|  |  |
| --- | --- |
|  | Microsoft Research Sequence Assembler: User’s Guide  Version 1.0 - June 2010 |

Abstract

The Microsoft Research Sequence Assembler is a proof-of-concept application that demonstrates how the Microsoft® Biology Foundation (MBF) can be used to create applications for bioinformatics research. The MBF Sequence Assembler uses rich user interface (UI) elements to enable the manipulation and visualization of genomic data.

The MBF Sequence Assembler implements several features of the Microsoft Biology Foundation:

* A set of parsers for common genome file formats.
* A set of algorithms for alignment and/or assembly of DNA, RNA, or Protein strands.
* A set of connectors to several Basic Local Alignment Search Tool (BLAST) Web services for genome identification.

Reports from BLAST services can be viewed as single-line reports or in the SilverMap visualization component integrated by the Queensland University of Technology. The MBF Sequence Assembler is available at <http://mbf.codeplex.com>.

Contents

[Introduction 3](#_Toc264984093)

[How to Install the MBF Sequence Assembler 3](#_Toc264984094)

[UI Overview 4](#_Toc264984095)

[How to Import Sequence Data 5](#_Toc264984096)

[How to Align Sequences 6](#_Toc264984097)

[How to Send Consensus Views to BLAST Services 9](#_Toc264984098)

[How to Configure the MBF Sequence Assembler 11](#_Toc264984099)

[Appendix A: Supported Sequence and File Formats 13](#_Toc264984100)

Disclaimer: This document is provided “as-is”. Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it.

This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.

© 2010 Microsoft Corporation. All rights reserved.

Microsoft, Silverlight, Visual Studio, and Windows are trademarks of the Microsoft group of companies.  All other trademarks are property of their respective owners.

# Introduction

The Sequence Assembler application is intended for use by a biologist or laboratory technician who is responsible for managing next generation genomic sequencing data for alignment, assembly and/or BLAST identification. Though many other full-featured applications provide similar functionality, the primary goal of providing this application is to expose the capabilities available in the Microsoft Biology Foundation.

When coupled with the rich user interface (UI) elements that the Windows Presentation Foundation supports, the Sequence Assembler provides a unique combination of advanced application interface, controls, and visualization of the data. These features can be useful to scientists, researchers and clinicians who work with genomic data, but more importantly the Sequence Assembler is a full-featured sample application that can be extended, modified, and expanded to meet the needs of a wide range of researchers.

With the MBF Sequence Assembler, you can:

* Import files containing sequence data. Supported file formats:

FASTA GenBank  
FASTQ GFF

* Align entire or partial sequences. Supported algorithms:

MUMmer 3.0 Pairwise-Overlap (Reference Implementation)  
Needleman-Wunsch PAMSAM  
NUCmer 3.0 Smith-Waterman

* Assemble a consensus view from aligned sequences.
* Send consensus views to biological Web services for identification.  
  Supported biological Web services:

NCBI QBLAST  
EBI WU-BLAST  
AzureBLAST

* View results of BLAST queries as single-line reports or in the SilverMap visualizer.

For more information, see these topics in the Microsoft Biology Foundation documents folder:

Microsoft Biology Foundation: An Overview [MBF\_Overview.docx]  
Microsoft Biology Foundation Programming Guide [MBF\_Programming\_Guide.docx]

See also Appendix A, “Supported Sequence and File Formats.”

# How to Install the MBF Sequence Assembler

This section describes the prerequisites, system requirements, and installation steps for the MBF Sequence Assembler.

#### Prerequisites

To use the basic capabilities of MBF, you should have a basic understanding of methods and nomenclature of genomics and bioinformatics.

To take advantage of programming and extensibility capabilities, you need at least:

* Basic programming skills.
* Familiarity with the use of Microsoft Visual Studio® to program .NET Framework applications with C#.
* Basic understanding of programming for Web services.

#### System Requirements

* Windows® XP Service Pack (SP) 3 or later versions of Windows
* Microsoft .NET Framework Version 4.0, available at <http://www.microsoft.com/downloads/details.aspx?FamilyID=9cfb2d51-5ff4-4491-b0e5-b386f32c0992>.

