BL!P User Guide

Microsoft Biology Initiative

Draft Version 0.2–December, 17, 2010.

Abstract

This document is a usage guide for the BL!P application. BL!P [*blip*], or BLAST in Pivot, is a computer program that automates the NCBI BLAST alignment of DNA or protein sequences and processes the results for visualization in Pivot. Pivot is a application developed by Microsoft that can visualize the relationships within datasets in a novel way. Specifically BL!P accomplishes the following:

1. Submits each query sequence for alignment to a NCBI BLAST database and save the results.
2. Fetches and saves the GenBank records for each BLAST hit that meets user-specified criteria.
3. Provides an interface to create customized images for each BLAST hit using data from the BLAST result and GenBank record.
4. Creates a Pivot collection that allows the user to interact with the BLAST/GenBank results and view the custom images.

About the Author

Vince Forgetta, M.Sc.

Microsoft Intern: Summer 2010

Ph.D. Candidate, McGill University, Department of Human Genetics, Montreal, Quebec, Canada

vincenzo.forgetta@mail.mcgill.ca

[Microsoft Biology Foundation is an open source, reusable .NET library and application programming interface (API) for bioinformatics research software project provided under Microsoft Public License (Ms‑PL)](http://mbf.codeplex.com/license). This documentation 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.

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

Contents

[Introduction 3](#_Toc280359881)

[Terminology 4](#_Toc280359882)

[Getting Started 5](#_Toc280359883)

[Prerequisites 5](#_Toc280359884)

[Hardware and Software Requirements 5](#_Toc280359885)

[Installation and Configuration 5](#_Toc280359886)

[How to Use BL!P 6](#_Toc280359887)

[Tutorial 6](#_Toc280359888)

[Troubleshooting Tips 16](#_Toc280359889)

[References 16](#_Toc280359890)

# Introduction

NCBI BLAST [1] is a popular software program used to find regions of similarity between biological sequences, and can be used to infer functional and evolutionary relationships between sequences. A NCBI BLAST search using multiple query sequences (e.g. gene predictions from a genome sequencing project) typically generates a large dataset that must be explored for functional or evolutionary patterns on interest. Current approaches to exploring NCBI BLAST results include automated filtering of the dataset using a priori significance thresholds followed by manual inspection. While this approach is satisfactory, novel data exploration and visualization software exists that allows for patterns to be identified more easily and with less bias. One such program is Pivot [2], which can visualize the relationship between pieces of information allowing for the discovery of hidden patterns. Pivot structures its data into “collections”, which combines groups of similar items based on values of certain attributes (facet categories), and represents each item using an image. We have created a software application, BL!P, that automates the NCBI BLAST search of multiple biological sequences and converts the results into a Pivot collection. BL!P also provides an interface to construct custom image layouts for the collection of Pivot items.

BL!P was developed using C# and .NET 4.0, and uses the Microsoft Biology Foundation [3] (MBF) bioinformatics toolkit to access NCBI resources such as NCBI BLAST and GenBank [4], as well as parsers to read/write biological sequence data.

BL!P automatically submits multiple FASTA formatted DNA or amino acid sequences to a NCBI BLAST database. Submissions are polled until complete, and the results are saved to disk. Upon completion of the NCBI BLAST search, the GenBank records for each BLAST hit that meets user specified criteria is downloaded and saved to disk. The results from BLAST and information in the GenBank records are parsed and converted to a Pivot collection. Using data from the Pivot collection, a custom image layout is constructed to represent each BLAST hit. The results are saved to disk and can be loaded into Pivot for exploration. BL!P is freely available for download via the [Microsoft Biology Initiative](http://research.microsoft.com/bio/)[5] website.

# Terminology

This section contains terminology that is specific to BL!P or Pivot. It provides a reference for terms that are used later in this user guide.

Pivot Item:

Is a collection of property/value pairings and an image. With respect to BL!P, these are property/value pairs for an individual BLAST. The paired values are parsed from the BLAST result and GenBank record. Each item also has an associated image. With respect to BL!P, this image is generated from a custom layout of aforementioned value pairs.

Pivot Collection:

Is a collection of Pivot items. With respect to BL!P, the items are individual NCBI BLAST hits.

Pivot Facet:

Is the property value of an item. An example with respect to BL!P is the Score from the BLAST hit, or the Organism field from the GenBank record.

