BioExtract Server

User Manual

University of South Dakota

Table of Contents

INTRODUCTION

The BioExtract Server (bioextract.org) is an open, Web-based system designed to aid researchers in the analysis of genomic data by providing a platform for the creation of bioinformatics workflows. Scientific workflows are created within the system by recording tasks performed by the user. These tasks may include querying multiple, distributed data sources, saving query results as searchable data extracts, and executing local and Web-accessible analytic tools. The series of recorded tasks can then be saved as a reproducible, sharable workflow available for subsequent execution with the original or modified inputs and parameter settings. Integrated data resources include interfaces to the National Center for Biotechnology Information (NCBI) nucleotide and protein databases, the European Molecular Biology Laboratory (EMBL-Bank) non-redundant nucleotide database, the Universal Protein Resource (UniProt), and the UniProt Reference Clusters (UniRef) database. The system offers access to numerous preinstalled, curated analytic tools and also provides researchers with the option of selecting computational tools from a large list of Web services including the European Molecular Biology Open Software Suite (EMBOSS), BioMoby, and the Kyoto Encyclopedia of Genes and Genomes (KEGG). The system further allows users to integrate local command line tools residing on their own computers through a client-side Java applet.

QUERYING DATA SOURCES

Executing a Query

Creating a Query

Queries are constructed in the BioExtract Server by first checking the box for one or more data sources and then composing a query using the Query Form. The screen grab below presents an example of querying the NCBI Protein Database for the R2R3-MYB gene in Pinus taeda (Loblolly pine) (Figure 1). Queries executed against data sources residing at the BioExtract Server respond quickly. For example, a query of *Viridiplantae* returns in a matter of seconds. Response times for queries against data sources residing at other sites (e.g. NCBI Nucleotide and Protein databases) are consistent with response times for queries made directly at those sites.

Figure 1. The BioExtract Server Query page showing the data source NCBI Protein Database selected and the query form set-up to find the R2R3-MYB gene in Pinus taeda.

When attempting to locate sequence records associated with a fairly long description, you may find it helpful to use Boolean operators to break down the description into smaller units. As an example, suppose you are interested in locating nucleotide records containing the description *Homo sapiens zinc finger BTB domain.* Using the Boolean operator AND, this could be broken into *Homo sapiens **AND** zinc finger **AND** BTB domain* as illustrated in the screen grab below (Figure 2).

Figure 2. The Query page showing the Query Form using the Boolean operator AND to find *Homo sapiens zinc finger BTB domain.* Search terms can be combined with Boolean operators to create precise queries.

Executing a Wild Card Search

The BioExtract Server offers limited wild card searching functionality where the asterisk "*" is used to represent any character. For example, to locate *Arabidopsis thaliana basic helix-loop-helix (bHLH)* family of proteins, the taxonomy search value would be specified as "Arabidopsis thaliana" and the gene value on which to search would be "bHLH*". This query will return all Arabidopsis thaliana protein records that have a gene annotation beginning with bHLH.

Viewing Query Results

Viewing a Set of Records

The results of a query are displayed on the Extracts page. For each record in the result set, there are two links. The Local Details "view record" link provides quick access to details pertaining to the selected record. The External Link allows navigation to the Website hosting the original data record (Figure 3).

Figure 3. The BioExtract Server Extracts page showing the results of a query performed on the Query page. Clicking a Local Details link displays the selected record's GenBank file. Clicking an External Link takes you to the Website hosting the original data record.

Filtering a Set of Records

Records under the Extracts tab may be easily filtered to specific subsets by using the *Select Records* button. Clicking the check box to the left of a desired record adds that record to the subset. Clicking the *Select All on Page* button selects all of the records on an individual page. Clicking the *Complete* button after selecting all desired records generates a new result set containing only the filtered portion.

Exporting a Set of Records

Records under the Extracts tab may be downloaded using the *Export Records* button. The *Export Records* feature has options for downloading records in FASTA format or in the default format, which depends on the data source queried (Figure 4). For example, if an NCBI data source were queried, the default format would be GenBank.
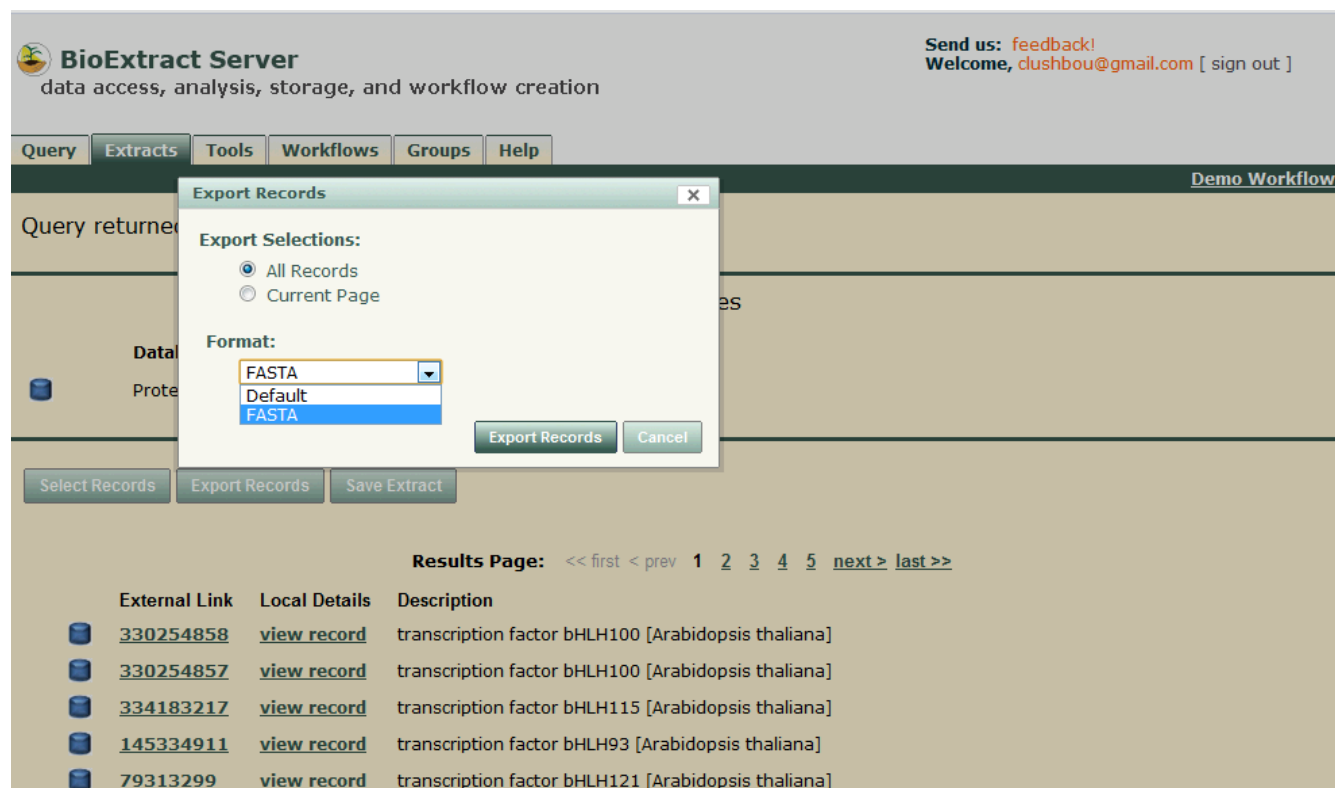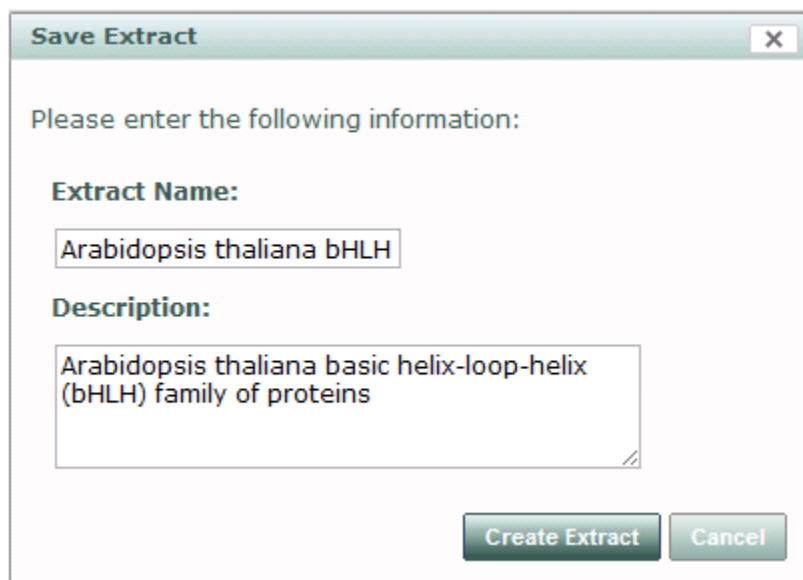
Figure 4. The BioExtract Server Extracts page showing the Export Records feature with downloading options FASTA and Default, which depends on the data source queried.

Saving a Set of Records

While most of the BioExtract Server's functionality is available to all users, the ability to save records on the BioExtract Server website is available only to users who have registered with the system.

Users who are signed into the BioExtract Server may save records as searchable data extracts by using the *Save Extract* button available from the Extracts page (Figure 5). Once saved, data extracts can be found on the Query page in Available Data Sources under the heading *Miscellaneous* (Figure 6).

All data extracts saved on the BioExtract Server website are privately owned by the user and are only made available to others by explicitly sharing them with a group. This is accomplished by: 1) clicking the Groups tab, 2) creating a group under additional actions, 3) clicking the new group name, 4) selecting the Extracts tab for the new group, and finally 5) clicking the "Add Elements" button to select the data extract to share.
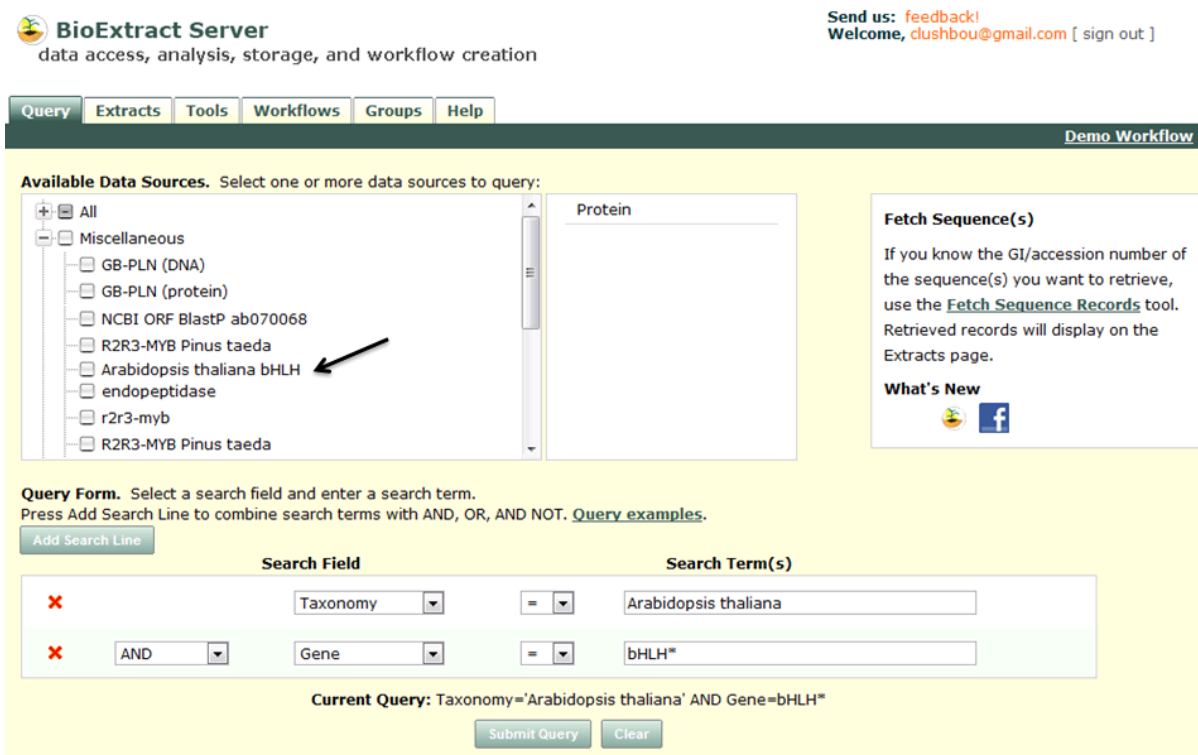
Figure 5. The Save Extract window showing Extract Name and Description entered. Clicking the *Create Extract* button creates a data extract which is stored on the BioExtract Server website.