#### Installation

The installer for the MBF Sequence Assembler is SequenceAssembler.msi, available at http://mbf.codeplex.com.

To install the MBF Sequence Assembler

1. Copy the MBF Sequence Assembler installer to a folder on your hard drive.

2. Navigate to that folder and double-click SequenceAssembler.msi, which runs the installation wizard.

3. Follow the directions in the installation wizard to install the MBF Sequence Assembler.

The installer copies the files to a directory named C:\Program Files\Microsoft Biology Initiative\1.0\MBT\Sequence Assembler.

After installation, the MBF Sequence Assembler appears in the Add and Remove Programs Control Panel applet as “Microsoft Research Sequence Assembler.”

# UI Overview

The MBF Sequence Assembler has three expandable panes and a menu bar, as shown in Figure 1:

* The menu bar contains the following menus:
* File: Commands to open and save sequence data files.
* Options: Commands to change sequence colors and associate file types.
* Help: Not functional in this release.
* Sequences pane: Used to assemble and align two or more sequences.
* Contigs pane: Displays the contigs of the assembled sequences.
* Blast pane: Displays the results of BLAST queries either as a single-line report or in the SilverMap visualizer.

  
Figure 1. The MBF Sequence Assembler

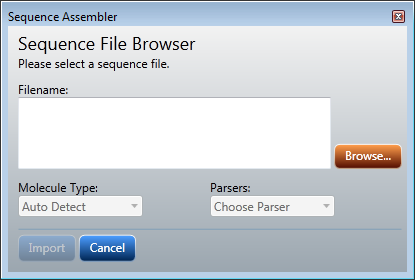
# How to Import Sequence Data

The MBF Sequence Assembler supports the following types and formats of genomic data:

* Sequences of DNA, RNA, or protein: FASTA, FASTQ, and GenBank formats
* Sequence metadata: GFF format

To import sequence data

1. Navigate to C:\Program Files\Microsoft Biology Initiative\1.0\MBT\Sequence Assembler and run SequenceAssembler.exe.
2. Click File on the menu bar, and click Open.
3. Click Browse in the Sequence File Browser, as shown in the following figure.

  
The Sequence File Browser window

1. Browse to a sequence file in the selected format and select it.
2. Select the sequence format from the Parsers drop-down menu, and click Import.

Sequence options are FASTA, FASTQ, Genbank, and GFF. The following figure shows the results of importing two FASTA sequences.

  
The Sequence Tree View

To unload a sequence

* Right-click the sequence in the Sequence Tree View, and click Unload from the pop-up menu.

# How to Align Sequences

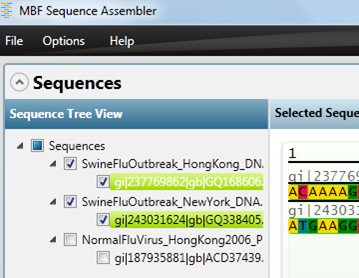
DNA, RNA, and protein sequences can be aligned using the following algorithms:

MUMmer 3.0  
Needleman-Munsch  
NUCmer 3.0  
PAMSAM  
Pairwise-Overlap (Reference Implementation)  
Smith-Waterman

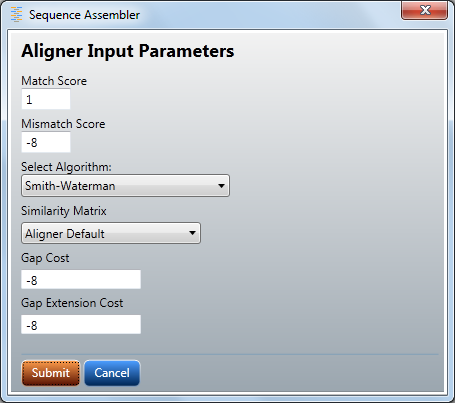
After choosing an algorithm and two or more sequences, you set the parameters for the alignment, which includes selecting a similarity matrix. The resulting alignment is displayed in the Contigs pane of MBF Sequence Assembler.

