**How to Create a Workflow within the BioExtract Server at** [**bioextract.org**](http://bioextract.org/)

Carol Lushbough

University of South Dakota

Carol.Lushbough@usd.edu

This tutorial will demonstrate how to perform a phylogenetic analysis on a set of proteins where the starting point is a query for specific nucleotide gene sequences. The necessary steps for creating this BioExtract Server workflow example involve first querying the NCBI Nucleotide Core data source (<http://www.ncbi.nlm.nih.gov>) for the *liliopsida chloroplast* gene *petB*. Next, the data extract resulting from this query is used as input into *Vmatch* (see <http://www.vmatch.de/>) to remove duplicate sequences. The resulting unique records are converted into GenBank format and the *fetchTranslation* tool is invoked. The execution of the fetch translation tool returns the protein translations from the GenBank-annotated coding sequence (CDS) regions (in FASTA format). Finally, the *ClustalW* ([http://www.clustal.org](http://www.clustal.org/)/) tool is executed to create the multiple sequence alignment with the input specified as coming from the previously executed tool (i.e., the extracted protein sequences) and to define and draw a dendrogram that represents how the sequences are related.

1. **Signing in as a Registered User**

In order to save the workflow, sign in as a registered user before continuing with the tutorial.



1. **Prepare for Recording**

BioExtract Server workflows are created by “recording” in the background steps executed by users as they work with the system. When first signing into the BioExtract Server, the recording area is empty. But, if there have been previously executed steps that are not meant to be part of the workflow being created, click on the ***Create and Import Workflows*** node on the Workflow tab. Then click the button.



Clear workflow recording area

1. **Executing the Query**
2. Click the “Query” tab
3. Select the *NCBI Core Nucleotide* data source from the list of available data sources.
4. Select the *Organism* search filed and enter ***liliopsida***
5. Click the  three times
6. Select the *Definition* search field and enter the term ***chloroplast***
7. Select the *Definition* search field and enter the term ***petB***
8. Select the *Definition* search field and enter the term ***gene***
9. Click the ** button
10. Click the “Extracts” tab to view the results. 
11. **Remove the Duplicate Records**
12. Click the “Tools” tab
13. Select the ***xmknr*** tool under the *Edit Tools* node in the Available tools tree
14. Select the *Use records on Extracts page formatted as FASTA* as the Input Data (Note: the input into the xmknr tool will be the records listed on the Extracts tab.)
15. Select ***dna*** for the *Sequence Type* parameter.
16. Click the  button at the bottom of the screen. (Note: this tool will modify the list of records listed on the “Extract” tab.)

**

1. **Retrieve Records in GenBank format**
2. Select the ***Format Conversion*** tool under the *Edit Tools* node in the Available Tools tree
3. Select the *Use records on Extracts page formatted as FASTA* as the Input Data (Note: the input into the format conversion tool will be the records listed on the Extracts tab.)
4. Set the *To Format* parameter to ***genbank***
5. Set the *From Format* parameter to ***fasta***
6. Click the **button.
7. **Retrieve Protein Translations from the CDS regions**
8. Select the ***Fetch Translations*** tool under the *Information tools* node in the Available Tools tree.
9. Select the *Use* *previously executed tool results* as the Input Data and select ***FormatConversion*** and ***result.txt***.
10. Click the **button.



1. **Create a Multiple Sequence Alignment and Dendrogram**
2. Select ***ClustalW*** under the *Alignment Tools* node in the Available Tools tree
3. Select the *Use* *previously executed tool results* as the Input Data and select ***Fetch Translations*** and ***fetchTranslations\_results.txt.***
4. Click the **button.
5. **Saving the Workflow**
6. Select the “Workflow” tab
7. Enter ***Liliopsida petb protein alignment*** in the Name field.
8. Enter ***Liliopsida petb protein alignment*** in the Description field.
9. Click the **button.



1. **Testing the Workflow**
2. On the “Workflows” tab, click the new workflow's name. The workflow opens in the right panel.
3. Click on the button. The workflow begins to run.
4. After a step has completed execution, it will turn blue. At that point the output can be viewed.



Click to view output

