|  |  |
| --- | --- |
|  | Microsoft Research Biology Extension for Excel: User’s Guide  Version 1.0 - June 2010 |

Abstract

This document describes how to use the Microsoft Research Biology Extension for Excel, an add-in for Microsoft® Office Excel 2007 and Excel 2010 that provides a simple and flexible way to work with genomic sequences, metadata, and interval data in an Excel document.

The Biology Extension implements several features of the Microsoft Biology Foundation: a set of parsers for common genome file formats; a set of sequencing algorithms for assembly of a consensus DNA strand; a set of connectors to several Basic Local Alignment Search Tool (BLAST) Web services for genome identification; and genomic interval functions that allow the manipulation of BED files inside Excel. The Biology Extension can be programmatically extended to use other features in the Microsoft Biology Foundation.

The Biology Extension is available at http://bioexcel.codeplex.com.

Contents

[Introduction 3](#_Toc264558413)

[How to Install the Biology Extension 4](#_Toc264558414)

[UI Overview 5](#_Toc264558415)

[How to Import Data 6](#_Toc264558416)

[How to Export Data 8](#_Toc264558417)

[How to Align Sequences 8](#_Toc264558418)

[How to Assemble Sequences 9](#_Toc264558419)

[How to Send Sequences to BLAST Services 10](#_Toc264558420)

[How to Create Charts of DNA Nucleotide Distribution 12](#_Toc264558421)

[How to Manipulate Genomic Interval Data 13](#_Toc264558422)

[How to Create Venn Diagrams from Genomic Interval Data 16](#_Toc264558423)

[How to Change Configuration Options 19](#_Toc264558424)

[Appendix A: Supported Sequence and File Formats 20](#_Toc264558425)

[Appendix B: Creating a Macro-enabled Workbook 21](#_Toc264558426)

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

# Introduction

The Microsoft Research Biology Extension for Excel is for biologists who are interested in, or already using, Microsoft Office Excel for their research. The extensible nature of Excel lends itself to the creation of a custom ribbon component labeled “Bioinformatics” in the top menu bar of the application. Through this customization, all the usual features of Excel remain available, but the biologist now has additional dedicated bioinformatics features available as well.

The Biology Extension can also be extended by developers to use other features in the Microsoft Biology Foundation or other custom bioinformatics capabilities found elsewhere. Developers can write custom bioinformatics applications using the libraries of the Microsoft Biology Foundation and add UI elements for those applications to the Biology Extension.

With the Biology Extension, you can:

* Import files containing sequence data. Supported file formats:

BED GenBank  
FASTA GFF  
FASTQ

* Align entire or partial sequences. Supported algorithms:

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

* 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

* Edit and export sequences. Supported export formats:

FASTA   
FASTQ   
GenBank

* Manipulate and display sequence data:

Manipulate genomic interval data in the UCSB BED format.

Perform operations such as Merge, Intersect, and Subtract on genomic interval data.

Create charts of sequence data.

Create Venn diagrams of genomic interval data using the NodeXL template for Excel.

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 Biology Extension

This section describes the prerequisites, system requirements, and installation steps for the Biology Extension.

#### Prerequisites

You should have a basic understanding of the following:

* Selecting ranges of cells in Excel.
* Methods and nomenclature of genomics and bioinformatics.

#### System Requirements

The Biology Extension can be installed on any computer that can run Microsoft Office 2007, as summarized at   
<http://office.microsoft.com/en-us/products/HA101668651033.aspx>

Your computer must have the following software installed:

* Any version of the Windows® operating system that can run Office 2007, which includes Windows XP Service Pack (SP) 3 and later versions of Windows
* Microsoft Office Excel 2007 or Excel 2010
* The NodeXL template for Excel, available at http://www.codeplex.com/NodeXL
* 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 Biology Extension is available at http://bioexcel.codeplex.com.

After installation, the Biology Extension appears in the Add and Remove Programs Control Panel applet as “Microsoft Research Biology Extension for Excel.”

To install the Biology Extension

1. Close all Excel documents.

2. Copy the Biology Extension installer to a folder on your hard drive.

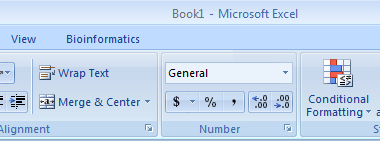
3. Navigate to that folder and double-click BioExcel.msi, which runs the installation wizard.

4. Follow the directions in the installation wizard to install the Biology Extension.

To verify the installation

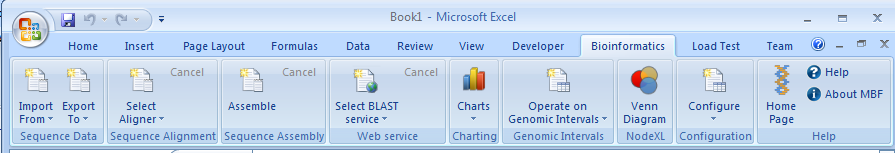
* Launch Excel 2007 or Excel 2010.

The ribbon should now include a Bioinformatics tab, as shown in Figure 1.

  
Figure 1. Excel ribbon with Bioinformatics tab

# UI Overview

The Bioinformatics tab, as shown in Figure 2, contains the primary user interface for the Biology Extension.

Figure 2. Bioinformatics tab

The Bioinformatics tab has seven command groups:

Sequence Data

Import and export sequences in the supported file formats (listed in Appendix A).

Sequence Alignment

Align entire or partial sequences with the supported algorithms.

Sequence Assembly

Assemble two or more sequences into a consensus view.

Web Service

Send consensus views of aligned sequences to a supported biological Web service for identification.

Charting

Create charts of the distribution of each nucleotide in DNA sequences.

Genomic Intervals

Manipulate genomic interval data in the UCSB BED format using operations such as Merge, Intersect, and Subtract.

NodeXL

Create Venn diagrams from aligned sequences.

Configuration

Specify the following options for the Biology Extension:

Change sequence data wraparound column width.

Change color scheme for DNA, RNA, and protein molecules.