Figure 6. The Query page showing a user's privately owned data extract *Arabidopsis thaliana bHLH* in Available Data Sources under the heading Miscellaneous.

BIOEXTRACT SERVER ANALYTIC TOOLS

Introduction

A number of well-established and unique bioinformatics analytic tools are made available through the BioExtract Server, with the majority integrated as curated Web services. Users access analytic tools through the list of Available Tools on the Tools page (Figure 7). The source of the analytic tool's input may be: 1) the records listed on the Extracts page, 2) the output from a previously executed tool, or 3) private data provided by the user (uploaded or entered in a text box). Analytic tool parameters may be selected or modified before execution and resulting output files may be viewed, downloaded and used as input into subsequently executed analytic tools.

The BioExtract Server offers users the ability to add analytic tools to their BioExtract Server workspace. Users may select such tools from a list of web services, including EMBOSS SoapLab, BioMoby and KEGG, with the integration of BioMart (www.biomart.org) currently in development. The BioExtract Server also offers generic support for other SOAP-based Web services and lets users integrate local command line tools residing on their own workstations through the use of a client-side Java applet. Analytic tools added by users may be annotated through the "Customize Your Tools" functionality, which allows users to provide detailed descriptions of the tools, as well as add a help link to additional information.

Figure 7. The BioExtract Server Tools Page showing the Available Tools with the Similarity Search Tools group expanded and the blastn tool selected.

Executing an Analytic Tool

Analytic tools are executed in the BioExtract Server by first selecting a tool from the list of Available Tools on the Tools page. One of the most important characteristics of the BioExtract Server is allowing the user to specify the input source when executing a tool. For each tool, four possible input methods are available: using the records listed on the Extracts page, using the output from a previously executed tool, entering data into a text box and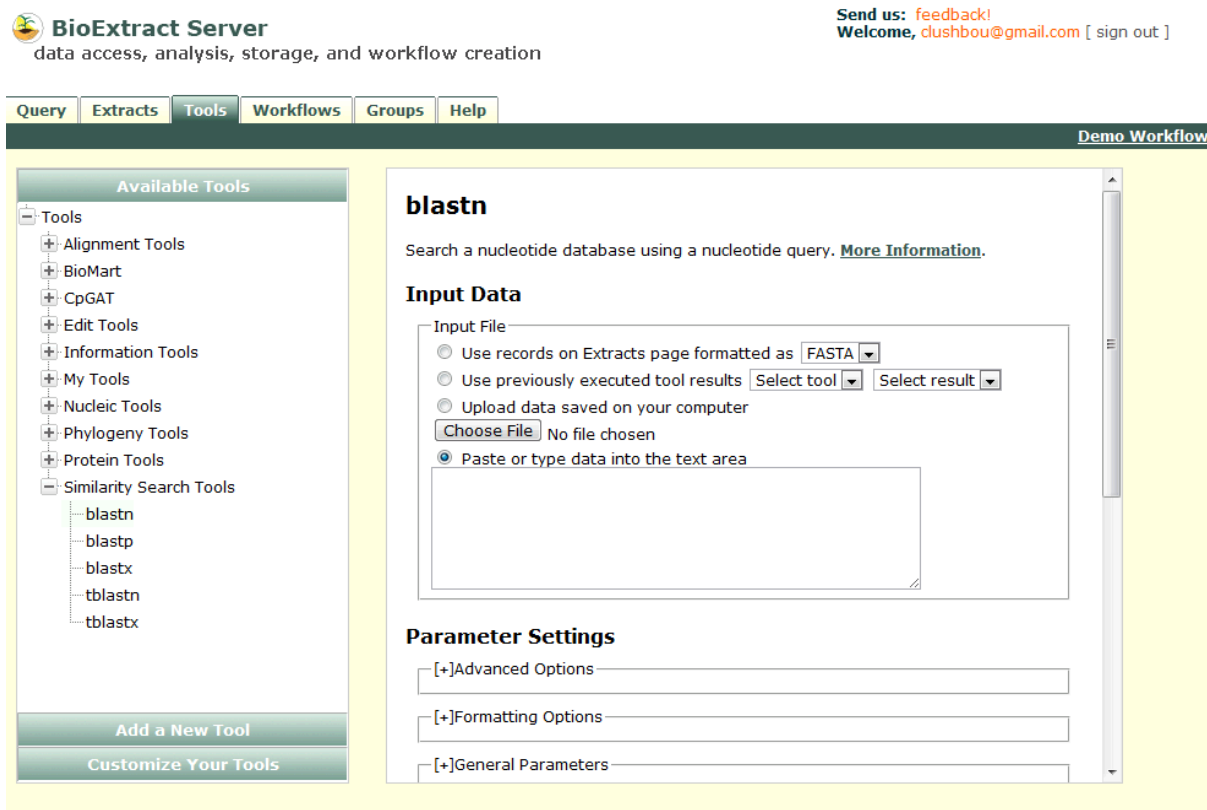 uploading a data file. When executing a tool using the "Use records listed on the Extracts page" option, the default file format is FASTA. For each record listed, the BioExtract Server system will retrieve it in FASTA format and create an input file for the tool.

Attempts have been made in the BioExtract Server to provide an appropriate level of abstraction to hide as much low-level format transformation as possible. In situations where this is not possible, users may run intermediate tools or "shims" to perform the necessary format conversions. The BioExtract Server has incorporated a predefined tool for format conversion and many of the Web services available to users have been explicitly defined for data filtering and transformation.

As an example, if the output from one analytic tool is in FASTA format and the input into another tool is required to be in GenBank format, the user may run the intermediate tool or "shim", *FormatConversion* to convert the FASTA formatted file to GenBank format. *FormatConversion* makes the conversion by parsing the id from the FASTA file and retrieving the record in GenBank format.

Creating Data Extracts Using Analytic Tools

The execution of some analytic tools results in the creation of a data extract, which displays on the Extracts page. As an example, let's create a data extract using a *Gallus gallus cadherin-19* protein sequence record and the analytic tool BLAST [25]. First, we'll query the NCBI Protein Database for Species = *Gallus gallus* and Definition = *cadherin 19*. From the query results, use the Select Records button to specify one record as the sequence of interest.

External Link	Local Details	Description
154152095	view record	cadherin-19 [Gallus gallus]

Next, click the Tools tab, open the "Similarity Search Tools" group and select the *blastp* tool. The tool form opens in the right panel. In the Input Data section select "Use records on Extracts page formatted as FASTA" and click the *Execute* button.

<u>Version 1.0</u>

A data extract will be created on the Extracts page containing sequence records similar to the *Gallus gallus cadherin-19* sequence. This data extract can be saved on the BioExtract Server website, shared with others and used as input into other analytic tools.

A Second Example With BioMart and Fetch Sequence Records

Now let's create a new data extract using BioMart and Fetch Sequence Records tools.

1. From the Tools page, expand the BioMart group and select ENSEMBL GENES 65 (SANGER UK).

2. From the expanded group, select the *Homo sapiens genes (GRCh37.p5)* dataset. The tool form opens.

3. Narrow the results by clicking the Region filter, specifying the X chromosome and q28 Band Start and Band End.

REGION:

☑ Chromosome   X ▶

☐ Base pair

Gene Start (bp)   1

Gene End (bp)   10000000

☑ Band

Band Start   q28

Band End   q28

4. Click the Attributes tab and select the Gene filter.

    a.   In Ensembl, deselect *Ensembl Gene ID* and *Ensembl Transcript ID*.
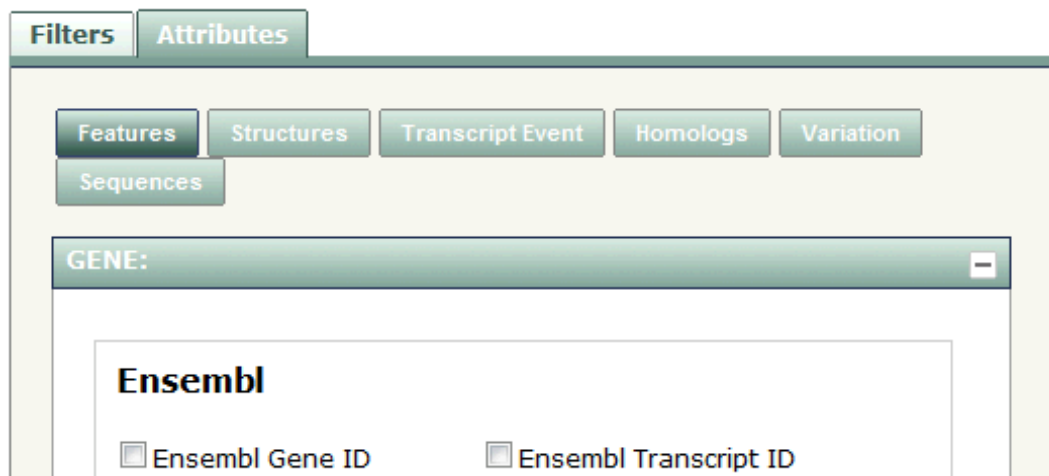
5. Click the External filter.