To align sequences

1. Import two or more sequences of the same type, as described in “How to Import Sequence Data” earlier in this document. The sequences are selected by default when you import them.
2. In the Sequences pane, clear the checkbox beside any sequence you want to omit from the alignment, as shown in the following figure.

  
Unselecting a sequence

1. In the **Sequences** pane, click Align and set the Aligner Input Parameters as shown in the following figure.

   
Aligner Input Parameters

1. Click Submit. The results are displayed in the Alignment pane as shown in the following figure.

  
The aligned sequences

To assemble sequences

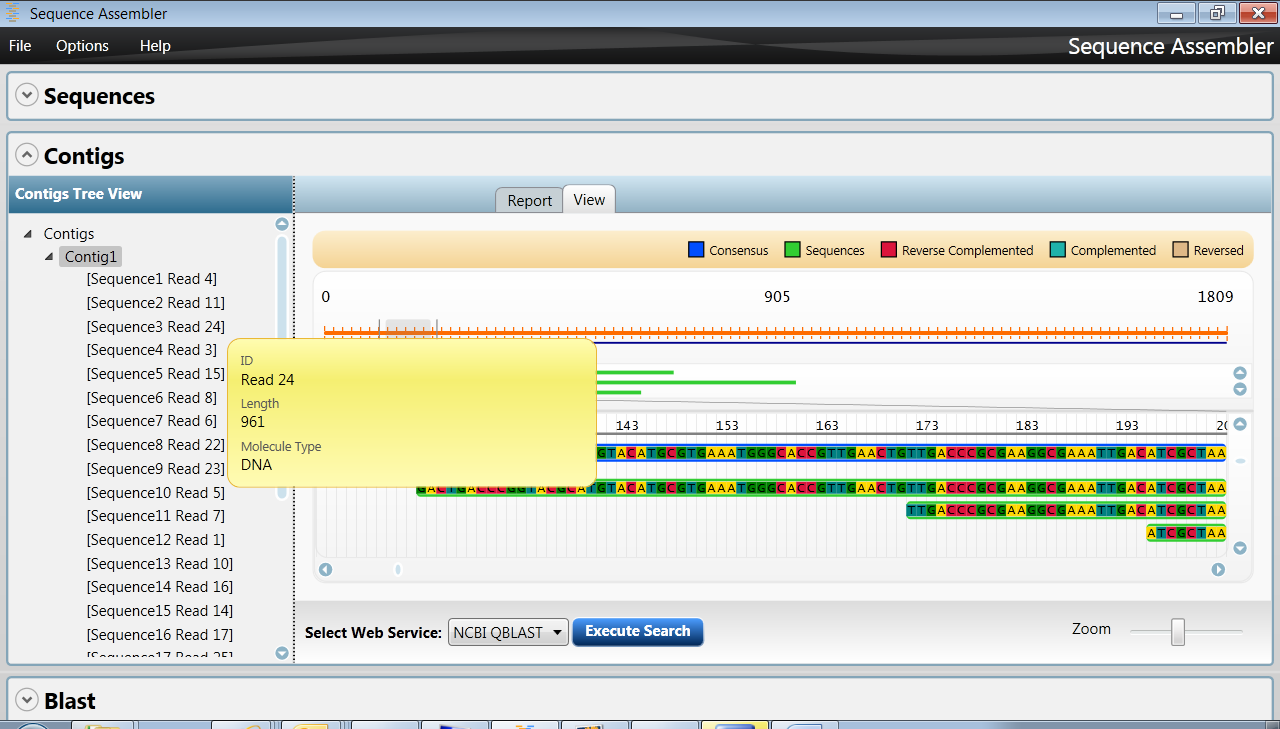
1. Align two or more sequences as described in “To Align Sequences” earlier.

2. In the Sequences pane, click Assemble.

The result is displayed in the Consensus pane, as shown in the following figure.

  
The Consensus pane

3. If a contig exists, it is displayed in the Consensus pane. Hover over a sequence in the contig to see its data in a pop-up, as shown in the following figure.