# How to Import Data

The Biology Extension supports the following types and formats of genomic data:

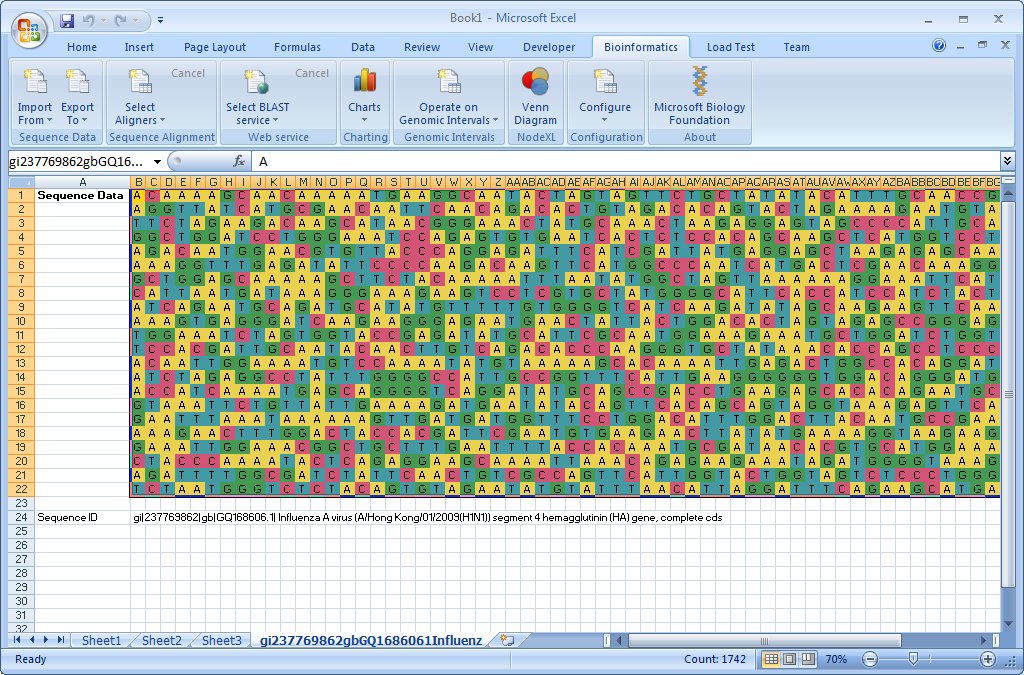
|  |  |
| --- | --- |
| Type | Format |
| Sequences of DNA, RNA, or protein | FASTA, FASTQ, GenBank |
| Sequence metadata | GFF |
| Genomic interval data | BED |

After the data is imported, you can edit it, manipulate it, send it to a BLAST Web service, and export it to new files.

To import a DNA, RNA, or protein sequence

1. Click Import From in the Bioinformatics ribbon.
2. Click the sequence format: FASTA, FASTQ, or GenBank.
3. Browse to a sequence file in the selected format and select it.
4. Click Open.

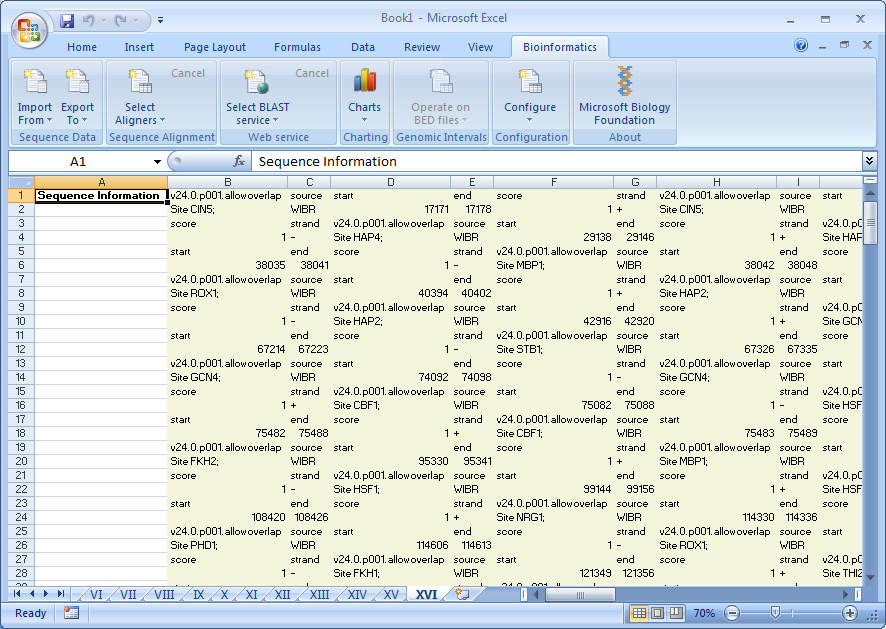
The sequence is imported into a new worksheet, as shown in the following figure.

  
A FASTA data sequence in Excel

To import GFF sequence metadata

1. Click Import From in the Bioinformatics ribbon.
2. Click GFF to select the GFF format.
3. Browse to a sequence file in the GFF format and select it.
4. Click Open.

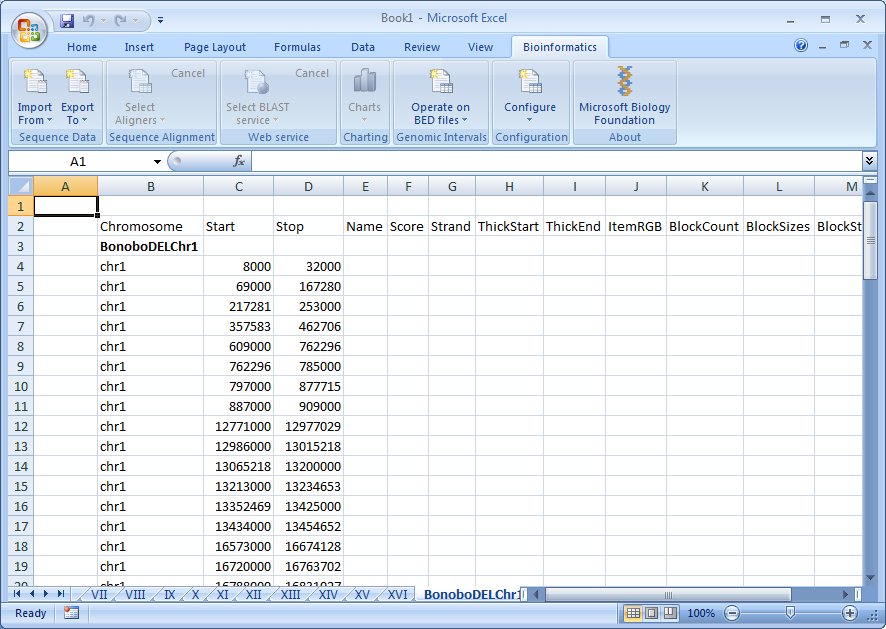
The GFF sequence is imported into a new worksheet, as shown in the following figure.

  
GFF sequence metadata in Excel

To import BED genomic interval data

1. Click Import From in the Bioinformatics ribbon.
2. Select the BED format.
3. Browse to an interval data file in the BED format and select it.
4. Click Open.

The interval data is imported into a new worksheet, as shown in the following figure.