    b.   In External References, select *RefSeq DNA ID*.
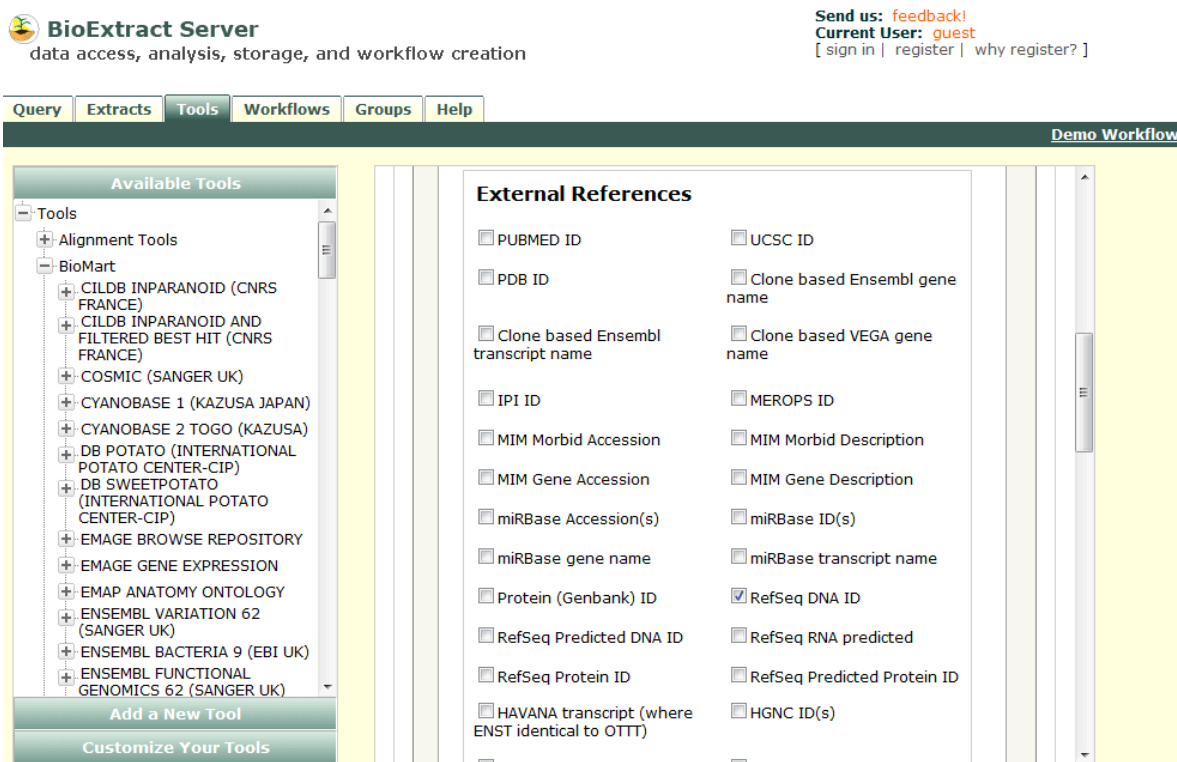
6. Scroll to the top of the tool form and click the *Execute* button.

Homo sapiens genes (GRCh37.p3)

Count   Execute   View Results

7. Now open and run the Fetch Sequence Records tool to create a data extract. From the Tools page, expand the Information Tools group and select the *Fetch Sequence Records* tool. The tool form will open in the right panel.

8. In the Input Data section, select "Use previously executed tool results". In the associated drop-down menu, select "BioMart Query".

9. In Parameter Settings > Database Options, set the database to *refseq*, then click the *Execute* button. A data extract will be created on the Extracts page. This data extract can be saved on the BioExtract Server website, shared with others and used as input into other analytic tools.

BioExtract Server
data access, analysis, storage, and workflow creation

Send us: feedback!
Current User: guest
[sign in | register | why register?]

Query   Extracts   Tools   Workflows   Groups   Help

Demo Workflow

Query returned **249** records.

Databases

	Database	Total	Database Description
	NCBI	249	

Select Records   Export Records   Save Extract

Results Page: << first < prev 1   2   3   4   5   next > last >>

	External Link	Local Details	Description
	NR_027456	view record	homo sapiens hypothetical loc100272228 (loc100272228), non-coding rna
	NM_021049	view record	homo sapiens melanoma antigen family a, 5 (magea5), mrna
	NM_001129826	view record	homo sapiens csag family, member 3 (csag3), transcript variant 1, mrna
	NM_001080848	view record	homo sapiens csag family, member 2 (csag2), transcript variant 1, mrna
	NM_001129826	view record	homo sapiens csag family, member 3 (csag3), transcript variant 1, mrna
	NM_001080848	view record	homo sapiens csag family, member 2 (csag2), transcript variant 1, mrna
	NM_001136273	view record	homo sapiens zinc finger protein 92 homolog (mouse) (zfp92), mrna
	NM_020061	view record	homo sapiens opsin 1 (cone pigments), long-wave-sensitive (opn1lw), mrna
	NM_000513	view record	homo sapiens opsin 1 (cone pigments), medium-wave-sensitive (opn1mw), mrna
	NM_001048181	view record	homo sapiens opsin 1 (cone pigments), medium-wave-sensitive 2 (opn1mw2), mrna
	NM_001586	view record	homo sapiens testis expressed 28 (tex28), transcript variant 2, mrna
	NM_005342	view record	homo sapiens high mobility group box 3 (hmgb3), mrna

Adding an Analytic Tool

Selecting From a List of Available Tools

Adding an analytic tool by selecting from a list of Web services is fairly simple, primarily because most of the information required by the BioExtract Server system can be obtained by parsing the XML Web service description and translating it into a BioExtract Tool Object. The particular GUI control that the system maps to a specific tool parameter is a function of the parameter type.

From the Tools page, use the "Add a New Tool" menu to view the list of tools, which you can add to your BioExtract Server account. The "search" text box above the list of tools lets you search for a tool by name. As an example, entering *needle* will highlight those tools containing the word "needle" (Figure 8). Clicking on the *Save Tool* button adds the tool to your BioExtract Server account.
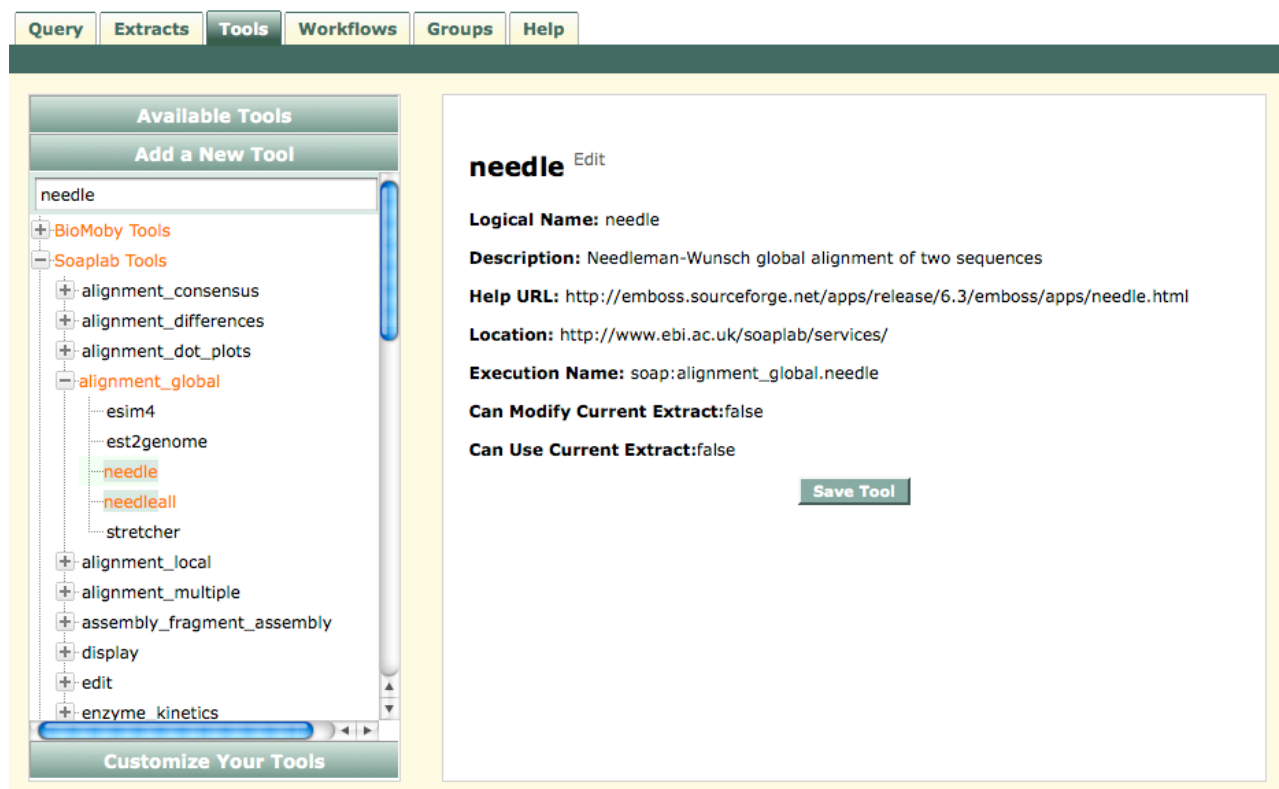
Figure 8. The BioExtract Server Tools page showing the "Add a New Tool" menu expanded with a search for tools containing the word *needle*. In this example, the Soaplab *needle* tool has been selected.

<u>Version 1.0</u>

Added tools can be found on the Tools page in the "Available Tools" menu in the "My Tools" group.
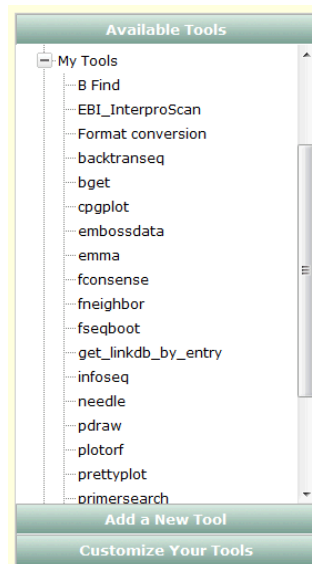
Modifying Added Tools

Users have the option to change the logical names of the tool attributes (e.g., parameters, inputs and tool name), which allows them to customize the tool interface. If the input to the tool being added represents sequence data, the user may indicate that the tool can use the BioExtract Server result set as input.

To modify an added tool:
1. From the Tools page, click the "Customize Your Tools" menu and select the tool to be modified. The selected tool opens in the right panel.
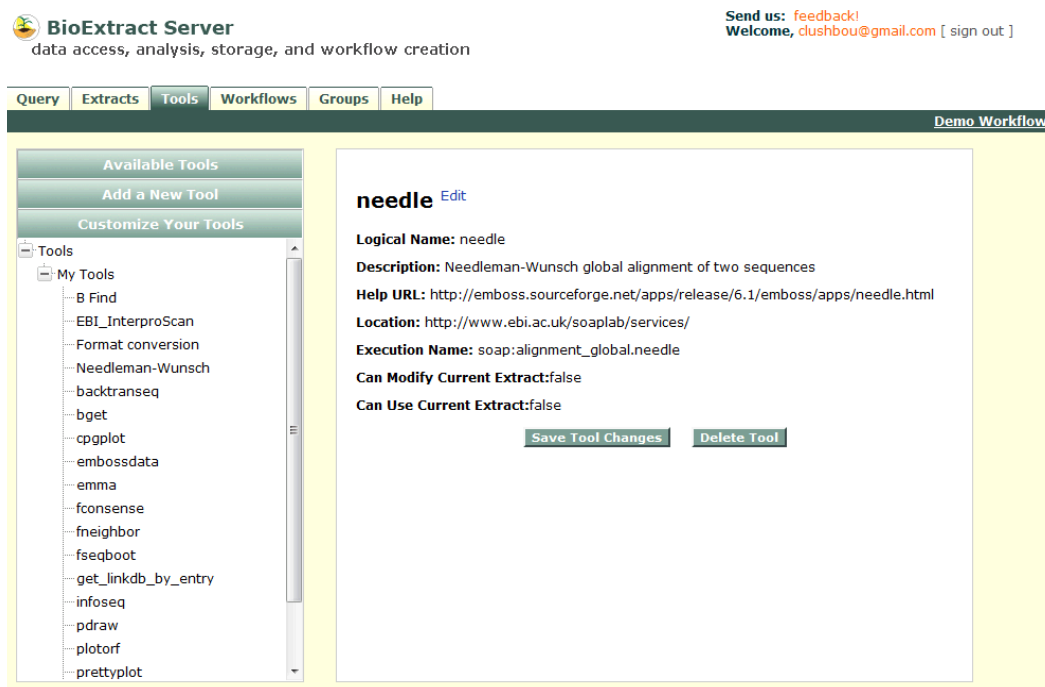
2. Click the **Edit** link adjacent to the tool name. The tool form opens.