  
The pop-up for a contig sequence

4. To save the contig, right-click it in the Consensus pane and click Save.

# How to Send Consensus Views to BLAST Services

You can use the MBF Sequence Assembler to send consensus views of aligned sequences to the following biological Web services for validation:

EBI WU-BLAST  
NCBI QBLAST  
AzureBLAST

Note: AzureBLAST is disabled by default. See readme.txt for instructions on how to enable it.

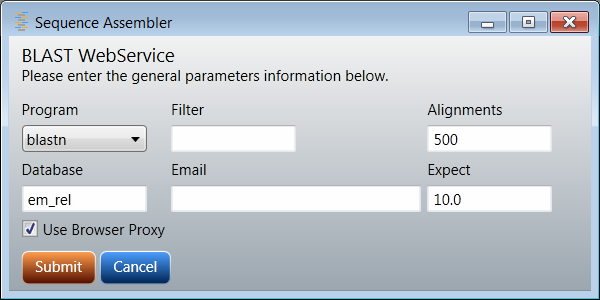
After selecting a sequence and choosing a service, you set the parameters for the query, which vary for each service. The results are displayed in the Web Service tab.

To send a consensus view to a BLAST service

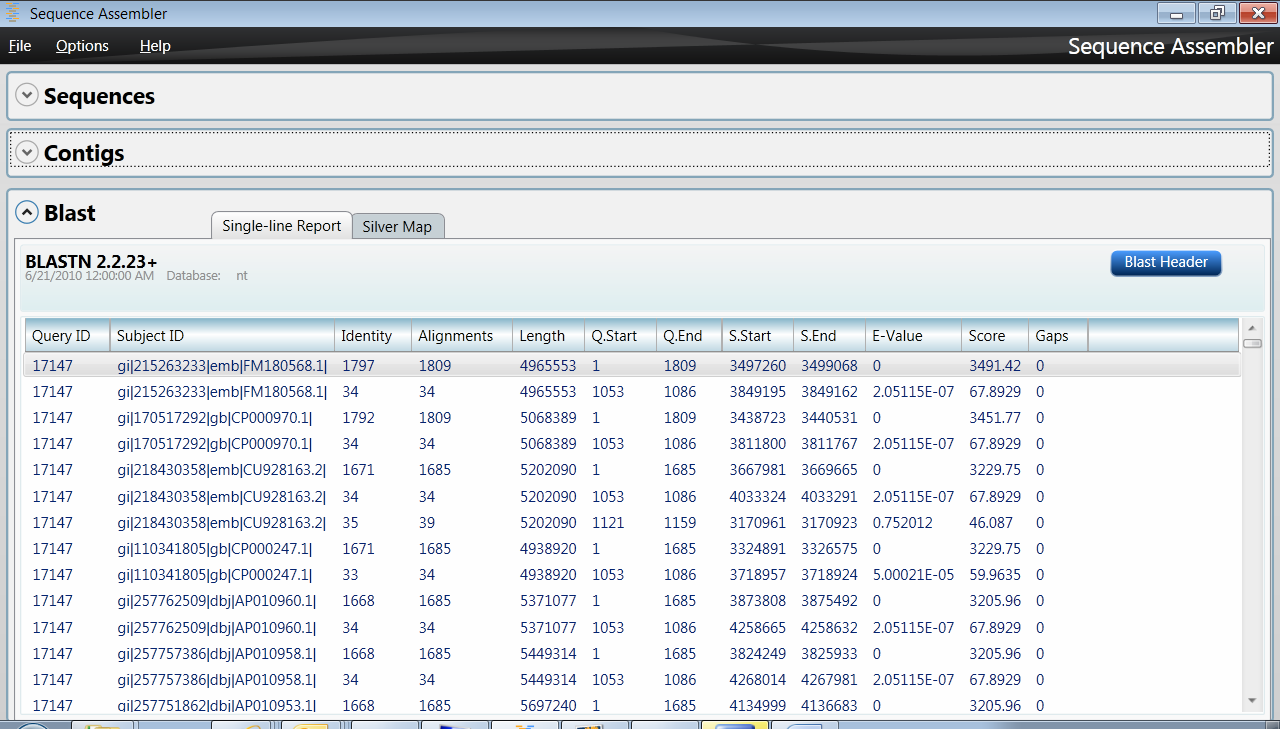
1. Select a service using Select Web service as shown in the following figure, and then click Execute Search.