  
BED interval data in Excel

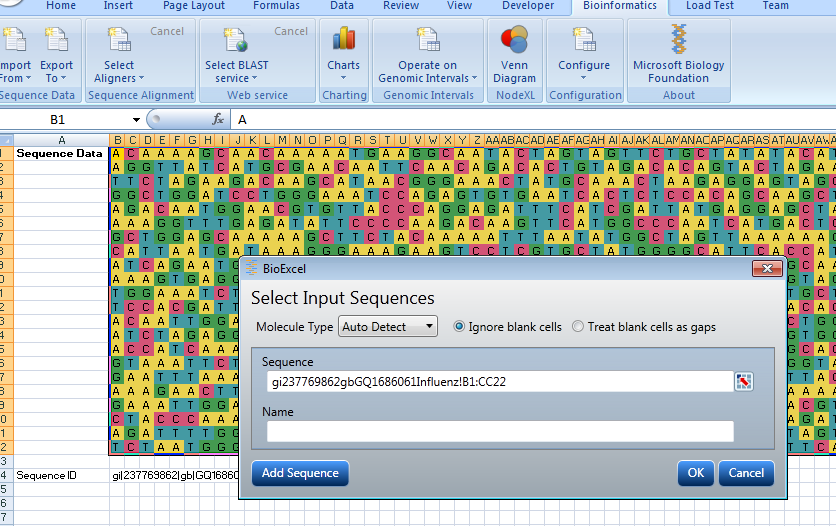
# How to Export Data

The Biology Extension supports the export of sequences of DNA, RNA, or protein in FASTA, FASTQ, and GenBank formats.

After the data is imported, you can edit it, align it, send it to a BLAST Web service, and export it to new files.

To export a DNA, RNA, or protein sequence

1. Click Import From in the Bioinformatics ribbon.
2. Click the sequence format: FASTA, FASTQ, or Genbank.
3. Browse to a sequence file in the selected format and select it.
4. Click Open.
5. Make your changes to the file and click Export To.
6. Use the Select Input Sequences window, as shown in the following figure, to select all or part of the sequences and click OK.

  
The Select Input Sequences window

1. Click Save As on the File tab, and save the file with a new name.

# How to Align Sequences

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

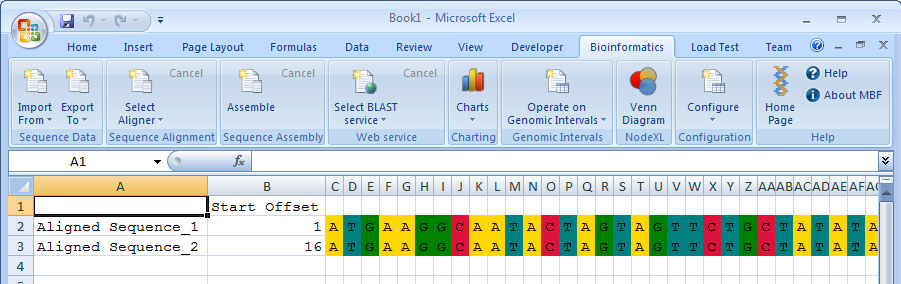
MUMmer 3.0 Pairwise-Overlap (Reference Implementation)  
Needleman-Wunsch Smith-Waterman  
NUCmer 3.0

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 a new worksheet.

To align sequences

1. Import two or more sequences of the same type, as described in “How to Import Data” earlier in this document. .
2. Click Select Aligner on the Bioinformatics ribbon.
3. Select two or more sequences in the Select Input Sequences window and click **OK**.
4. Set the parameters for the alignment in the Align Inputs Parameters window and click **OK**.

The resulting alignment is displayed in a new worksheet, as shown in the following figure.

The aligned FASTA sequences

# How to Assemble Sequences

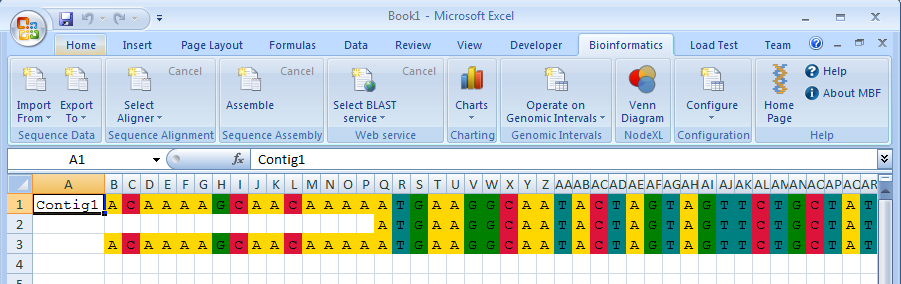
DNA, RNA, and protein sequences can be assembled into consensus views.

After selecting two or more sequences of the same type, you set the parameters for the assembly and choose an alignment algorithm. The resulting assembly is displayed as a consensus view in a new worksheet.

To assemble sequences

1. Import two or more sequences of the same type, as described in “How to Import Data” earlier in this document.
2. Click **Assemble** on the Bioinformatics ribbon.
3. Add the imported sequences in the Select Input Sequences window and click **OK**.

4. Set the parameters for the alignment in the Align Inputs Parameters window and click **OK**.

  
The aligned FASTA sequences

# How to Send Sequences to BLAST Services

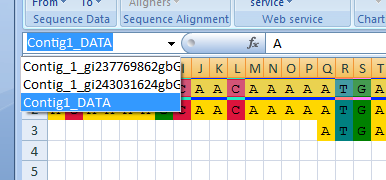
You can use the Biology Extension to send consensus views of aligned sequences to the following biological Web services for validation:

EBI WU-BLAST  
NCBI QBLAST  
AzureBLAST

After selecting an entire or partial sequence and choosing a service, you set the parameters for the query, which vary for each service. The results are displayed in a new worksheet.

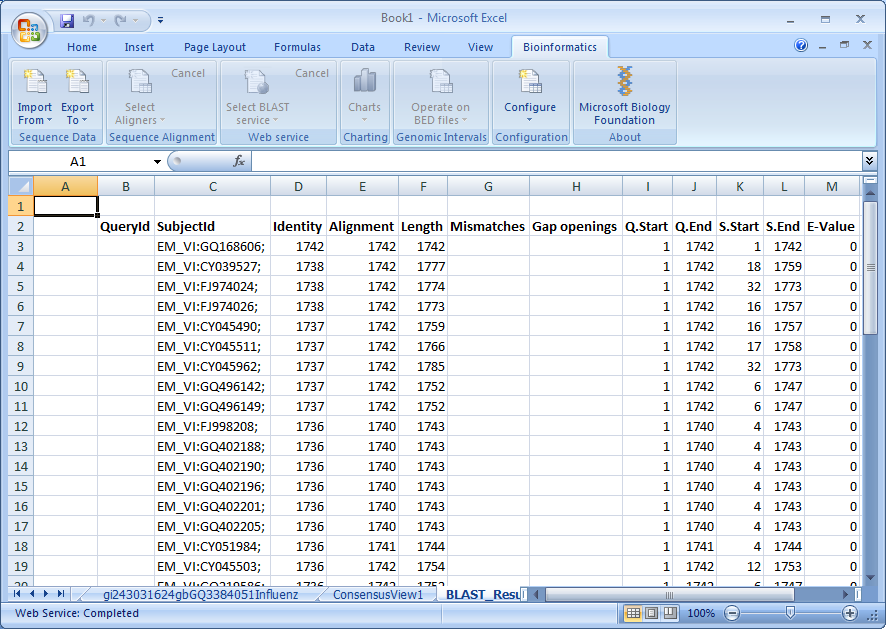
To send an entire sequence to a BLAST service

1. Select the sequence by clicking Contig1\_DATA in the Name Box drop-down of the worksheet, as shown in the following figure.