In the tool form, you can change the Logical Name, Description and HelpURL. As an example, let's modify the tool name by entering *Needleman-Wunsch* in the Logical Name box.

3. Click the **Save** link at the top of the tool form to keep your changes. Then click the **Save Tool Changes** button to complete the process.

Adding a Local Tool

A local tool is a program on your own computer. The tool itself is not uploaded to the BioExtract Server. Rather, the BioExtract Server uses the information given about the tool to execute it on your own system. Once the tool has been added, it can be used like any of the other tools in the BioExtract Server and can even be included in workflows.

The tool must meet the following criteria:
- The tool can be executed from the command line (either a DOS prompt or Linux shell). If the tool only has a graphical or window-like interface, BioExtract Server cannot use it.
- If you wish to use the output of the local tool as input for another tool on the BioExtract Server, the name and location of any output files produced by the tool must be known before the tool is executed.

If the tool meets the criteria above, the following information is required to add the tool:
- The full path to the location of the tool on your system. For example:
  "C:\biotools\mytool.exe" on Windows
  "/usr/local/bio/mytool" on Linux

- Any required command line arguments. These are pieces of information typed after the program's name, usually specifying the name of the file it should use as input, the file name its output should be written to, etc. For example, let's say we have a tool that can convert sequences in one format to another format. The documentation for the tool may read: usage: seqconv -i filename -if formatName -of formatName
  Where:
  -i filename is the file to be converted
  -if is the format of the input file
  -of is the requested format of the output file
  Valid format names: fasta, blast, genbank We have a file in fasta format, located at "C:\temp\fasta.txt", that we wish to convert to the genbank format. According to the documentation, we could use the following command to accomplish this: seqconv -i C:\temp\fasta.txt -if fasta -of genbank

- Any additional requirements or constraints the tool may have in regards to input and output files. For example, some tools may require that all files used for input be located in a particular directory, while others may allow files in any location. Some tools may allow you to specify the name and location for each output file, while others may automatically give a name to the file and place it in a predetermined location.

As an example, suppose we have a tool called "Fetch Records", which takes a file containing a list of ids and the name of a data source (i.e. ncbi, embl, refseq). The tool accesses the data source and retrieves the records associated with the ids.

Step 1: Click on the Tools tab, then select **Add a New Tool** from the menu that appears on the left.

Step 2: Click on **Add a Local Tool**, then **New Local Tool**. Now, click on the **Edit** link next to "New Tool" in the right hand panel.

- **Logical Name:** a name for the tool that will be used in the BioExtract Server. It does not have to match the actual name of the program. For our tool, names such as *Fetch Records.*

- **Description:** Optional. A description of the tool.

- **HelpURL:** Optional. If there is a website associated with your tool, you can enter a link to it here.

- **Location:** Not used for local tools. The field should be left blank.

- **Execution Name:** Enter the full path of your tool here.
  If your tool is a Java JAR file (e.g. ResultSetDB.jar), this should be "java -jar" followed by the full path of your JAR file. For example, on a Linux system the Execution Name might be "java -jar /usr/local/tools/ResultSetDB.jar", and on a Windows system it might be "java -jar C:\ResultSetDB\ResultSetDB\dist\ResultSetDB.jar". Depending on your system configuration, you may need to replace 'java' with the full path to the java command (e.g. C:\Program Files\Java\jre6\bin\java.exe).

- **Can Use Current Extract:** Not used for local tools, and should be left unchecked. As a note, local tools *can* use the current extract as input, but this checkbox has no effect on that at all.

Step 3: Once this information is entered, click the **Save** link at the top of the form.

Step 4: Specify input files.

Data from the BioExtract Server is streamed to the local tool by using one or more input files, created on your computer when BioExtract Server runs the tool. The data stored in each input file can come from one of four sources, selected when the tool is executed:

1. The current extract (records on Extracts page)

2. A previous tool's output

3. An input file uploaded from your computer

4. Text typed or pasted directly into a text box

Not all tools require input files, so you may not need to add one at all. Our example tool "Fetch Records" requires a single input file containing a list of Ids representing sequence records in an external data source.

- **Physical Name:** We're going to skip this one for now. It will make more sense later after going through the rest of the fields.

- **Logical Name:** As with the tool itself, the logical name is only used within BioExtract Server. It may help to use the type of data as the logical name (especially if the tool has more than one input). Since Fetch Records expects this input file to be a list of Ids, we'll use *ids* for the logical name.

- **Description:** Any additional information or notes about the tool can be written here.

- **Record Number Limit:** If the current extract is being used as the input, the number of records included will be truncated to this amount. 0 means "no limit".

- **File Name:** Enter the name (with full path) that BioExtract Server should use to create this file. For Fetch Records' input file, we will use *C:\biotools\id_input.txt*.

- **Data Types:** Not used for local tools.

- **Uses Current Extract:** This control isn't used for local tools, and has no effect at all on whether the input can come from the current extract. It's best to leave it unchecked.

Inputs Create New

New Input

Save Delete Cancel

Physical Name: -i

Logical Name: Ids

Description:

Unique sequence identifiers

Record Number Limit: 0

File Name: C:\biotools\id_input.txt

Data Types:
Amino Acid
Nucleotide Sequence
Protein Sequence
Sequence

Uses Current Extract: ☐

File Size Limit: 0

Include in Command Line: ☑

- **File Size Limit:** If the size of the input file should be restricted to a certain number of bytes, you can enter that number here. This should only be needed if your tool has a limit on the size of the input files given to it. Most tools can accept files of any size, so if you're not sure, it's probably safe to leave this at the default value of 0 (no limit).

- **Include in Command Line:** If your tool expects the name of the input file to be given after its name in the command line, this box should be checked. If **Include in Command Line:** ☐ is checked, BioExtract Server adds the file name to the command used to run to the tool. If **Include in Command Line:** ☐ is left unchecked, the file will still be created with the name

given by **File Name:** C:\biotools\id_input.txt , but its name will not be added to the command line.

- Some tools require a "switch" (like -i or -f) before files given in the command line. For example, Fetch Sequence requires a "-i" before the input file. The command line would contain: " -i C:\biotools\id_input.txt "

- Remember the **Physical Name** field we skipped earlier? If the tool requires a switch before the input file, enter the switch in the Physical Name field.

- Once all the information about this input has been entered, click the **Save** link at the top of the form. Any number of inputs can be added, depending on the requirements or limitations of your tool.

Step 5: Specify output files.

The interface for defining output files is very similar to the one used to define input files. As with input files, output files are optional and need only be defined if you would like to use the output of the tool as input to other tools on BioExtract Server.

Physical Name, Logical Name, Description, File Name, and Include in Command Line behave exactly the same as they do for Input files.

Record Number Limit, Description File Name, and Modify Current Extract are not used for local tool output files and can be ignored.

Step 6: Command Line Parameters/Arguments

Most tools have a set of options whose values are given by arguments on the command line. Specifying the input and output files for the program is just one example of such arguments. For our purposes here, the terms "argument" and "parameter" are identical.

BioExtract Server uses the information given about each parameter to add an element in the completed tool's interface where the value for the parameter can be given. This is the interface shown for Fetch Records' parameters:
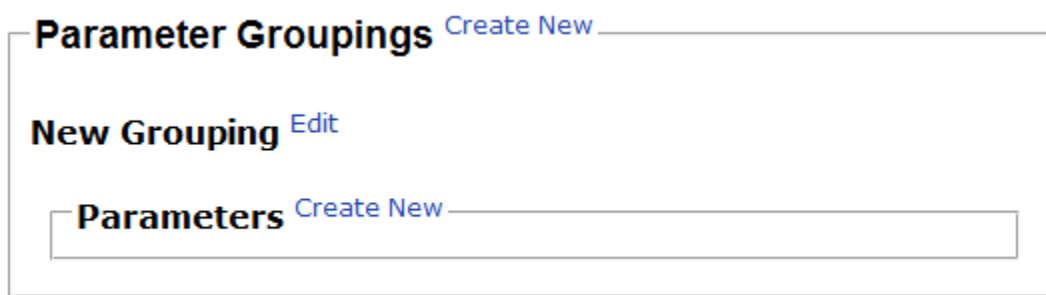

Before adding a Parameter, you must first add a Parameter Group. Multiple parameter groups are allowed, and can be used to keep related parameters together. For example, a tool may have a set of parameters

that affect the appearance of the tool's output. These parameters could be placed within a group called "Output Options", and will be displayed together in the menu used to run the tool from BioExtract Server.

Our example tool, Fetch Records, has one parameter, which specifies the name of the database containing the desired sequence records. On the command line it would appear as *–db ncbi.*

- To add parameters, begin by clicking on the **Create New** link next to "Parameter Groupings". Then, click **Edit** next to "New Grouping".
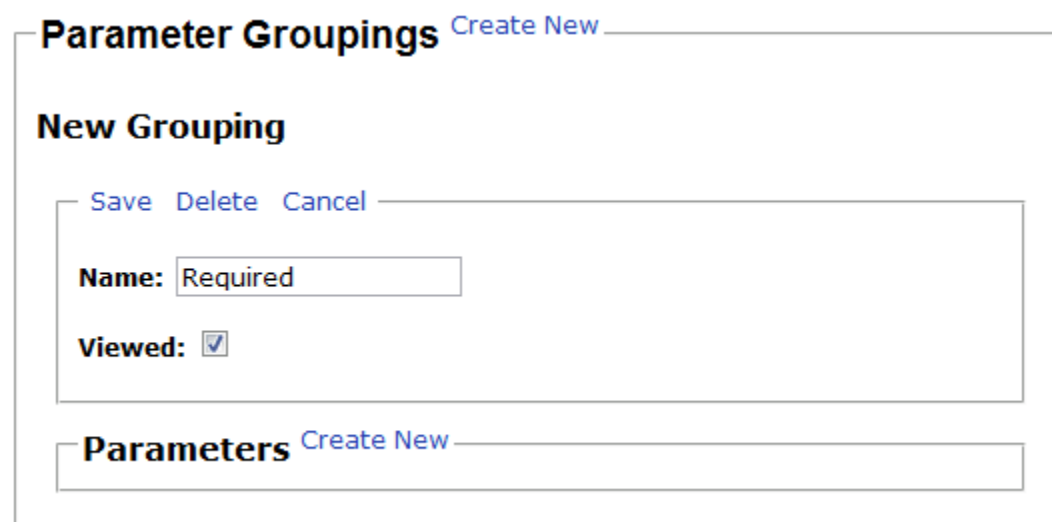

- Assign a name to this parameter grouping. The **Viewed:** ☑ check box allows you to specify if you would like to have the parameter group expanded when the tool is being executed.
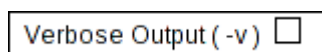

- Now, click **Create New** next to Parameters, and then click the **Edit** link next to "New Parameter". Below is a description of each field required to define a parameter. After describing these fields in general, we will demonstrate how they were used to add parameters for the mutate tool.

