Bio.Net Bug Bash Help Document

# Introduction

This document is meant to serve as an introduction to Bio.Net Demo application from Microsoft Research. This will likely be helpful for programmers & testers who are wishing to test/use the demo application.

# Sequence Assembler

Bio.NET is a library of reusable bioinformatics algorithms and functions using the .NET platform. Sequence Assembler is built to demonstrate the power of the Bio.NET libraries. This is a primary executable demo application which hosts the application framework, the WPF UI components as well as hosts the graphical visualization components.

# What is a Sequence?

A sequence is a succession of letters representing the [primary structure](http://en.wikipedia.org/wiki/Primary_structure) of a real or hypothetical [molecule](http://en.wikipedia.org/wiki/Molecule) or strand, with the capacity to carry [information](http://en.wikipedia.org/wiki/Information) as described by the [central dogma of molecular biology](http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology).

The possible letters for Dna Sequence is A, C, G, and T

# What is Sequence Alignment?

Sequence alignment is a way of arranging the sequences of [DNA](http://en.wikipedia.org/wiki/DNA), [RNA](http://en.wikipedia.org/wiki/RNA), or [protein](http://en.wikipedia.org/wiki/Protein) to identify regions of similarity that may be a consequence of functional, [structural](http://en.wikipedia.org/wiki/Structural_biology), or [evolutionary](http://en.wikipedia.org/wiki/Evolution) relationships between the sequences.

Alignment can be done with the help of different algorithms below:

* [Needleman-Wunsch algorithm](http://en.wikipedia.org/wiki/Needleman-Wunsch_algorithm) – Global Alignment
* [Smith-Waterman algorithm](http://en.wikipedia.org/wiki/Smith-Waterman_algorithm) – Local Alignment
* Pair wise Overlap algorithm

# File Formats

There are three different file formats which can be passed to Sequence Assembler:

### FastA Format

FastA format (a.k.a. Pearson format) is a text-based [format](http://en.wikipedia.org/wiki/File_format) for representing either [nucleotide sequences](http://en.wikipedia.org/wiki/Nucleotide_sequence) or [peptide sequences](http://en.wikipedia.org/wiki/Peptide_sequence), in which [base pairs](http://en.wikipedia.org/wiki/Base_pair) or [amino acids](http://en.wikipedia.org/wiki/Amino_acid) are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences.

### 

### GenBank Format

The GenBank [Format](http://en.wikipedia.org/wiki/Sequence_database) is also a text-based for representing either [nucleotide sequences](http://en.wikipedia.org/wiki/Nucleotide_sequence) or [peptide sequences](http://en.wikipedia.org/wiki/Peptide_sequence).



### GFF Format

The general feature format (gene-finding format, generic feature format, GFF) is a [file format](http://en.wikipedia.org/wiki/File_format) used for describing [genes](http://en.wikipedia.org/wiki/Gene) and other features of [DNA](http://en.wikipedia.org/wiki/DNA), [RNA](http://en.wikipedia.org/wiki/RNA) and [protein](http://en.wikipedia.org/wiki/Protein) sequences.



# Web Services

### BLAST

BLAST, is an [algorithm](http://en.wikipedia.org/wiki/Algorithm) for comparing [primary](http://en.wikipedia.org/wiki/Primary_structure) biological sequence information, such as the [amino-acid](http://en.wikipedia.org/wiki/Amino_acid) sequences of different [proteins](http://en.wikipedia.org/wiki/Protein) or the [nucleotides](http://en.wikipedia.org/wiki/Nucleotide) of [DNA sequences](http://en.wikipedia.org/wiki/DNA_sequence). A BLAST search enables a researcher to compare a query sequence with a library or [database](http://en.wikipedia.org/wiki/Database) of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

### SilverMap

SilverMap is a visual representation of BLAST output. It basically shows how similar a particular gene is to a given set of genes with all genes coming as BLAST output.

# Flow of the Application

* Open the SequenceAssembler.exe by double clicking

### Load Sequences

The below steps would load the sequences into Sequence assembler application.

* Click on Menu ‘Load’ -> ‘Open’
* Click on ‘Browse’ and select any two valid files in the formats (FastA, GenBank & GFF).
* Select the Molecule Type i.e., DNA and then click on ‘Import’ Button.

### Validate Loaded Sequences

The loaded Sequences can be validated following the steps below.

* Expand the ‘Sequence View’ Tab (Bottom tab) and on the left hand side one can find the sequence files with file names imported.
* One can place the mouse cursor over the sequence and can find the details about the sequence i.e., the length, ID, Molecule Type and File path.

### Unload Loaded Sequences

* To unload a particular sequence, right click on a particular sequence and select ‘Unload’.

### Assemble Loaded Sequences

To Assemble sequences loaded follow the below steps.

* Double click/Drag and drop the sequences from the left hand side of ‘Sequence View’ to the Centre pane.
* The imported sequences would now be visible under the ‘Selected Sequences’.
* In the ‘Select Algorithm’ drop down in the bottom of the window select the algorithm (Needleman-Wunsch, Smith Waterman & Pairwise Overlap) which is to be used for alignment and assemble.
* Once the Algorithm is selected click on ‘Assemble’ button next to drop down.

### Validate Assembled Sequences

To Validate the assembled sequences follow the below steps.

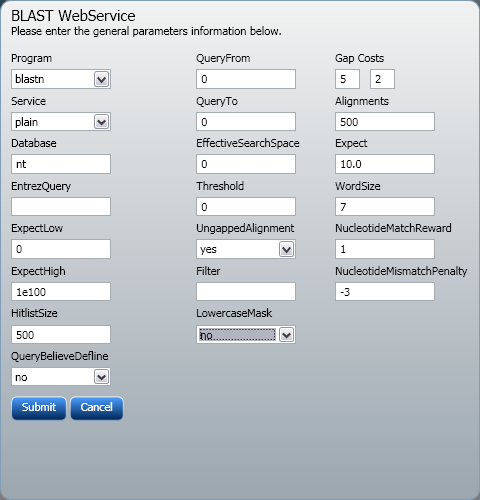
* Once the ‘Assemble’ button is clicked, the application would take to ‘Consensus view’ Tab (Top tab) and should provide the information about the assembled sequences i.e., time taken for run, number of contigs found, total length, number of assembled sequences, etc.,
* The summary would be provided in the tab called ‘Alignment Report’ and the graphical view of the assembled sequence would be visible in the Custom alignment view if there are any valid contigs, we can change the color scheme for the same.
* One can place the mouse cursor over the Contig sequence on the left pane and can find the details about the sequence i.e., the length, ID, Molecule Type and Contig sequence.

### Save/Edit the Contigs

* On the left hand side of ‘Consensus View’, the Contigs for the sequences should be visible which on right click would provide an option of ‘Save’ and ‘Edit’.

### Validate Blast Web service

* To validate the Blast result for the contig received can be done by selecting web service in the drop down “Select Web service” (Ncbi and NCBI QBlast) in the bottom of the window.
* Once the Web service type is selected, click on the “Execute Search” button on the left of the drop down.
* Now Blast web service parameter window would pop-up, update the values as per the screen shot below and click on ‘Submit’ button, this would return the Blast results.



**Screen Shot for values to be update when selected ‘Ncbi’ Option**

****

**Screen Shot for values to be update when selected ‘NCBI QBLAST’ Option**

* The Blast result would be provided in the ‘Web Service’ tab which would be having two results one is Single line report and the other is Silver Map.
* On double clicking on any item in the Single line report, would provide the Pair wise Alignment.