  
Selecting the consensus view of the two sequences

1. Click Select BLAST Service in the Bioinformatics ribbon.
2. Click the appropriate BLAST service to query.
3. Set the query parameters in the BLAST WebService window and click OK.

The results are displayed in a new worksheet, as shown in the following figure.

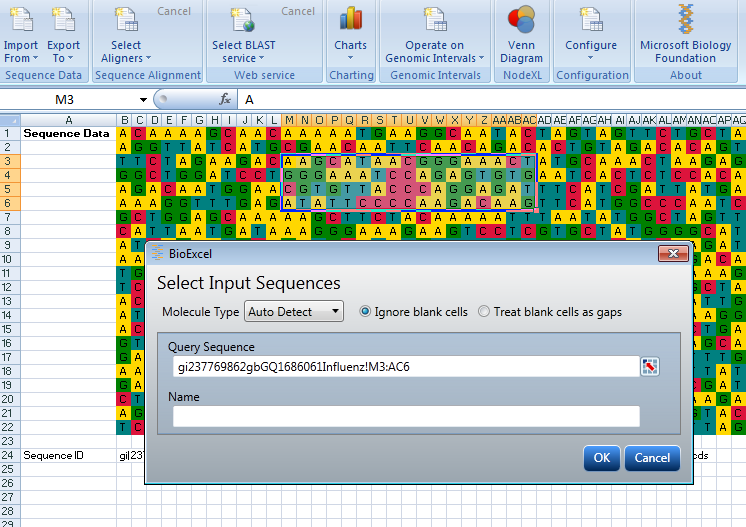
  
Results of an EBI-WU query using the consensus view of the two sequences

To send a partial sequence to a BLAST service

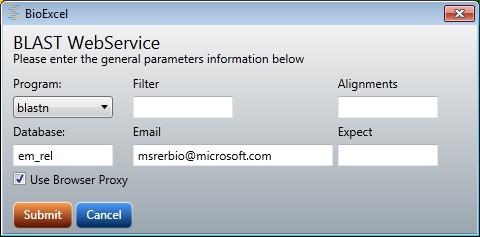
1. Using your mouse, select the specific cells of the sequence.
2. Click Select BLAST Service in the Bioinformatics ribbon and select a BLAST service.

The Select Input Sequences window displays the selection in the Query Sequence field, as shown in the following figure.

Notice that the last characters are the Excel reference to the selected cells: M3:AC6.

  
The Select Input Sequences window

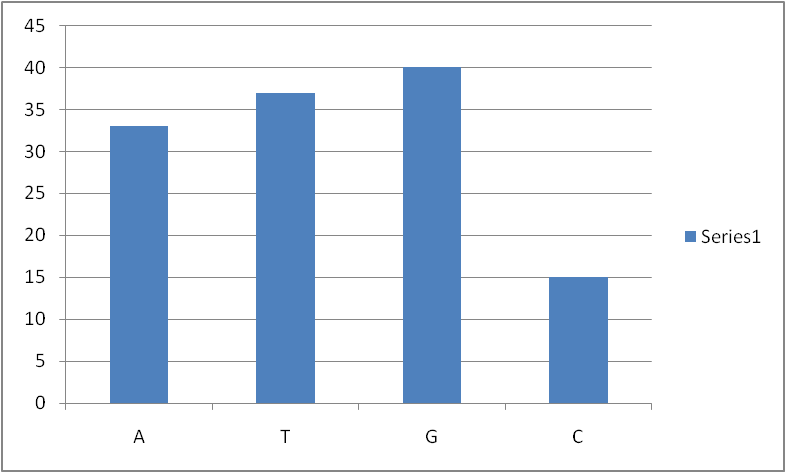
1. Set the query parameters in the BLAST WebService window and click Submit, as shown in the following figure.

  
The BLAST WebService window for EBI WU-BLAST

If the query is successful, the results are displayed in a new worksheet.

# How to Create Charts of DNA Nucleotide Distribution

With the Charting function, you can generate charts of DNA nucleotide distribution from sequence data, as shown in Figure 3.

  
Figure 3. A chart of nucleotide distribution

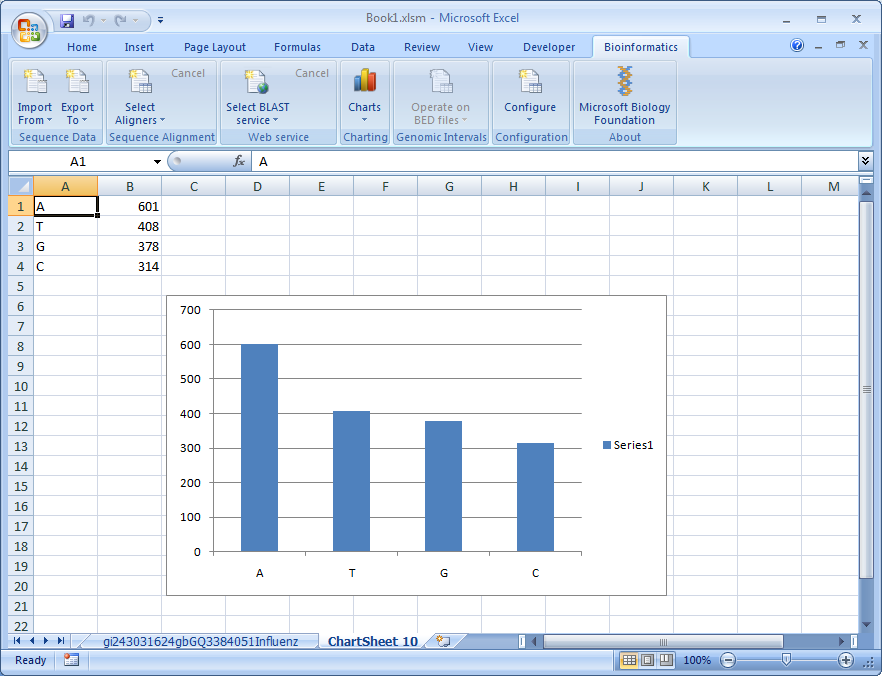
To use the Charting function, you must enable macros in Excel and add an Excel macro to the workbooks that you use with the Biology Extension. The macro is named DisplayDNASequenceDistribution.bas, and it is installed with the Microsoft Biology Foundation.