Pivot Facet Category:

It describes the type and visibility settings of a particular facet. With respect to BL!P, these are the categories, such as Score, Identity and Species, that are available in the Filter or appear in the Info Panel (see Pivot website [2]for more information details on the Filter and Info Panel). In addition, these categories are available for creating the custom image layout.

# Getting Started

This section describes what you need to use BL!P.

## Prerequisites

To use BL!P you should be familiar with:

* FASTA sequence format: <http://www.ncbi.nlm.nih.gov/blast/fasta.shtml>
* NCBI BLAST: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>
* NCBI GenBank: <http://www.ncbi.nlm.nih.gov/genbank/>
* Pivot: <http://www.silverlight.net/learn/pivotviewer/>

## Hardware and Software Requirements

Hardware. You must have the following hardware to use BL!P:

* A PC capable of running Windows® XP SP2 or later.
* Recommended System Configuration: Windows 7 with Aero enabled, 2-GHz 32-bit (x86) processor, 2 gigabytes of random access memory. For hardware requirements specific to Pivot please refer to <http://www.silverlight.net/learn/pivotviewer/>.

Software. You must have the following software to use BL!P:

* Windows® XP SP2 or later, x86 or x64 versions.
* Microsoft .NET 4.0: <http://www.microsoft.com/net/>.
* Microsoft Silverlight 4.0: [http://www.silverlight.net/getstarted/overview.aspx](http://www.google.com/url?sa=D&q=http://www.silverlight.net/getstarted/overview.aspx&usg=AFQjCNGGIGQuEELM74B-8rPD-LMQmICK5g).
* Pivot: <http://www.silverlight.net/learn/pivotviewer/>.

## Installation and Configuration

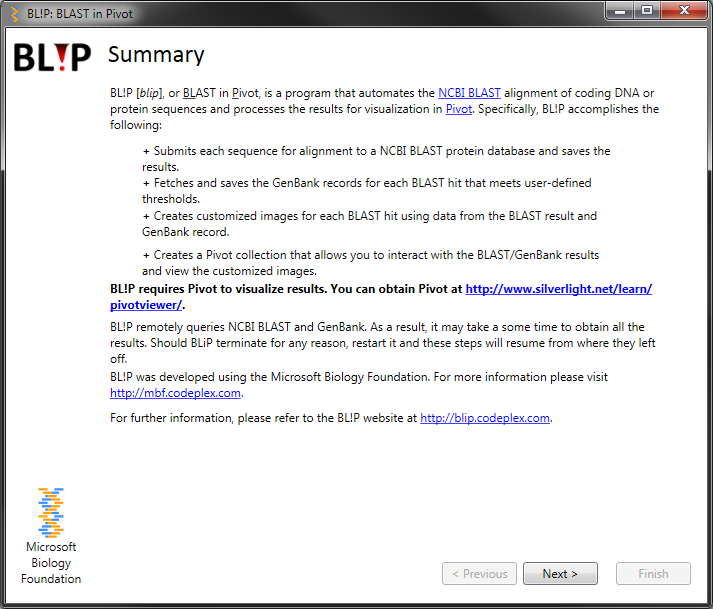
* Download and install the Microsoft .NET 4.0 Framework available at <http://www.microsoft.com/net/>
* Download and install Microsoft Silverlight 4.0 available at [http://www.silverlight.net/getstarted/overview.aspx](http://www.google.com/url?sa=D&q=http://www.silverlight.net/getstarted/overview.aspx&usg=AFQjCNGGIGQuEELM74B-8rPD-LMQmICK5g).
* Download and install the Microsoft Silverlight PivotViewer Control available at [http://www.silverlight.net/learn/pivotviewer/](http://www.google.com/url?sa=D&q=http://www.silverlight.net/learn/pivotviewer/&usg=AFQjCNEFwjiJYzvUWrBukz3B1wZqVpZyQQ).
* Download and install BL!P from <http://blip.codeplex.com>.

## How to Use BL!P

Usage of BL!P is similar to other Microsoft Windows programs known as wizards. Wizards instruct the user to complete a task, such as the installation of a program, through a series of linear steps. Please consult the Tutorial section in this document for a more detailed explanation.