  - **Logical Name:** As with Inputs and Outputs before, the logical name is the name used for this parameter within BioExtract Server.

  - **Physical Name:** This is the actual parameter name, as it should appear on the command line. Note that BioExtract Server will not add a "-" automatically, so please remember to do so if your tool requires a "-" before the parameter name.

  - **Description:** Optional; Additional information about the parameter.

- **Parameter Type:** This specifies the way values are given for the parameter. Depending on the type, they may be entered directly by the user or chosen from a list of pre-defined values. Below are the details for each type:
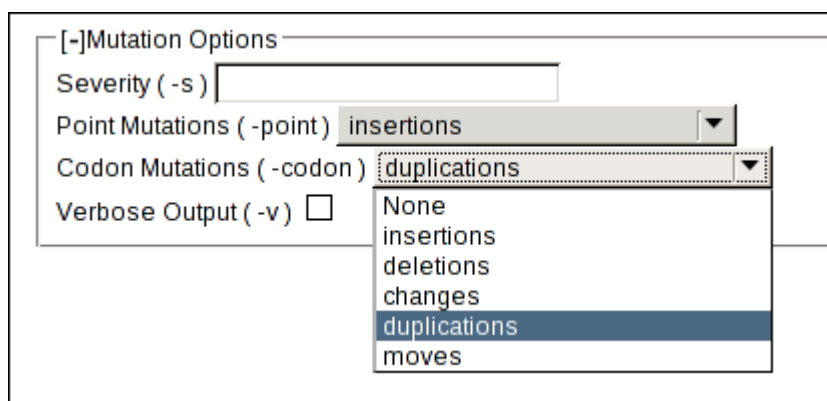   - **text:** Creates a field where the value can be typed in directly. Also useful for numeric values.

      Severity (-s) [   ]

    - **checkbox:** Creates a checkbox. Useful for parameters that don't have any additional values (for example, some programs will print extra information if a -v is present on the command line). If checked, the parameter (specifically, the **Physical Name**) will appear on the command line.

      Verbose Output (-v) ☐

    - **select:** Creates a "drop-down" menu from which one of several possible values can be chosen.

      [-]Mutation Options
      Severity (-s) [   ]
      Point Mutations (-point) insertions ▼
      Codon Mutations (-codon) duplications ▼
      Verbose Output (-v) ☐
      None
      insertions
      deletions
      changes
      duplications
      moves

      When using the "select" type, the possible values must be defined. This is not required for any of the other types. To define a set of values, click on the **Create New** link next to "Values". Then, click **Edit** next to the "New Parameter Value". "Value" is the value as it should appear on the command line. If "Is Default" is checked, this value will be selected by default when the tool's interface is shown.
   - **textarea:** Identical to "text" type above.
- **Tab Order:** Optional; If you would like the parameters to be displayed in a certain order, this number gives the rank of this particular parameter. The ordering is from lowest-to-highest, so a parameter with a tab order of 2 appears above one with an order of 3.
- **Is Mandatory:** If checked, the tool will not be able to execute unless a value for this parameter has been entered.
- Click the **Save** link under "New Parameter".

This is the completed "-db" parameter for Fetch Records:

Step 7: Saving the Local Tool

Once all of the necessary inputs, outputs, and parameters have been defined, click on the **Save Tool** button at the bottom of the screen. Please note that all of the "sub forms" opened for each input, output, and parameter must be saved before clicking **Save Tool**. If any of the sub forms are still open, the following error message will be displayed:
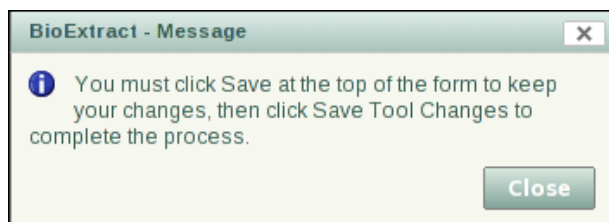
Running a Local Tool

Once the tool is saved, it will appear under the "My Tools" group in the "Available Tools" menu.

When you select the tool, the interface presented is the same one used by all BioExtract Server tools.

A few moments after clicking "Execute", a popup window should appear, followed by another one that looks like this:

This is normal. A Java applet is used to execute the local tool, and applets cannot execute programs or write files without permission, which can only be granted if the applet is placed in a digitally signed .jar file. Since we have signed the file ourselves withou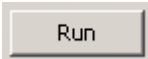t using a certificate from one of the third-party Certificate Authorities, the browser reports that the signature cannot be verified.

Once you click the button, the applet will download the input files from the server, execute the tool, and upload the output files, displaying a short message for each step. Once it is finished, you may close the applet window.

BIOEXTRACT SERVER WORKFLOWS

Introduction

Users do not explicitly create workflows in the BioExtract Server, but implicitly do so by working with the system. As you work in the BioExtract Server, all of your tasks (such as executing queries against selected data sources, applying analytic tools, and saving data extracts) are recorded. At any point, a workflow comprising the performed set of tasks can be saved and subsequently executed as a single unit. Individual tasks within the workflow may also be modified or deleted by the workflow owner.

To illustrate a BioExtract Server workflow, consider the task of carrying out a phylogenetic analysis for a set of proteins where the starting point is a particular genomic coding sequence representing only one member of the gene family in a given species. In the BioExtract Server, the steps for accomplishing this task involve:

- Selecting the *tblastx* tool and providing the accession number of the nucleotide sequence record as input.

- The output from *tblastx*, a BLAST report along with a set of records representing similar sequences, is parsed using a formatting template to produce an initial data extract (i.e. a set of matching nucleotide sequences).

- The resulting data extract is saved.

- The resulting data extract is used as input into the tool *Vmatch* (see http://www.vmatch.de/) to remove duplicate sequences.

- The tool *fetchTranslation* is run. This tool is defined to use the current data extract as input (in GenBank format) and returns the protein translations from the GenBank-annotated coding sequence (CDS) regions (in FASTA format).

- The *ClustalW* tool is run to create the multiple sequence alignment with the input specified as coming from the previously executed tool (*fetchTranslation*) and to define and draw a dendrogram that represents how the sequences are related.

- The *TCoffee* tool is run to create the multiple sequence alignment with the input specified as coming from the previously executed tool (*fetchTranslation*) and to define and draw a dendrogram that represents how the sequences are related.

Creating a Workflow

As an example, let's create the workflow outlined above.

Step 1: Preparing to create the workflow

If you have just signed into the BioExtract Server, it's not necessary to "clear" the workflow memory. But if you have been working with the system and want to begin creating a workflow, it is necessary to clear any previously executed tasks from memory.

To clear previously executed tasks from memory:

1. Click the Workflows tab and select the **Create and Import Workflows** heading in the Workflows tree. The "Create and Import Workflows" form opens in the right panel.

2. In the Create Workflow section, click the **Record Workflow** button. This will erase any previously executed tasks.

Step 2: Execute the *tblastx* tool from the Tools tab

1. Click the Tools tab. In the Available Tools menu expand the ⊟ Similarity Search Tools group and select the *tblastx* tool. The tool's form opens in the right panel.

2. In the Input Data section, specify **Paste or type data into the text area** for the input source and enter the accession number of the nucleotide sequence record as input, "L16896" (without the quotes).

3. Click the Execute button. The output from the execution of the *tblastx* tool is a blast report; furthermore the *tblastx* report is turned into a data extract viewable on the Extracts page.

Step 3: Save the resulting data extract

1. Click the Extracts tab. The tblastx data extract displays.

2. Click the [Save Extract] button. The "Save Extract" dialog box opens.
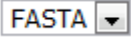
3. Enter a name and description for the data extract and click the [Create Extract] button. The data extract is saved and becomes a searchable data source available on the Query page In Available Data Sources under the heading Miscellaneous.

> **Save Extract** ✕
>
> Please enter the following information:
>
> **Extract Name:**
>
> tblastx results
>
> **Description:**
>
> nucleotide sequence query is translated in
> six reading frames, and the resulting six-
> protein sequences are compared, in turn,
> to those in a protein sequence database.
>
> [Create Extract] [Cancel]

Step 4: Remove duplicate sequences in the data extract using the *xmknr* tool
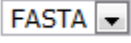
1. Click the Tools tab. In the Available Tools menu expand the ⊟ Edit Tools group and select the *xmknr* tool. The tool's form opens in the right panel.

2. In the Input Data section, specify ◉ Use records on Extracts page formatted as FASTA ▾ for the input source.

3. In the Parameter Settings section, set Sequence type (protein or DNA) (-a) dna ▾ to dna.

4. Click the Execute button. A number of result files are created. In addition, the data extract visible on the Extracts page has been modified to remove any duplicate sequence records.

Step 5: Convert the data extract to GenBank format using the *FormatConversion* tool (as the next tool requires GenBank format to run)

1. Click the Tools tab. In the Available Tools menu expand the ⊟ Edit Tools group and select the *FormatConversion* tool. The tool's form opens in the right panel.

2. In the Input Data section, specify ⦿ Use records on Extracts page formatted as FASTA ▾ for the input source.

3. In the Parameter Settings section, set **To Format** to *genbank* and **From Format** to *fasta*.

4. Click the Execute button. The records in the data extract are now converted to *genbank* format.

Step 6: Retrieve the protein translations from the CDS regions of the DNA sequences using the *FetchTranslation* tool

1. From the Tools page, in the Available Tools menu expand the ⊞ Information Tools group and select the *FetchTranslation* tool. The tool's form opens in the right panel.

2. In the Input Data section, specify **Use previously executed tool results** for the input source and select *FormatConversion* and *result.txt* from the associated drop-down menus.

3. Click the Execute button. The tool *FetchTranslation* runs and returns the protein translations in FASTA format.

Input Data

Input File

○ Use records on Extracts page formatted as FASTA ▾

⦿ Use previously executed tool results 3 FormatConversion ▾

  result.txt ▾

  Select tool
  1 tblastx

○ Upload data saved on your computer 2 xmknr

  3 FormatConversion

Choose File No file chosen

○ Paste or type data into the text area

Step 7: Create a *ClustalW2* multiple sequence alignment and dendrogram with the protein translations retrieved by *FetchTranslation*

1. From the Tools page, in the Available Tools menu expand the ⊞ Alignment Tools group and select the *ClustalW2* tool. The tool's form opens in the right panel.

2. In the Input Data section, specify **Use previously executed tool results** for the input source and select *FetchTranslation* and *fetchTranslation_results.txt* from the associated drop-down menus.

