nrseq-fasta-importer Documentation

Release 4

Michael Riffle

September 29, 2016

Contents

To learn more about the innards and working of the associated databases and software, see the About nrseq-fasta-importer page. To learn about how to install the nrseq-fasta-importer application, see the Installing nrseq-fasta-importer page. And, to learn more about how to use the nrseq-fasta-importer web application, see the Using nrseq-fasta-importer page.

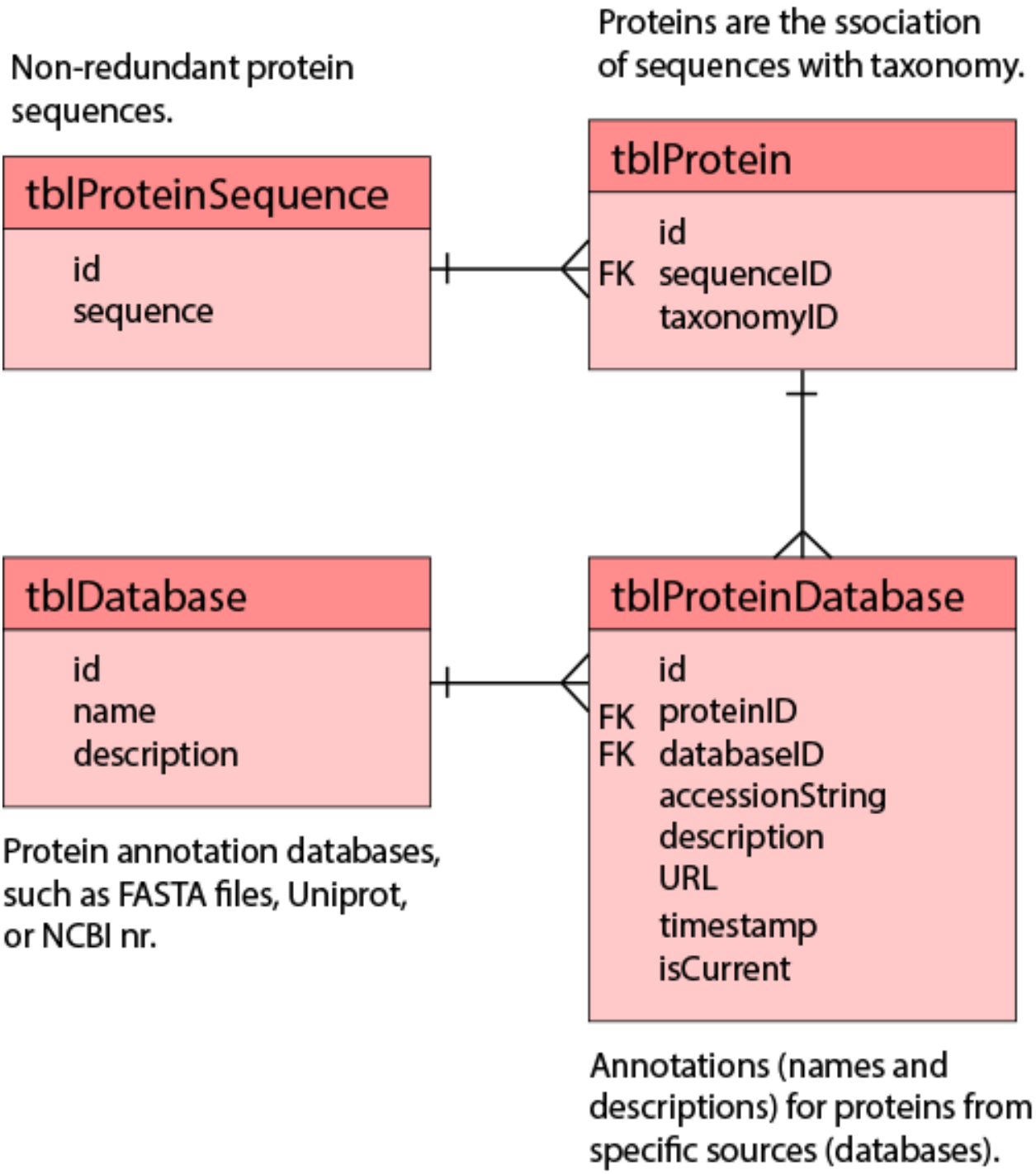
CHAPTER 1

Table of Contents

1.1 About nrseq-fasta-importer

1.1.1 The YRC_NRSEQ database

The purpose of this application is to parse a FASTA file into the YRC_NRSEQ database. The YRC_NRSEQ database stores non-redundant protein sequences, proteins that reference those sequences (a protein being a given sequence in a particular species), annotations that reference those proteins (names and descriptions from different sources), and annotations for the sources (e.g., NCBI, Uniprot, or individual FASTA files). This is accomplished using four simple tables:



Proteins are regarded simply as a specific protein sequence from a specific organism. This allows for unambiguously determining if two different experiments that may use different naming conventions are referring to the same protein– allowing for the ability to compare and contrast that is independent of the naming convention (e.g., FASTA file) used in the respective experiment. It also allows for knowing the organism associated with a protein in an experiment, and for showing the names or descriptions for those proteins from any parsed naming database.