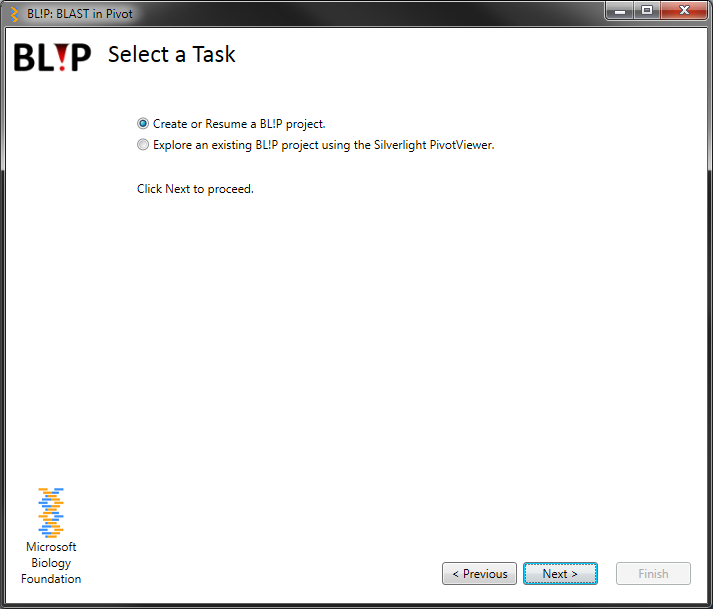
# Tutorial

This tutorial will instruct you how to use BL!P by demonstrating each step in the wizard. Each step in the wizard in represented pictorially and is accompanied by descriptive text. You may want to try BL!P while reading this tutorial using a small (< 10 query sequences) dataset of gene or protein sequences.



Double-click on **BL!P.exe** (or the Desktop shortcut) to launch the application. This will present you with a welcome screen that describes BL!P and provides additional information.

Click **Next** to proceed.

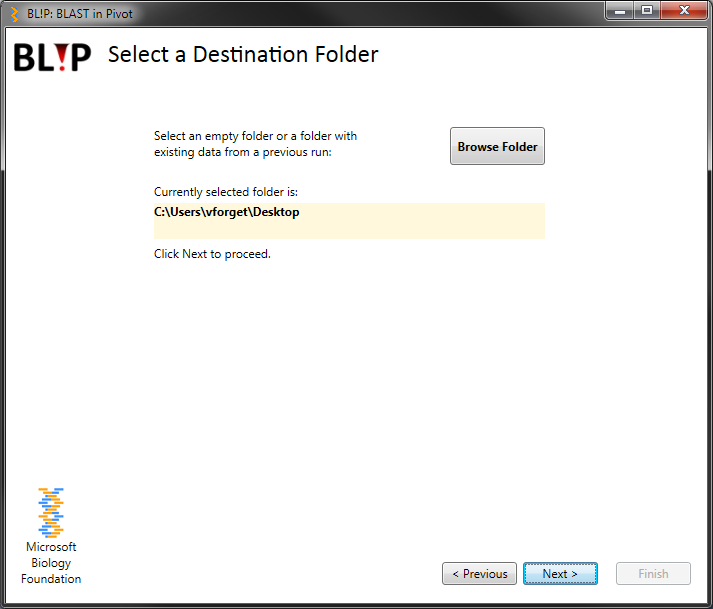


Select the task to perform:

* Create or Resume a BL!P project: Select this option of you want to submit query sequences to NCBI and convert the results into a Pivot collection, or you want to resume an incomplete process.
* Explore an existing BL!P project using the Silverlight PivotViewer: Select this option to load an existing project into the PivotViewer.

**NOTE:** The remainder of the tutorial will focus exclusively on creating a BL!P project.

Click **Next** to proceed.