3. Click the Execute button. The tool *ClustalW2* runs and creates a multiple sequence alignment and draws a dendrogram that represents how the sequences are related.

ClustalW2

Multiple Sequence Alignments. **More Information**.

Input Data

```
┌ Input file ─────────────────────────────────────────────┐
│  ○ Use records on Extracts page formatted as  FASTA ▼    │
│  ● Use previously executed tool results  4 FetchTranslation ▼ │
│  fetchTranslation_results.txt ▼                          │
│  ○ Upload data saved on your computer                    │
│  [Choose File]  No file chosen                           │
│  ○ Paste or type data into the text area                 │
│  ┌──────────────────────────────────────────┐           │
│  │                                          │           │
│  │                                          │           │
│  │                                          │           │
│  │                                          │           │
│  └──────────────────────────────────────────┘           │
└──────────────────────────────────────────────────────────┘
```

```
Untitled - Google Chrome                                    _ □ X

bioextract.org/clean/ViewFile.jsp?filename=/opt/biostreamserver/tmpFiles/ClustalW2/1309284560032/tmp02817573426370076.aln

55960937        ----------------------------------------------------
55960943        ------------------------------------------MERPAPLAVLPFSD
123248414       ------------------------------------------MERPAPLAVLPFSD
55960945        ------------------------------------------MERPAPLAVLPFSD
18027854        ----------------------------------------------------

1177229.28.1    VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
296206596       VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
15488886        VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
55960942        VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
1177229         VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
13529614        VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQRIYGDGTGGSVV
159572780       VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQRIYGDGTGGSVV
60552423        VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSRFFQRIYGDGSGGSVV
55960940        VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
55960941        VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
189303569       VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSRFFQRIYGDGSGGSVV
126366001       VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQRIYGDGTGGSVV
159572779       VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQRIYGDGTGGSVV
108996288       VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
73956793        IQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
301776961       IQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
300794654       VQHSVRVLQELNKQRERGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
335290456       VQHSVRVLQELNKQRERGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
149695571       VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGSGGSVV
291399608       VQHSVRVLQELNKQREKGQYCDATLD-VGGLVFKAHWSVLACCSHFFQSLYGDGPGGSVV
```

Step 8: Create a *TCoffee* multiple sequence alignment and dendrogram with the protein translations retrieved by *FetchTranslation*

1. From the Tools page, select the *TCoffee* tool in the ⊞ Alignment Tools group. The tool's form opens in the right panel.

2. In the Input Data section, specify **Use previously executed tool results** for the input source and select *FetchTranslation* and *fetchTranslation_results.txt* from the associated drop-down menus.

3. Click the Execute button. The tool *TCoffee* runs and creates a multiple sequence alignment and draws a dendrogram that represents how the sequences are related.

TCoffee

Computes a multiple sequence alignment and the associated phylogenetic tree for Protein, RNA and DNA sequences. **More Information**.
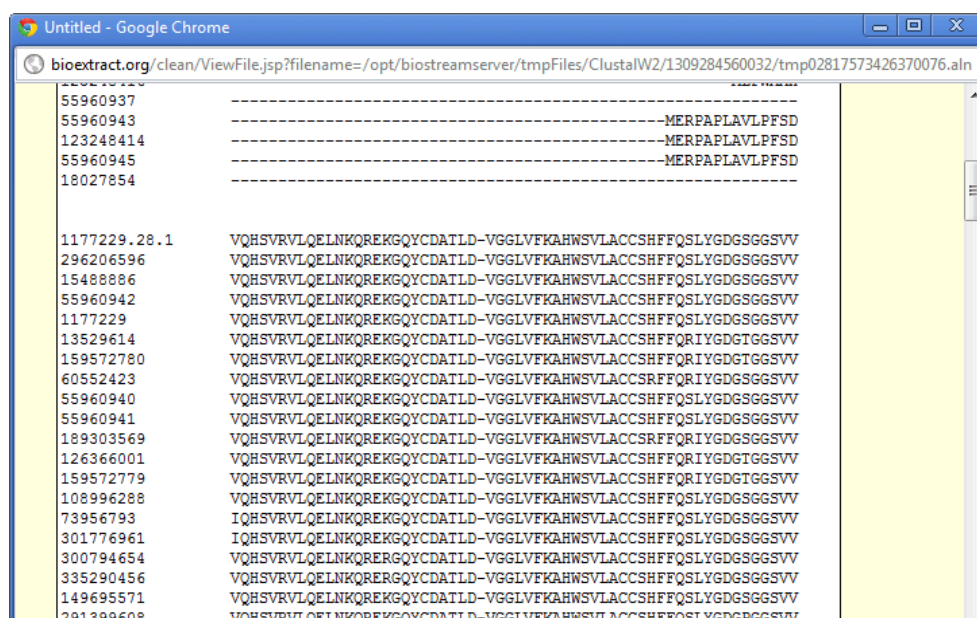
Input Data

Input sequences (FASTA format)
- ○ Use records on Extracts page formatted as FASTA ▾
- ◉ Use previously executed tool results 4 FetchTranslation ▾
  fetchTranslation_results.txt ▾
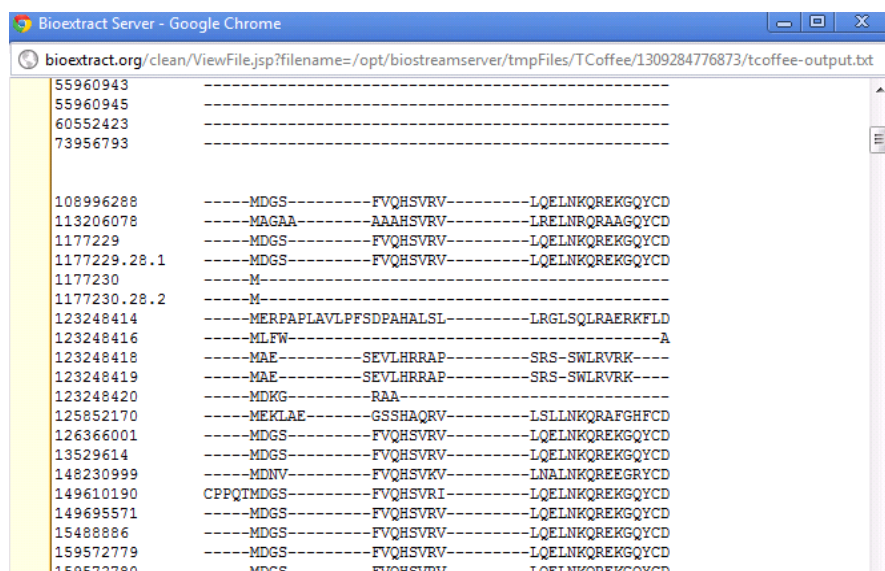- ○ Upload data saved on your computer
 Choose File No file chosen
- ○ Paste or type data into the text area

Bioextract Server - Google Chrome

bioextract.org/clean/ViewFile.jsp?filename=/opt/biostreamserver/tmpFiles/TCoffee/1309284776873/tcoffee-output.txt

```
55960943          ------------------------------------------------
55960945          ------------------------------------------------
60552423          ------------------------------------------------
73956793          ------------------------------------------------


108996288         -----MDGS---------FVQHSVRV---------LQELNKQREKGQYCD
113206078         -----MAGAA--------AAAHSVRV---------LRELNRQRAAGQYCD
1177229           -----MDGS---------FVQHSVRV---------LQELNKQREKGQYCD
1177229.28.1      -----MDGS---------FVQHSVRV---------LQELNKQREKGQYCD
1177230           -----M------------------------------------------
1177230.28.2      -----M------------------------------------------
123248414         -----MERPAPLAVLPFSDPAHALSL---------LRGLSQLRAERKFLD
123248416         -----MLFW----------------------------------------A
123248418         -----MAE---------SEVLHRRAP---------SRS-SWLRVRK----
123248419         -----MAE---------SEVLHRRAP---------SRS-SWLRVRK----
123248420         -----MDKG---------RAA----------------------------
125852170         -----MEKLAE-------GSSHAQRV---------LSLLNKQRAFGHFCD
126366001         -----MDGS---------FVQHSVRV---------LQELNKQREKGQYCD
13529614          -----MDGS---------FVQHSVRV---------LQELNKQREKGQYCD
148230999         -----MDNV---------FVQHSVKV---------LNALNKQREEGRYCD
149610190         CPPQTMDGS---------FVQHSVRI---------LQELNKQREKGQYCD
149695571         -----MDGS---------FVQHSVRV---------LQELNKQREKGQYCD
15488886          -----MDGS---------FVQHSVRV---------LQELNKQREKGQYCD
159572779         -----MDGS---------FVQHSVRV---------LQELNKQREKGQYCD
159572780         -----MDGS---------FVQHSVRV---------LQELNKQREKGQYCD
```
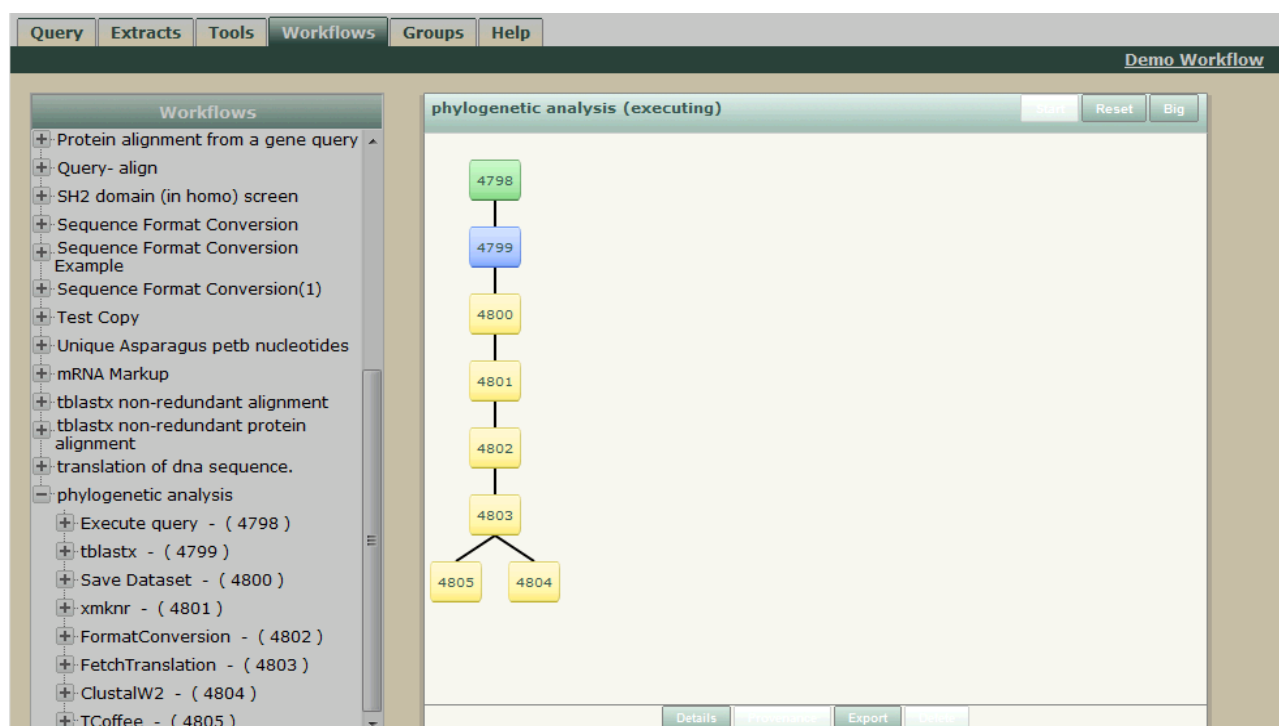

Step 9: Save the workflow

1. Click the Workflows tab and select the ┊┈Create and Import Workflows heading in the Workflows tree. The "Create and Import Workflows" form opens in the right panel.