Important: Follow the procedures in Appendix B, “Creating a Macro-enabled Workbook.” before using the Charts control on the ribbon.

To create a DNA sequence chart

1. Open an Excel macro-enabled workbook that contains the DisplayChart macro.
2. Select a worksheet containing sequence data.
3. Click the Charts icon on the Bioinformatics tab and click DNA Sequence Distribution Table.

The chart is displayed in a new worksheet, as shown in the following figure.

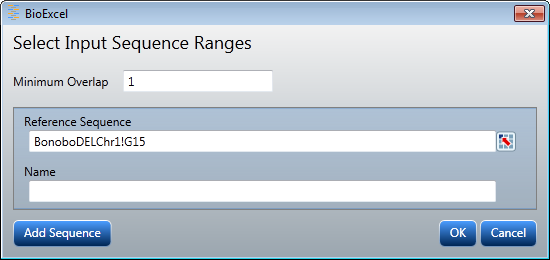
  
The new worksheet with chart

# How to Manipulate Genomic Interval Data

With the Operate on Genomic Intervals feature, you can perform the three basic genomic interval operations: Merge, Intersect, and Subtract. Using files in the BED format, you define one or more queries as ranges of worksheet cells and then select one of the three operations. The selected ranges contain one or more base pairs of chromosome coordinates. The output of each operation is a worksheet with the results of the operation.

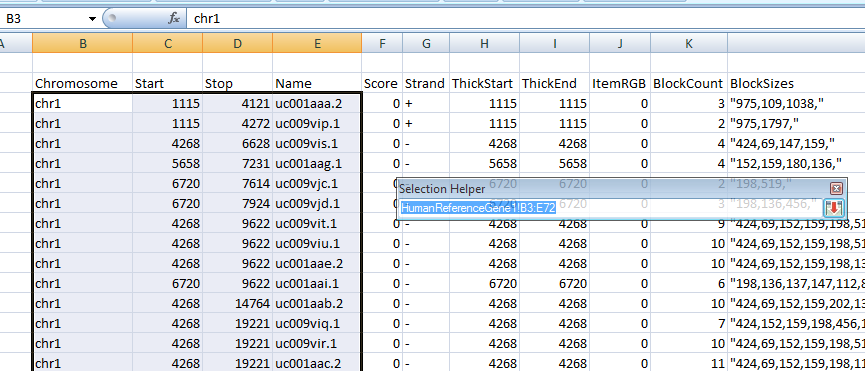
To merge the overlapping intervals of a query

1. Click Import From and click BED.
2. Select one or more files and click Open.
3. Select a worksheet and click Operate on Genomic Intervals.
4. Click Merge. The Select Input Sequences Ranges window is displayed, as shown in the following figure.

  
The Select Input Sequence Ranges window

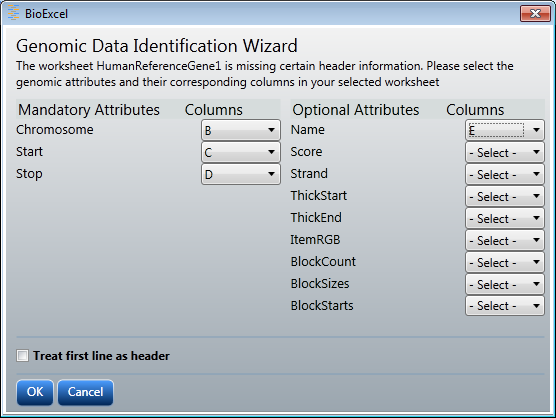
1. Click the selection icon at the right of the Reference Sequence field, and select a range of base pairs.

Notice that four columns are selected in the example shown in the following figure.

  
Selecting a range of base pairs

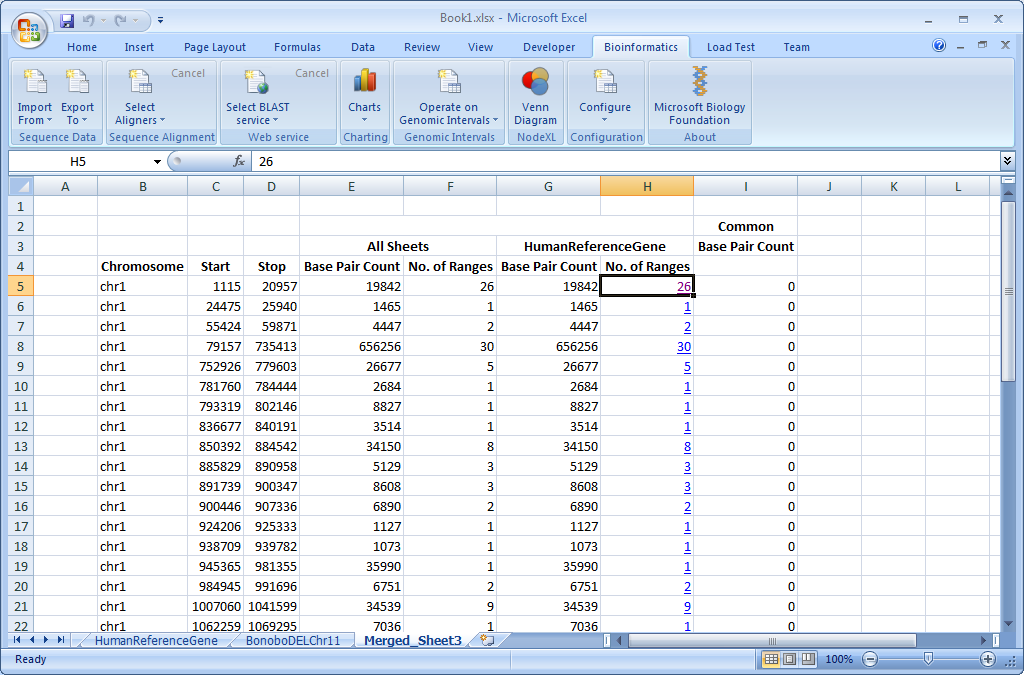
1. Click the selection icon in the Selection Helper window or press Enter to return to the Select Input Sequences Ranges window.
2. Enter a name for the query and click OK.

The Genomic Data Identification Wizard is displayed, as shown in the following figure.