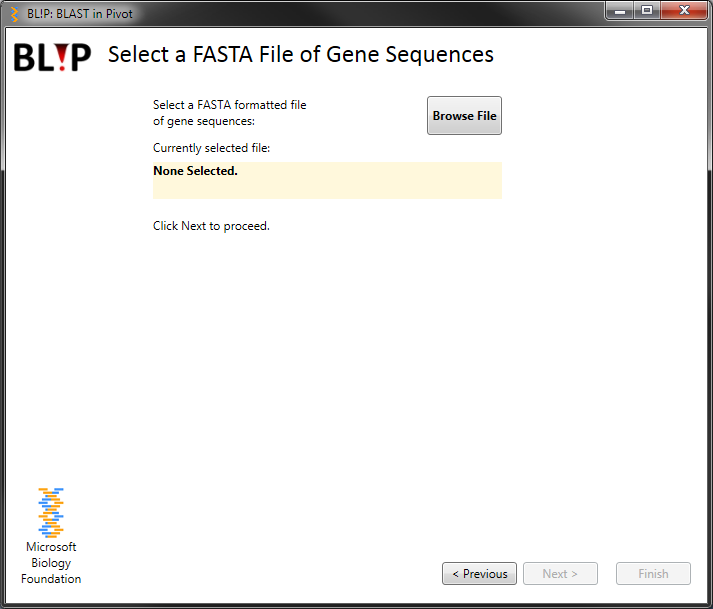


To start a new project, click the “Browse Folder” button and select a pre-existing empty folder or create a new folder.

*Alternatively, to load an existing project, select a folder that has existing BL!P project data i.e. a folder which contains a genes.fasta file, and an xml, gb, and txt folder.*

*NOTE: Loading an existing folder allows you to resume a crashed session, or to create additional Pivot collections using existing data. Selecting a pre-existing project will also resubmit incomplete BLAST searches and download missing GenBank records.*

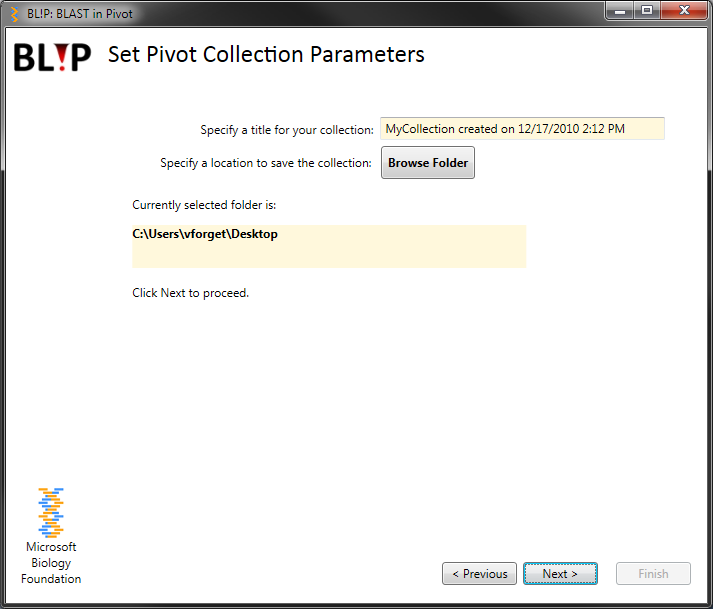
Click **Next** to proceed.



If you selected an empty project folder, you will be prompted to select a FASTA file of gene or protein sequences as input to NCBI BLAST.

*NOTE: If an existing project folder is selected, you will be unable to select a FASTA file and will be prompted to proceed to the next step.*

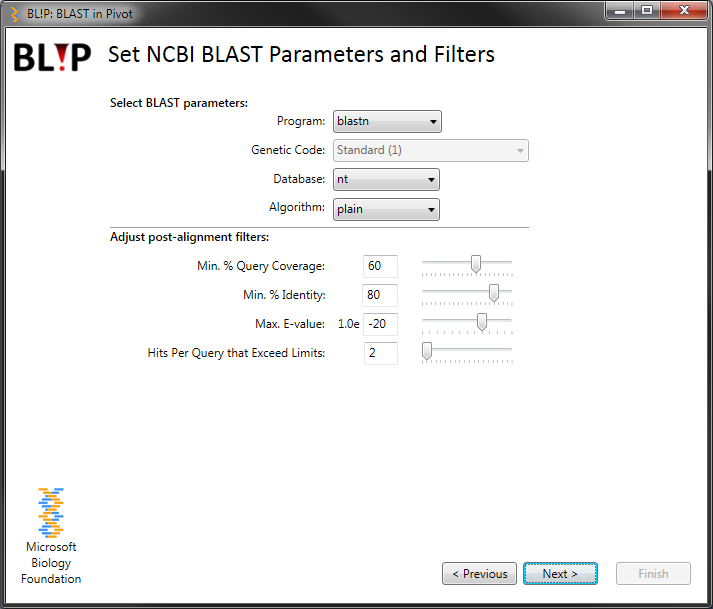
Click **Next** to proceed.