2. In the Save Workflow section, enter a name and description for the new workflow, and then click the **Save** button. The workflow is saved and displays in the Workflows menu.
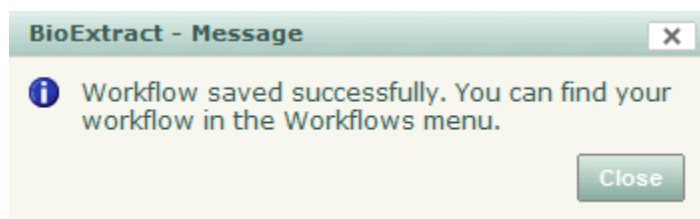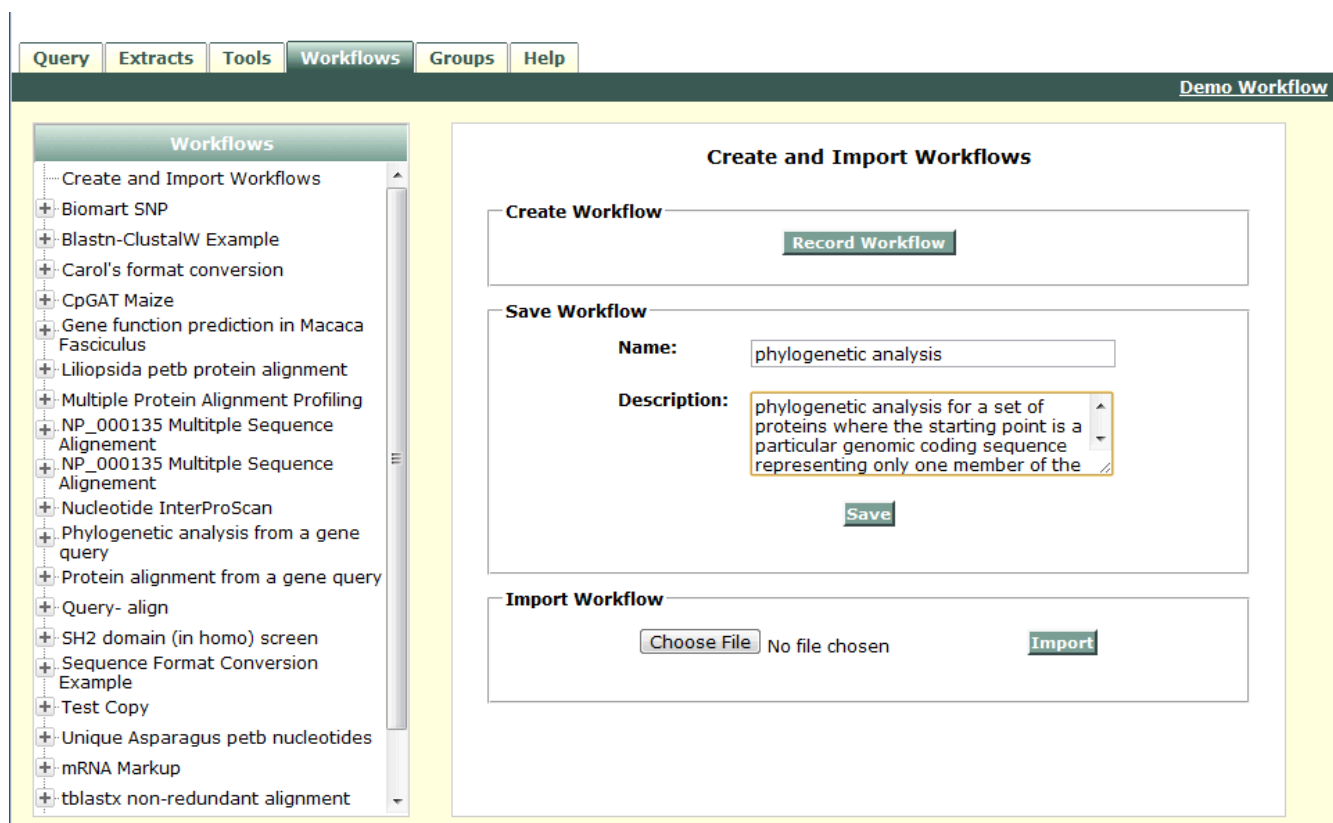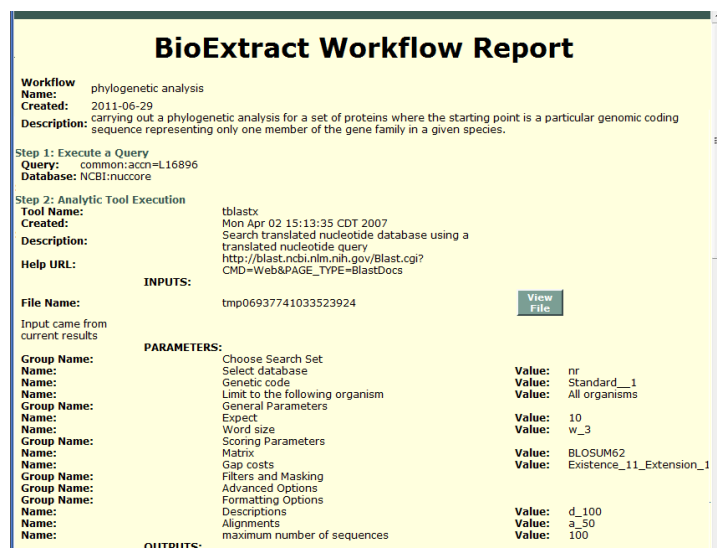
Executing a Workflow

To execute a workflow, select the desired workflow from the Workflows menu on the Workflows page. A graphical representation of the workflow displays in the right panel.

Click the [Start] button at the top of the workflow panel. When a step within the workflow begins to run, it will turn blue. When the execution completes, the color of the step will change to green.

When execution completes, the **Provenance** button becomes enabled. Click the **Provenance** button to see the provenance information associated with the execution of the workflow.

Modifying a Workflow

One of the major advantages of BioExtract Server workflows is they can be modified and executed using different data and parameter settings.

Warning: Workflow modifications are saved permanently if you are the owner of the workflow. Otherwise, they are saved temporarily. Before you modify a workflow, you may want to make a copy, so you don't overwrite the original workflow. See "Copying a Workflow" for more information.

As an example, assume the original workflow contained a query step using accession number "L16896" and you're interested in running the same workflow using accession number "NM_005341". You would also like to change some of the *tblastx* tool parameter settings.

To modify a query and tool step in a workflow:

1. From the Workflows page, select the workflow you want to modify (i.e. ![phylogenetic analysis]). This will show you all of the steps included in that workflow. Don't forget to sign in to see your workflows.

If you want to change the query request, click the plus sign to expand the "Execute query" step. The step will expand and display modifiable sub-steps (Query and Databases). Select the Query heading. A Query Information form opens in the right panel. You can change the search field and the search term. Change the search term "L16896" to NM_005341 and click the Save button to keep your changes. Table 1 displays valid search terms that may be used when modifying a query.

Query Information

Properties

Query: common:accn=L16896

Save   Reset

Query examples.

Query Information

Properties

Query: common:accn=NM_005341

Save   Reset

Query examples.

Field	Definition	Search Term	Search Term Examples
Accession	Contains the unique accession number of the sequence or record	common:accn	common:accn=NM_005341
All Text	Contains all terms from all searchable database fields in the database.	common:all	common:all='BTB domain'
Author	Contains all authors from all references in the database records	common:author	common:author=Zhang
Definition	Includes only those words found in the definition line of a record.	common:defn	common:defn='Homo sapiens' AND common:defn='BTB domain' AND common:defn=' zinc finger '
Feature Key	Contains the biological features assigned or annotated to the nucleotide sequences and defined in the DDBJ/EMBL/GenBank	common:fkey	common:fkey=gene

	Feature Table		
Gene	Contains the standard and common names of genes found in the database records	common:gene	common:gene=Zbtb16
Id		common:id	common:id=157694498
Keywords	Contains special index terms from the controlled vocabularies associated with the GenBank, EMBL, DDBJ, SWISS-Prot, PIR, PRF, or PDB databases.	common:keyword	common:keyword= Zbtb8
Species		common:species	common:species= 'Camphor tree'
Taxonomy	Contains the scientific and common names for the organisms associated with protein and nucleotide sequences.	common:taxonomy	common:taxonomy='Mus musculus'
Title	Title of the journal abbreviation	Common: title	Common: title=" Plant Physiol"

Table 1. List of valid search fields and example search terms.

If you want to change *tblastx* tool parameters, click the plus sign to expand the "tblastx" tool step. The step will expand and display the modifiable sub-step Tool. Select the Tool heading. The tool form opens in the right panel. Change parameters according to your preferences. Input into a tool may be modified if the original data was entered in the text box or uploaded as a file. Click the Save button to keep your changes.

Copying a Workflow

To copy a workflow:

1. Click the Create and Import Workflows heading on the Workflows page. The "Create and Import Workflows" form opens in the right panel.

2. In the Create Workflow section, click the Record Workflow button.

3. Next, open the workflow you want to make a copy of and run it. The BioExtract Server will record the running workflow in the background.

4. Once execution is complete, click the Create and Import Workflows heading on the Workflows page. The "Create and Import Workflows" form opens in the right panel.

5. In the Save Workflow section, enter a new Name and keep or change the Description, then click the Save button. The name of the new "copied" workflow will appear in the Workflows menu. You can now modify the original workflow or the copy.

BIOEXTRACT SERVER GROUPS

About Groups

Groups provide a collaborative environment to facilitate the sharing of data extracts, analytic tools and workflows.

Registered users can create new groups by using the Create Group option on the Groups page. When you create a group, you'll be able to invite others to join. Plus you'll be able to share your tools, workflows and data extracts with others. In addition, you'll be able to modify your shared workflows, and remove tools, data extracts and workflows from group access. Groups you own are listed under Owned Groups.

Groups of which you are a member are listed under Member Groups. As a group member, you can view and execute any of the tools, workflows and data extracts that you've been given permission to use.

To use group tools, sign in and click the Tools tab. In Available Tools, expand **My Tools**. By default, group tools are listed along with your private tools. Locate and select the group tool you want to use. The tool form will open in the right panel.

To use group workflows, sign in and click the Workflows tab. By default, all workflows (public, private and group) are given in one list. Locate and select the group workflow you want to use. The workflow graph and its control buttons will open in the right panel.

Group data extracts can be found on the Query page in Available Data Sources under the Miscellaneous heading.

Creating a Group

The BioExtract Server provides a facility to create groups of registered users. Groups allow registered users to share tools, workflows and data extracts.