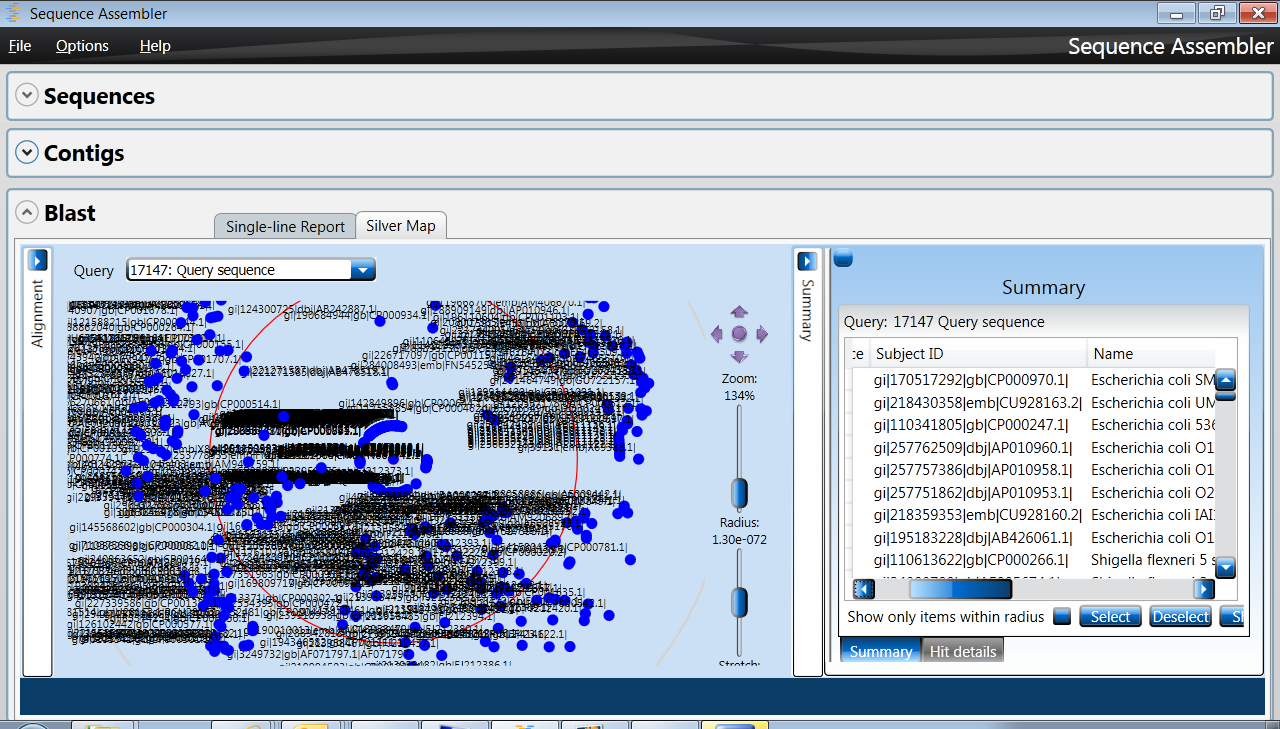
  
The Select Web service drop-down menu

1. Set the query parameters in the BLAST WebService window, as shown in the following figure, and click Submit.

  
The BLAST WebService window

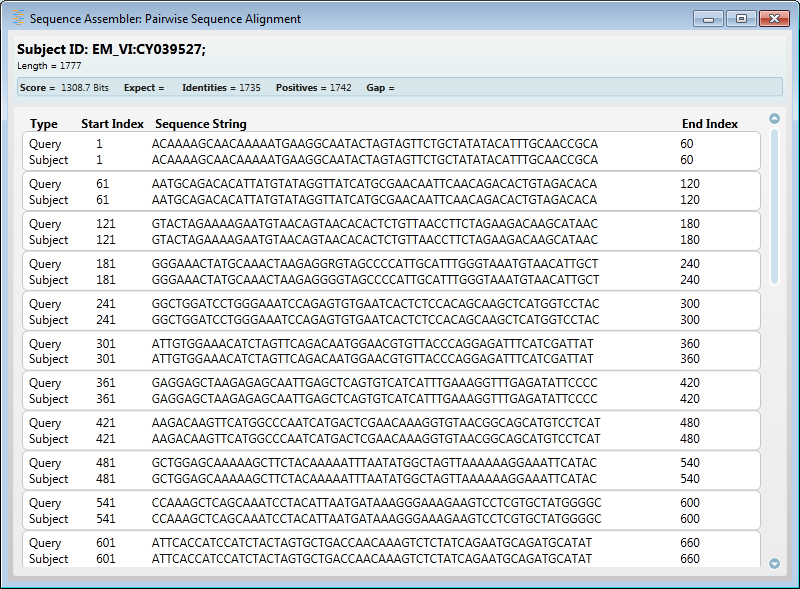
The results are displayed in the Blast pane either as a single-line report or in the SilverMap visualizer as shown in the following figures.

  
Single-line report from the BLAST service

   
The SilverMap visualizer

For information on how to use the SilverMap visualizer, see http://qutbio.codeplex.com/.

3. In the single-line report, double-click any subject ID to see the pairwise sequence alignment for that subject ID, as shown in the following figure.

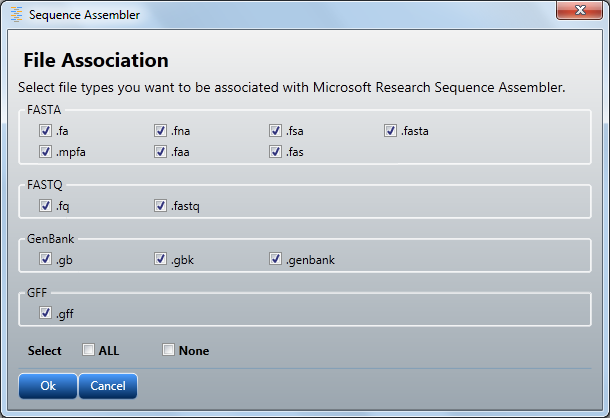
  
Pairwise sequence alignment

# How to Configure the MBF Sequence Assembler

The MBF Sequence Assembler has two configurable options: file type associations and the color scheme for the sequence data.

To associate file types

1. On the **Options** menu, click Associate File Types.

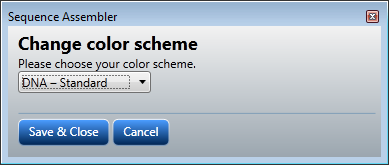
  
File Association window

1. In the File Association window, click the file types to associate with the MBF Sequence Assembler, and click OK.

To configure the color scheme

1. Click Change Colors in either the Sequence View pane or the Consensus View pane.

2. Select DNA – Standard or Protein – Standard in the Change color scheme window, as shown in the following figure.

  
The Change color scheme window

3. Click Save & Close to save your changes.

To create a custom color scheme

1. Open the Sequenceassembler.exe.config file in the C:\Program Files\Microsoft Biology Initiative\1.0\MBT\Sequence Assembler folder.
2. In the <Colors> section, change the color values for one or more symbols using standard color names.
3. Add a new color scheme by creating a new <ColorScheme> section. Be sure to duplicate the formatting and structure of an existing <ColorScheme> section.
4. Save the Sequenceassembler.exe.config file.