You will be prompted to provide the information required to create a Pivot collection:

1. Pivot title: The title that appears in the top left corner of the Pivot interface.
2. Location: Specify a folder where to save the Pivot collection.

Click **Next** to proceed.



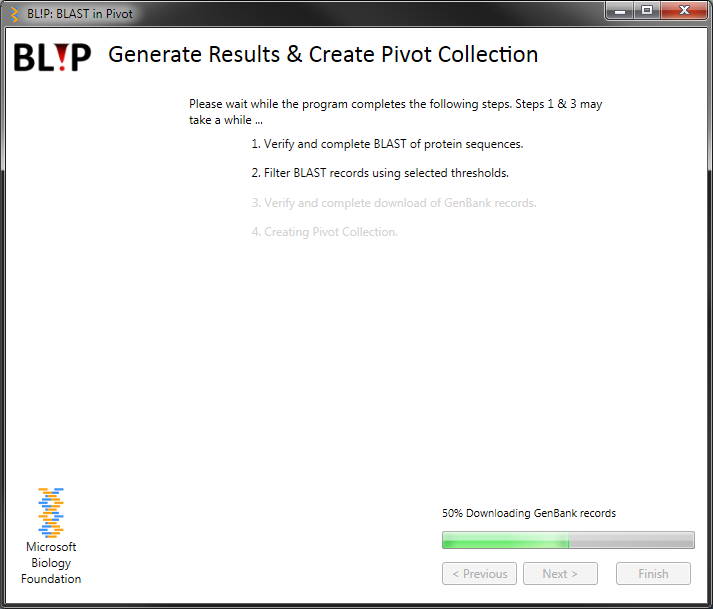
You will be prompted to select BLAST parameters:

1. Select the BLAST program appropriate for the input sequences: *blastx* for gene sequences, and *blastp* for protein sequences.
2. If you selected *blastx*, specify the genetic code to use to translate the gene sequences.
3. Select the target search database for NCBI BLAST.

Adjust the post-BLAST search parameters:

*These parameters ONLY affect the post search filtering of the BLAST results. Initially, BL!P submits query sequences to BLAST using default parameters (i.e. e-value = 10.0), and allows you to change the filters on the saved results, avoiding the need to use the BLAST service again.*