  
The Genomic Data Identification Wizard

1. Use the drop-down controls to set the column names for the four selected columns—B, C, D, and E—and click OK.

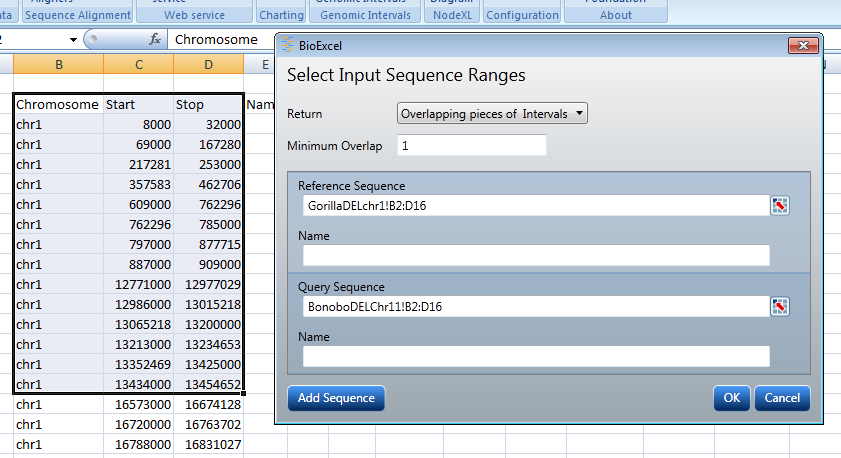
The results are displayed in a new worksheet named Merged\_Sheet1, as shown in the following figure.

  
The results of the merge operation

1. Click any of the hyperlinked values in the No. of Ranges column to see which ranges in the original worksheet were merged.

To intersect the intervals of two queries

1. Click Import From and select BED.
2. Select one or more files and click Open.
3. Select a worksheet and click Operate on Genomic Intervals.
4. Click Intersect and use the Select Input Sequences Ranges window to select two ranges of base pairs: the Reference Sequence and the Query Sequence, as shown in the following figure.

The Reference Sequence and the Query Sequence

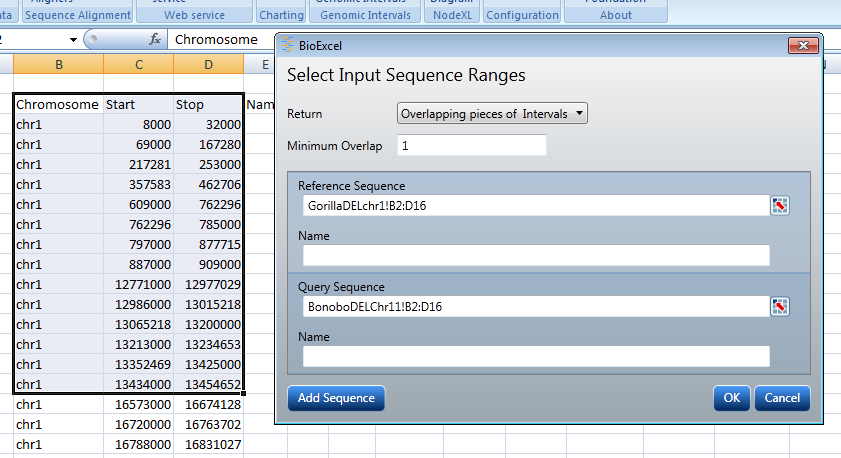
In this example, two worksheets are used, one for each sequence.

Tip: By adding the header row to the selections, you can bypass the Genomic Data Identification Wizard. The Biology Extension automatically populates the column values.

1. Click OK. The results are displayed in a new worksheet called Intersect\_Sheet1.

To subtract the intervals of two queries

1. Click Import From and select BED.
2. Select one or more files and click Open.
3. Select a worksheet and click Operate on Genomic Intervals.
4. Click Intersect and use the Select Input Sequences Ranges window to select two ranges of base pairs: the Reference Sequence and the Query Sequence, as shown in the following figure.

The Reference Sequence and the Query Sequence

1. Click OK. The results are displayed in a new worksheet called Subtract\_Sheet1.

# How to Create Venn Diagrams from Genomic Interval Data

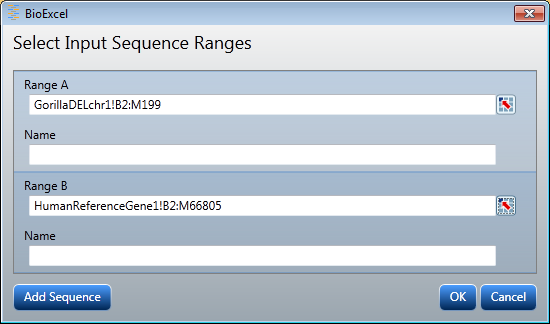
With the Venn Diagram feature, you can create two- or three-region Venn diagrams from genomic interval data in the BED format. This enables you to visualize the proportionate areas and overlaps of the data.

Note: The Venn Diagram feature requires the NodeXL template for Excel 2007 or Excel 2010, available at <http://www.codeplex.com/NodeXL>.

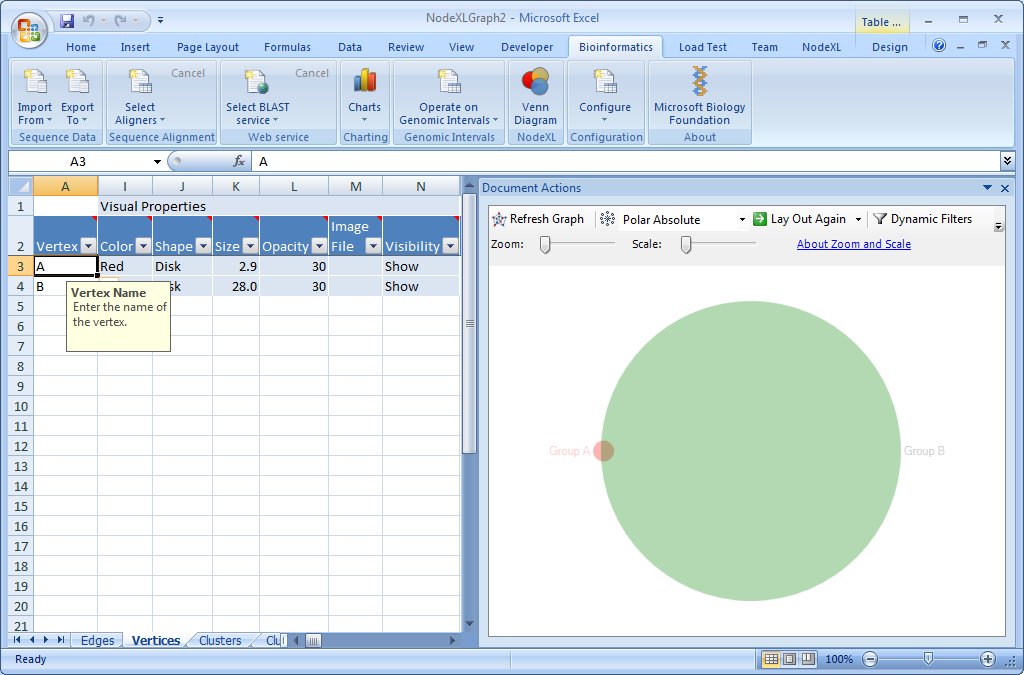
To create a two-region Venn diagram

1. Click Import From and select BED.
2. Select one or more files and click Open.
3. Select a worksheet and click Venn Diagram.
4. Use the Select Input Sequence Ranges window, as shown in the following figure, to select two ranges of base pairs.

