of the mismatched nucleotides.
- Print_best_alignment – prints the best alignment to the console.
- best_alignment – return the best alignment in the formatted version.
- Plot_matter_sequence_by_indexes – gets left_index, right_index, file_path, name and returns a PNG file, at the file_path with the name that was given from left to right indecis.
- Generate_pdf – in case of given file_path, it generates a PDF file with a plot of the whole sequence.
- Generate_nucleotides_info_file – gets a file_path and a string name, and generates a text file at the file path. It contains all the AB1 nucleotides information.
- Generate_alignment_file – gets file_path and a string name, and generates text file with the alignment.

In addition, the library includes several functions that are already implemented for future use:

- Plot_heat_map – gets start and finish sequence indices, and shows a heat map of the close-range changes.
- Is_noisy – gets amplitude_ratio_threshold and noise_percent_threshold. it returns a Boolean- if the ratio between the noisy elements and the size of the sequence is greater than noise_percent_threshold. Noisy elements are nucleotides that the amplitude_ratio is grater then amplitude_ratio_threshold.