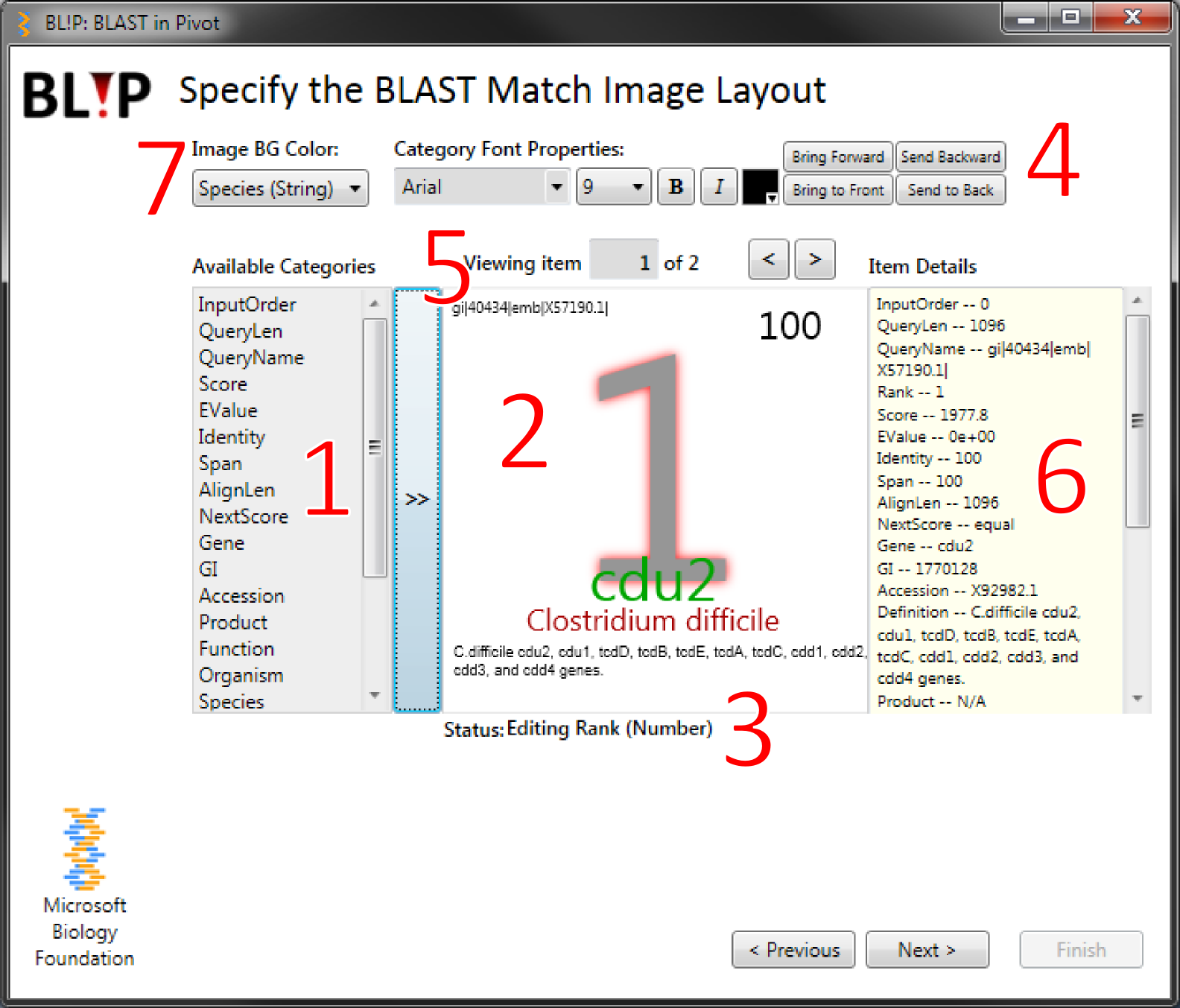
Click **Next** to proceed.



This step tracks the progress of the query sequences submitted to NCBI BLAST, the filtering of the BLAST results and the downloading of GenBank records.

*NOTE: For new projects with many query sequences this step may take some time (~30 sec/query sequence). Should the program crash during this step, the progress will resume from where it left off. Just be sure to select the same project folder during the "Select a destination folder" step.*

Click **Next** to proceed.



This step allows you to create a custom image layout from the Pivot collection. The collection contains data passed from the BLAST result and GenBank record for each BLAST hit. The interface contains the following controls:

1. Lists the Pivot facet categories available for addition to the image. Click the button to the right (**>>**) to add the selected facet category to the image. This will add the facet value for the currently loaded item (i.e. BLAST hit (5)) into the image.
2. A preview of the image. Values from facet categories selected from [1.] will appear in this area. The currently active facet category value has a red drop shadow and will be identified on the status line (3). The value in the image can be dragged with the mouse. To delete a facet category value from the image preview, and send it back to the “Available Categories” list, select it and press the Delete or Backspace key.
3. The status line shows the currently selected facet category or other mouse events.
4. These controls allow you to modify the currently selected facet category. The controls behave in a similar manner as other Microsoft Windows applications, such as Word and PowerPoint.
5. Change the currently loaded item (i.e. BLAST hit). Use the arrows to view how different BLAST hits will appear as an image, which is useful for adjusting the layout.
6. The facet names and values for the currently loaded item (i.e. BLAST hit).
7. The facet category that dictates the background color of the image. The facet values in the selected category are tabulated across all items in the collection and sorted by count. The background color is selected from a 15 color scheme according to each values rank in total count.