IMPORTANT: The ranges of base pairs must have overlapping values. This example uses Chr1 of the human and gorilla genomes.

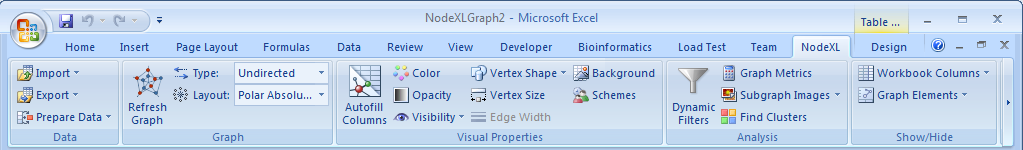
  
The Select Input Sequence Ranges window

1. The resulting Venn diagram is displayed in a new workbook called NodeXLGraph1, as shown in the following figure.

  
Venn diagram of Chr1 of the human and gorilla genomes

1. Click the NodeXL menu to see the NodeXL ribbon, as shown in the following figure.

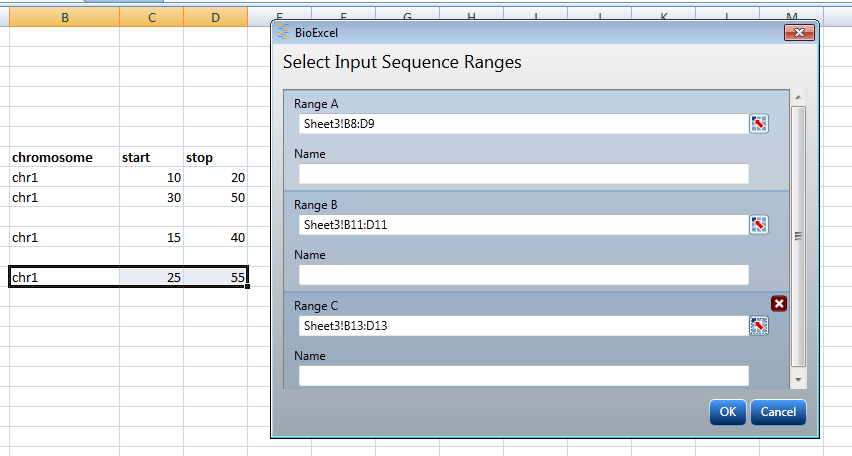
For documentation on the commands, see http://www.codeplex.com/NodeXL

  
The NodeXL ribbon

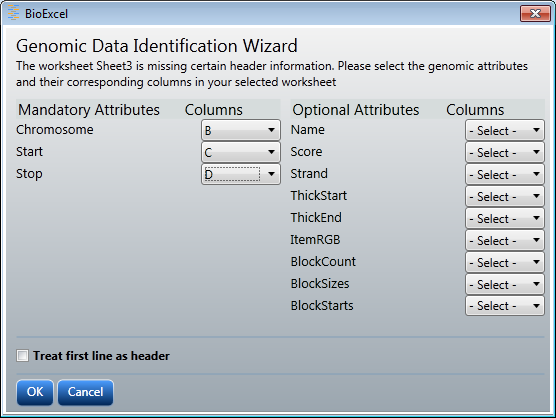
To create a three-region Venn diagram

1. Click Import From and select BED.
2. Select one or more files and click Open.
3. Select a worksheet and click Venn Diagram.
4. Use the Select Input Sequence Ranges window, as shown in the following figure, to select three ranges of base pairs and click OK.

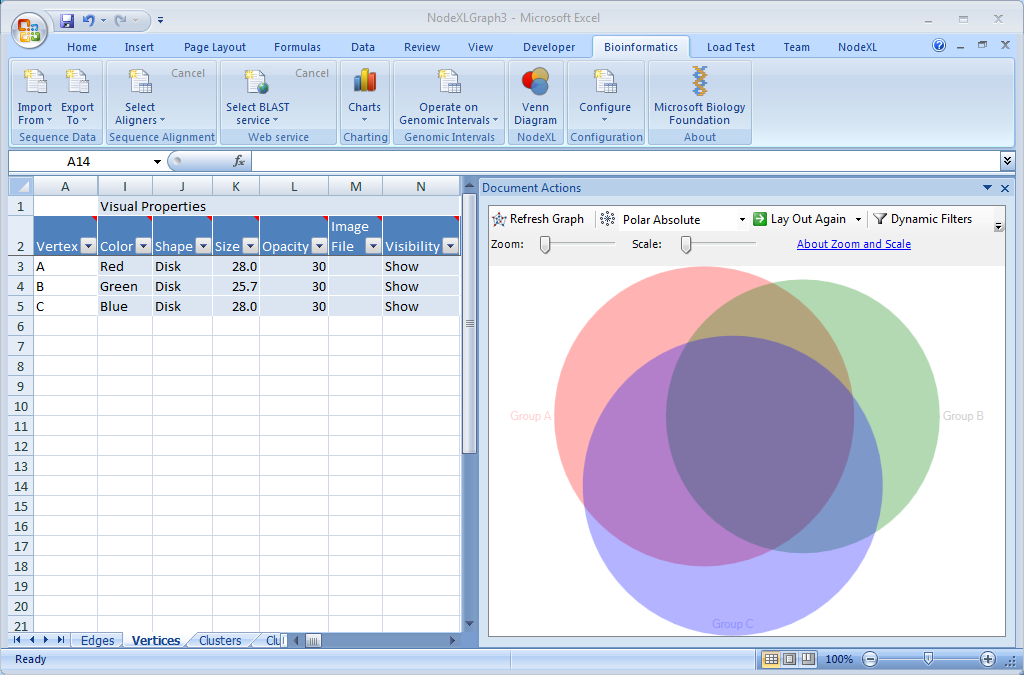
In this example, three overlapping ranges of base pairs are selected.

  
Three overlapping ranges of base pairs

1. Set the column values with the Genomic Data Identification Wizard, as shown in the following figure, and click OK.

  
The Genomic Data Identification Wizard

1. The resulting Venn diagram is displayed in a new workbook called NodeXLGraph1, as shown in the following figure.

  
Three-region Venn diagram

# How to Change Configuration Options

The Biology Extension has two configurable options:

* The number of the sequence data wraparound column.