Listing 1. The Sequenceassembler.exe.config file

<?xml version="1.0"?>

<configuration>

<configSections>

<section name="Colors" type="SequenceAssembler.ColorSchemeConfigHandler, SequenceAssembler"/>

</configSections>

<Colors>

<!--Represents the DNA\RNA color scheme-->

<ColorScheme Name="DNA – Standard">

<Symbol Char="A" Color="Red"/>

<Symbol Char="T" Color="Red"/>

<Symbol Char="G" Color="Red"/>

<Symbol Char="C" Color="Red"/>

<Symbol Char="U" Color="Red"/>

<Symbol Char="-" Color="Red"/>

<Default Color="Red"/>

</ColorScheme>

<!--Represents the Protein color scheme-->

<ColorScheme Name="Protein – Standard">

<Symbol Char="A" Color="Blue"/>

<Symbol Char="R" Color="Blue"/>

<Symbol Char="N" Color="Blue"/>

<Symbol Char="D" Color="Blue"/>

<Symbol Char="C" Color="Blue"/>

<Symbol Char="E" Color="Blue"/>

<Symbol Char="Q" Color="Blue"/>

<Symbol Char="G" Color="Blue"/>

<Symbol Char="H" Color="Blue"/>

<Symbol Char="I" Color="Blue"/>

<Symbol Char="L" Color="Blue"/>

<Symbol Char="K" Color="Blue"/>

<Symbol Char="M" Color="Blue"/>

<Symbol Char="F" Color="Blue"/>

<Symbol Char="S" Color="Blue"/>

<Symbol Char="T" Color="Blue"/>

<Symbol Char="W" Color="Blue"/>

<Symbol Char="Y" Color="Blue"/>

<Symbol Char="V" Color="Blue"/>

<Symbol Char="-" Color="Blue"/>

<Default Color="Blue"/>

</ColorScheme>

</Colors>

<startup><supportedRuntime version="v4.0" sku=".NETFramework,Version=v4.0"/></startup></configuration>

# Appendix A: Supported Sequence and File Formats

This appendix describes the formats supported in the Microsoft Biology Foundation, with links to references and resources for more information.

### FASTA: Sequence Data

Simple text-based format for representing peptide or nucleotide sequences, so that it is easy to parse and manipulate sequences by using scripting languages such as Iron Python.

Format is a series of lines, usually at 80 characters per line, but not exceeding 120 characters per line.

Specification

FASTA Format Specification   
<http://www.ncbi.nlm.nih.gov/blast/fasta.shtml>

Resources

Overview, links to format converters, and references on Wikipedia   
<http://en.wikipedia.org/wiki/FASTA_format>

### FASTQ: Sequence Data with Quality

Text-based format that stores biological sequences and Phred quality scores in a single file. Often considered the *de facto* standard for storing the heuristic and scoring data from high-throughput sequencing analyzers.

Format is usually four lines per sequence.

Common file extensions include .fq, .fastq, .txt.

Specification

FASTQ Format Specification   
<http://maq.sourceforge.net/fastq.shtml>

Resources

Overview, links to format converters, and references on Wikipedia   
<http://en.wikipedia.org/wiki/FASTQ_format>

### GenBank: Format for Nucleotide Sequence Database

Flat-file format that describes nucleotide and nucleotide sequences from the open-access GenBank database.

Specification

“Chapter 1, GenBank: The Nucleotide Sequence Database,” Ilene Mizrachi; *NCBI Handbook*, 2007

<http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/handbook/ch1.pdf>

Resources

GenBank on the NCBI database web site  
<http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide>

Overview, links to format converters, and references on Wikipedia   
<http://en.wikipedia.org/wiki/GenBank>

### GFF: Generic Feature Format

Line-based, tab-delimited format for a record in a genome database. The GFF record represents a substring in a biological sequence, such as a gene or protein sequence, while allowing “moderately verbose” annotation.

The filename extension for such a file is .gff.

Earlier specifications translated the acronym as Gene-Finding Format.

Specification

GFF (General Feature Format) specifications document, Version n2  
Initially proposed by Richard Durbin and David Haussler, with amendments proposed by Lincoln Stein, Suzanna Lewis, Anders Krogh and others   
<http://www.sanger.ac.uk/resources/software/gff/spec.html>

Resources

Overview on the Wellcome Trust Sanger Institute site   
<http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml>

Overview of GFF on Encode Project at UCSC site  
<http://genome.ucsc.edu/goldenPath/help/customTrack.html#GFF>