The categorical color scheme used is adapted from [6]:

http://geography.uoregon.edu/datagraphics/color/Cat_12.gif

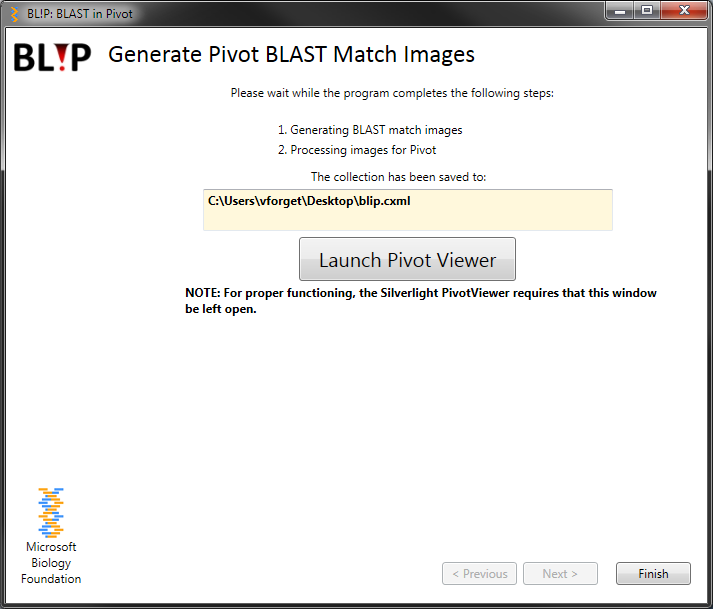
Rank (Counts):

High >>>>>>>>>>>>>>>>>Low

*For example, the image preview above displays the facet category values from item 8:*

* *Rank of the BLAST hit (large font, gray and bottom-most layer).*
* *Name of the query sequence (top left, black).*
* *Percent identity of the pairwise alignment (top-right, dark red).*
* *The gene name as found in the GenBank record for the BLAST hit. (black, bottom and centered).*
* *The species name as found in the GenBank record for the BLAST hit(red, bottom and centered).*
* *The protein product as found in the GenBank record for the BLAST hit(green, bottom and centered).*

Click **Next** to proceed.



This step creates the images for each BLAST hit and converts them to a format that is viewable in Pivot. Once complete, the location of the Pivot collection will be provided. Click the button to launch the Silverlight PivotViewer in your default web browser. Note that BL!P must remain open for the browser-based PivotViewer to function properly.

Click **Finish** to exit the program.

**Please be sure to record the location of the collection for future use.**

# Troubleshooting Tips

Problem: Double-clicking the image or clicking links in the Image Details panel does not navigate to the desired resource e.g. NCBI GenBank, or the pair-wise alignment in text format.

**Solution:** Enable popup windows in the web browser originating from address 127.0.0.1.

**Problem:** The Silverlight PivotViewer Control interface fails to load in the web browser i.e. stalls on spinning orbs.

**Solution:** Clear the web browser cache and delete web browser cookies. Reload the page or restart BL!P and load the collection using the second option in the "Select a Task" step.

# References

1. NCBI BLAST

Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ (1990). Basic local alignment search tool.J. Mol. Biol., 215, 3:403-10.

WWW: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

1. Pivot

Pivot makes it easier to interact with massive amounts of data in ways that are powerful, informative, and fun.

WWW: [http://www.silverlight.net/getstarted/overview.aspx](http://www.google.com/url?sa=D&q=http://www.silverlight.net/getstarted/overview.aspx&usg=AFQjCNGGIGQuEELM74B-8rPD-LMQmICK5g)

1. Microsoft Biology Foundation

The Microsoft Biology Foundation (MBF) is a language-neutral bioinformatics toolkit built as an extension to the Microsoft .NET Framework, initially aimed at the area of Genomics research. Currently, it implements a range of parsers for common bioinformatics file formats; a range of algorithms for manipulating DNA, RNA, and protein sequences; and a set of connectors to biological web services such as NCBI BLAST. MBF is available under an open source license, and executable, source code, demo applications, and documentation are freely downloadable.

WWW: <http://mbf.codeplex.com>

1. NCBI GenBank

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. **GenBank.** Nucleic Acids Res. 36(Database issue), D25-30 (2008)

WWW: <http://www.ncbi.nlm.nih.gov/genbank/>

1. Microsoft Biology Initiative

The Microsoft Biology Initiative (MBI) is an effort in Microsoft Research to bring new technology and tools to the area of bioinformatics and biology. This initiative is comprised of two primary components, the Microsoft Biology Foundation (MBF) and the Microsoft Biology Tools (MBT).

WWW: <http://research.microsoft.com/bio/>

1. Color Schemes Appropriate for Scientific Data Graphics

WWW: <http://geography.uoregon.edu/datagraphics/color_scales.htm>