To create a group:

1. You'll first need to create a user account and sign in. At the top right corner of any BioExtract Server page, select **register**. Fill in the required fields and click **Add User**. Your account will be created immediately. Return to the BioExtract Server and sign in by clicking the **sign in** link at the top right corner of any page.

2. Click the **Groups** tab.

3. In the left panel, under the **Additional Actions** heading, click **Create Group**. A new group form opens in the right panel.

4. Click **Edit** to the right of the heading "A New Group." Enter a Name and Description for this new group, then click the **Save** button. A new group is created. You'll see your new group under the **Owned Groups** heading in the left panel.

The "Name and Description" of the newly created group can be edited by selecting the group and editing the information. Once the information has been edited, click the **Save** link at the top of the panel to keep your changes.

Adding Tools, Workflows and Data Extracts to a Group

After you create a group, you can add elements, such as analytic tools, data extracts and workflows to your group.
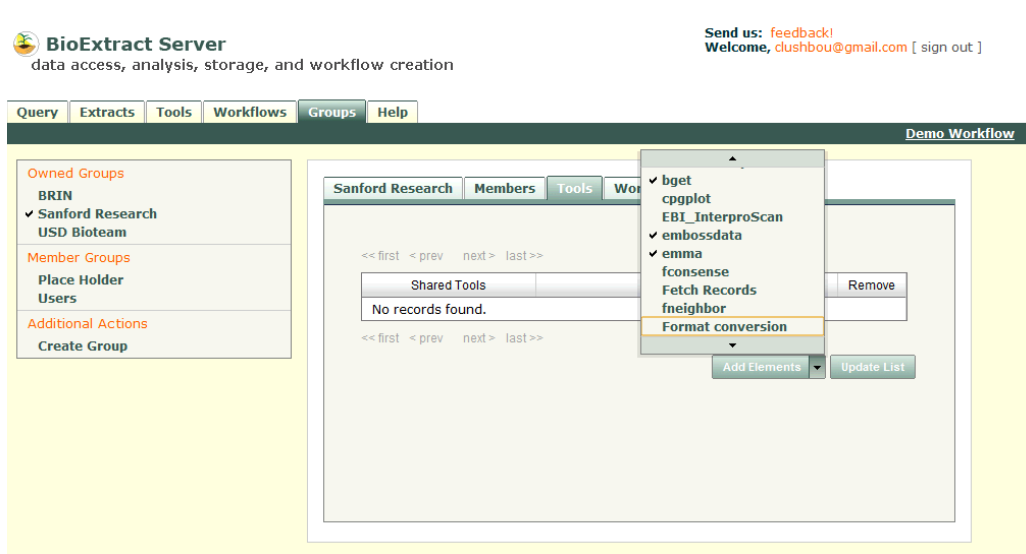
Note: Before adding tools, data extracts and workflows to your group, you must first add them to your private BioExtract Server account.

To add tools, data extracts and workflows to a group:

1. Sign in to the BioExtract Server and click the Groups tab.

2. In the left panel, under the **Owned Groups** heading, select the name of the group that you want to add an element to. The group form opens in the right panel.

3. Click the Tools, Workflows or Extracts tab depending on what you'd like to add. The Add Elements form displays.

4. Click the black arrow on the **Add Elements** button. A list of tools, workflows and data extracts appears. Select the name of the element that you'd like to add. Choose as many elements as you like, then click the **Add Elements** button. The selected element will be added to the form.

Removing Tools, Workflows and Data Extracts from a Group

If you're the owner of a group, you can remove tools, data extracts and workflows from that group.

To remove tools, data extracts and workflows from a group:

1. Sign in to the BioExtract Server and click the Groups tab.

2. In the left panel, under the **Owned Groups** heading, select the name of the group you want to remove an element from. The group form opens in the right panel.

3. Click the Tools, Workflows or Extracts tab depending on what you'd like to remove. The Add Elements form displays.
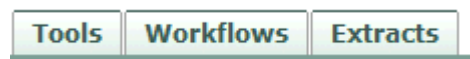
4. In the Remove column, select the check box for the element you'd like to remove. Choose as many elements as you like. Then click the **Update List** button. The selected element will be removed from that group.

Inviting Members to Join a Group

To invite a new user to a group, the group owner sends an invitation to that user. Once the recipient accepts the invitation, they will have access to the tools, workflows and data extracts owned by the group.

To invite other people to your group:

1. Sign in to the BioExtract Server and click the Groups tab.

2. In the left panel, under the **Owned Groups** heading, select the desired group name. The group form opens in the right panel.

3. Click the **Members** tab. The Invite Members form displays.

4. Click the **Invite Members** button. An Invite New Members dialog box appears.

5. Select "Please enter a recipient's email address". A text box will appear.

6. Enter the email address of whomever you want to invite to the group, click the **Save** button, then click the **Send Invitation** button to complete the process. A message appears stating the invitation was sent.

When you invite people to join a group, we will immediately send email invitations to the addresses you provided. Once the recipient accepts the invitation, they will have access to the tools, data extracts and workflows owned by that group.
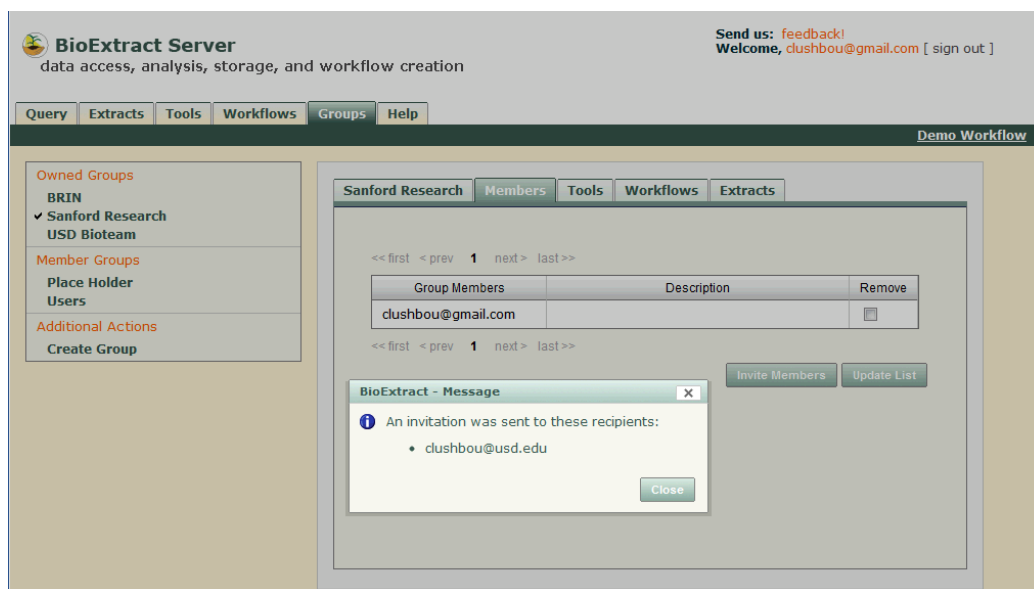
To accept an invitation to join a group:

1. If you don't have a BioExtract Server account, you'll need to create one before accepting the invitation. Open bioextract.org. On the top right corner of any BioExtract Server page, select **register**. Fill in the required fields and click **Add User**. Your account will be created immediately.

2. Return to the invitation email. The invitation email tells you that a BioExtract Server group wants to add you as a member. It gives you the name of the group that you've been invited to join and gives you a link to accept the invitation. Take note of the group's name for future reference.

3. Click the **Invitation** link. A BioExtract Server login window opens. Enter your BioExtract Server user name and password.

4. Click the **Accept Invitation** button. A window displays "Thank you for accepting the invitation."
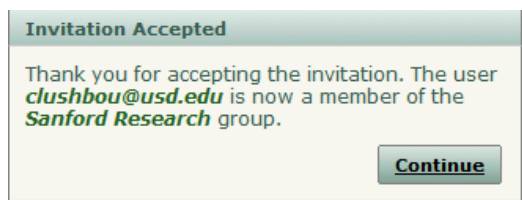
5. Click the **Continue** button. The BioExtract Server web site opens. Sign in by clicking the sign in link on the top right corner of any BioExtract Server page.

6. Click the **Groups** tab. Under Member Groups you'll see the name of the group you joined. Select the **group name**. The group interface opens in the right panel.

7. Click the **Members** tab. In the Group Members list you'll see your email address. You now have access to the tools, data extracts and workflows owned by that group.

WORKFLOWS THROUGH MYEXPERIMENT

About myExperiment

my experiment (http://www.myexperiment.org) is a collaborative environment where scientists can publish their workflows and experiment plans, share them with groups and find those of others. Workflows, other digital objects and bundles (called Packs) can be swapped, sorted and searched like photos and videos on the Web.

Importing a BioExtract Server Workflow into myExperiment

To import a BioExtract Server Workflow into myExperiment:

1. Click the **Workflows** tab and select the desired workflow. The workflow graph and its control buttons opens in the right panel.

2. Click the **Export** button. The Open dialog box opens. Save the file to your computer desktop.

Version 1.0

3. Open the myExperiment website (http://www.myexperiment.org).

4. Click the Workflows tab. Click the **GO** button in the upper right corner under the **New/Upload** heading.
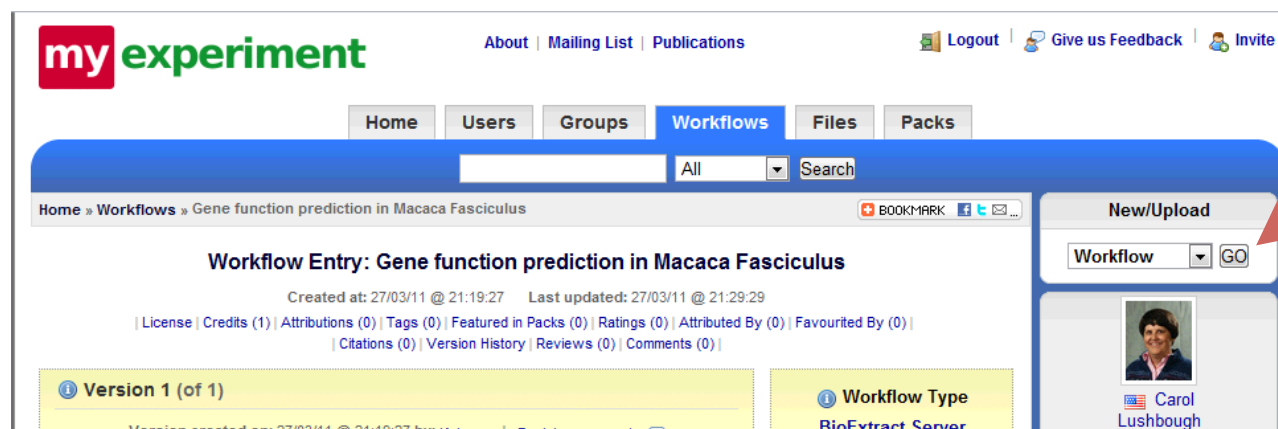
5. Click the **Choose File** button and select the file representing the workflow you exported from the BioExtract Server.

6. Click the Upload and Continue button at the bottom of the screen.

7. Select the desired tags and click the [Finish] button. The workflow has been uploaded to myExperiment.