Set this to change how Excel displays the sequence data. The sequence data default wraparound is 80 and the maximum value is 16000.

* The color coding scheme for the molecules in the sequence data.

Each molecule can have an associated color to enhance the visualization of a sequence. By default, only five molecules have a specified color: A, T, C, G, and U.

To configure the wraparound column

1. Click Configure on the Bioinformatics ribbon.
2. Click Sequence Data Wrapround Column.
3. Enter a new value in the Enter the maximum number of columns field.

To configure the color scheme

1. Click Configure on the Bioinformatics ribbon.
2. Click Change Color Scheme for Molecules.
3. Click one of the Change Color buttons in the Configure Color window.
4. Select a color in the Format Cells window and click OK.
5. Click OK in the Configure Color window to save your changes.

# 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’s easy to parse and manipulate sequences 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

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>

### Browser Extensible Data (BED) Format

The BED format provides a flexible way to define the data lines that are displayed in an annotation track. A BED format submission consists of a main file with fields separated by tabs and records separated by spaces.

The filename extension for such a file is .bed.

FAQ

Browser Extensible Data (BED) Format FAQ  
http://genome.ucsc.edu/FAQ/FAQformat.html#format1

Resources

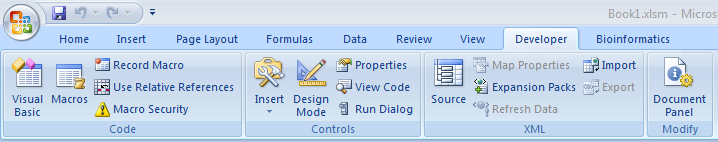
Overview of Genome Browser Database at UCSS site   
http://users.soe.ucsc.edu/~kent/gbd.html#BED

# Appendix B: Creating a Macro-enabled Workbook

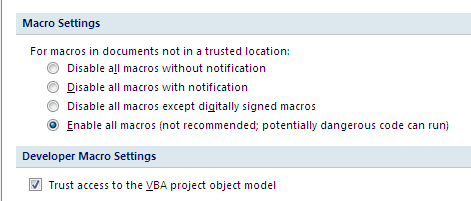
To use the Charting function, you must enable macros in Excel and add an Excel macro to the workbooks that you use with the Biology Extension. The macro is named DisplayDNASequenceDistribution.bas, and it is installed with the Microsoft Biology Foundation.

To enable macros in Excel

1. With your current workbook open, click the Office Button in Excel, and click Excel Options.
2. Click Show Developer tab in the Ribbon and click OK.
3. Click the Developer tab in the ribbon, as shown in the following figure.

  
The Developer tab

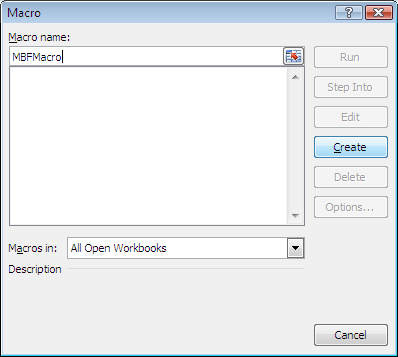
1. Click Macro Security in the Developer ribbon.
2. Click Macro Settings in the Trust Center window.
3. Click Enable all macros and check Trust access to the VBA project object model as shown in the following figure.

  
Macro Settings

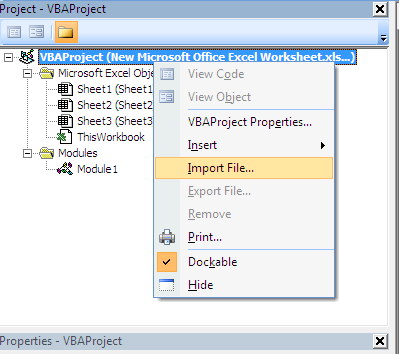
1. Click OK.
2. Close and reopen the workbook for the changes to take effect.

To add the charting macro

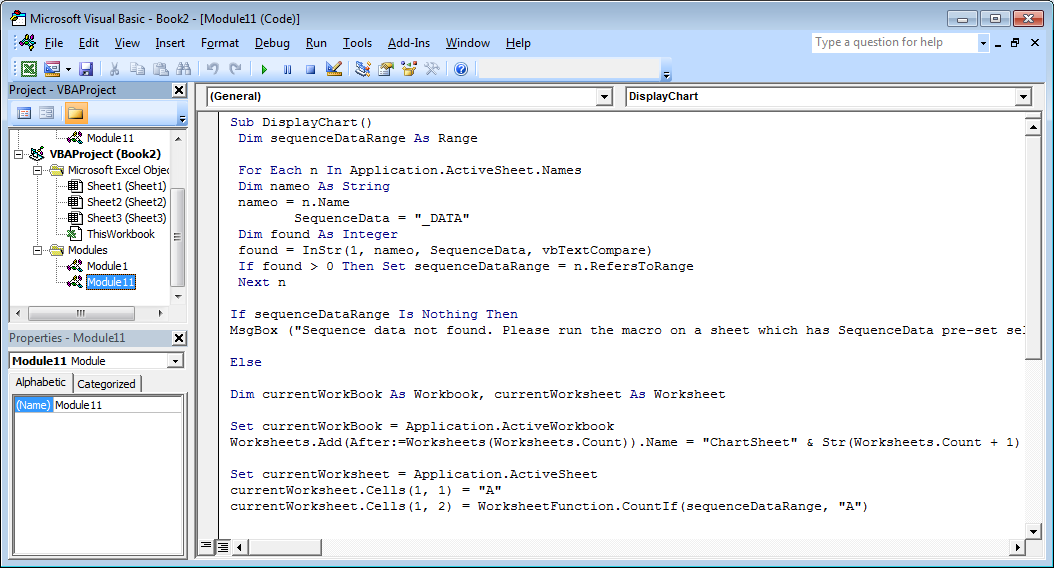
1. Click the Developer tab in the ribbon and click Macros.
2. Enter a macro name such as “MBFMacro” in the Macro name field of the Macro window, as shown in the following figure, and click Create.

  
The Macro window

1. Right-click VBAProject in the Microsoft Visual Basic window and click Import File… as shown in the following figure.

  
The Import File… command

1. Navigate to C:\Program Files (x86)\Microsoft Biology Initiative\Microsoft Biology Foundation. Select DisplayDNASequenceDistribution.bas and click Open.
2. Double-click Module 11 to display the macro as shown in the following figure.

  
The DisplayChart macro

1. Click Save and save as type Excel Macro-Enabled Workbook (\*.xlsm).

Now that you have enabled macros in Excel and added the DisplayDNASequenceDistribution.bas macro, you can use the Charting function in the Biology Extension.