1.1.2 Why do we have to parse FASTA files?

In order to reap the benefits of the YRC_NRSEQ database, a FASTA file must be parsed and inserted into the database. A FASTA file has this basic structure:

```
>NAME1 Descriptive text
SEQUENCE
>NAME2 Descriptive text
SEQUENCE
>NAME3 Descriptive text
SEQUENCE
```

In a proteomics experiment, proteins are typically referenced by the names associated with protein sequences in the FASTA file (NAME1, NAME2, and NAME3 in the above example). By parsing the FASTA file in advance, an application built to interface with YRC_NRSEQ knows which protein sequence and organism is being referenced by that name. When combining results from multiple experiments, the application will know if two names are describing the same protein. When showing results, the name from that FASTA file or from a preferred naming database (that has also been parsed into the YRC_NRSEQ) may be shown. And other protein annotations, such as domain predictions, that are associated with YRC_NRSEQ protein IDs may also be shown along with experimental data to provide biological context.

1.1.3 How does parsing work?

In general terms, parsing works as follows:

1. **Validate the FASTA file**
    - Each FASTA name only appears once
    - Each FASTA header has a sequence
    - This FASTA file hasn't already been parsed
2. **Confirm a NCBI taxonomy ID can be determined for each FASTA header**
    - If this isn't possible, import should not proceed
    - User input may be required
    - This is the trickiest part of the import process and described in more detail below
3. **For each FASTA entry (header + sequence) in the FASTA file:**
    (a) Insert an entry for this FASTA file in tblDatabase, get ID
    (b) **Get the sequence ID for the current sequence**
        - If not in the database, insert it and get the resulting ID
    (c) Determine the NCBI taxonomy ID
    (d) **Determine the protein ID (from sequence ID and taxonomy ID)**
        - If not in the database, insert it and get the resulting ID
    (e) Insert the annotation for this protein from this FASTA file in tblProteinDatabase

1.1.4 How does the taxonomy lookup work?

Software, such as the nrseq-fasta-importer web application, may employ any logic to determine the NCBI taxonomy ID that should be associated with a given FASTA header. The logic employed by the nrseq-fasta-importer works roughly as follows:

1. Does the FASTA header look like a reversed or shuffled sequence? Use 0 for the taxonomy ID.
2. Does the FASTA header contain `Tax_Id=###` or `Taxonomy_Id=###`? Use ###.
3. Does the FASTA header end with `[species name]`? Lookup the taxonomy ID for that species name.
4. Does the FASTA header contain `organism=species_name`? Lookup the taxonomy ID for that species name.
5. Does the FASTA header contain `OS=speces name`? Lookup the taxonomy ID for that species name.
6. Does the FASTA header name look like a NCBI accession string? If so, lookup the taxonomy ID from NCBI.
7. Does the FASTA header name look like a swiss-prot accession string? If so, lookup the taxonomy ID from swiss-prot.
8. Does the FASTA header name look like a Uniprot accession string? If so, lookup the taxonomy ID from Uniprot.
9. Does the FASTA header name look like a SGD accession string? If so, lookup the taxonomy ID from SGD.
10. Does the FASTA header name look like a Wormbase accession string? If so, lookup the taxonomy ID from Wormbase.
11. Does the FASTA header name look like a Flybase accession string? If so, lookup the taxonomy ID from Flybase.

1.1.5 What if the taxonomy ID cannot be found?

The nrseq-fasta-importer web application provides an interface for providing NCBI taxonomy IDs for any FASTA headers for which it could not be determined. (See Using nrseq-fasta-importer.) If the taxonomy ID cannot be determined for a very large number of entries, several strategies are available:

1. Write a script to modify the FASTA file and add `Tax_Id=XXX` to each FASTA header, where XXX is the NCBI taxonomy ID. This is often a good choice if your FASTA file is of your own making and the headers require custom taxonomy lookup logic.
2. Write a new taxonomy lookup module for the nrseq-fasta-importer web application. Although harder, this is often a good choice if the FASTA headers include names or description from a protein naming database you often use or encounter. The authors of nrseq-fasta-importer are happy to work with you to add a new lookup module.
3. Write your own parser script. Although not generally recommended, importing into the YRC_NRSEQ database is not conceptually complex and may be the right choice if you are an advanced user and are comfortable with scripting.

1.2 Installing nrseq-fasta-importer

1.2.1 1. Install MySQL, Java, and Apache Tomcat (if necessary)

This documentation assumes that Java (JDK version, 1.7 or later) and the Apache Tomcat (7 or later) servlet container are installed on the same computer on which you are installing ProXL. (Note: Apache Tomcat requires the JDK version of Java be installed.)

This documentation also assumes that MySQL (5.6 or later) has been installed and is accessible by the installation of Apache Tomcat. This does not need to be on the same machine as Apache Tomcat.

These software (and nrseq-fasta-importer) should work equally well on any operating system for which MySQL and Java are available (MS Windows, Apple OS X, or Linux). Other servlet containers and database server software may work as well, though this documentation assumes that the above are installed. Please refer to the respective websites for more information about MySQL, Java, or Apache Tomcat installation.

You may need to download and install the MySQL JDBC driver. This is available from the MySQL Connector/J website. To install, copy the downloaded jar file into `$CATALINA_HOME/lib` directory on the server on which Apache Tomcat is installed (e.g. /usr/local/apache-tomcat-7.0.65/lib).

1.2.2 2. Set up the YRC_NRSEQ database

To set up this database, first download `YRC_NRSEQ_create.sql`. To run this SQL script, you may log into your MySQL server and either paste in the contents of this file to MySQL or use `source /location/to/YRC_NRSEQ_create.sql`. To use the latter, the .sql file must be in a directory to which MySQL has read access.

1.2.3 3. Set up the nrseq_fasta_importer database

To set up this database, first download `database_schema_create.sql`. To run this SQL script, you may log into your MySQL server and either paste in the contents of this file to MySQL or (preferably) use `source /location/to/database_schema_create.sql`. To use the latter, the .sql file must be in a directory to which MySQL has read access.

Update configuration table

Download and open `config_inserts.sql` in a text editor. Edit the SQL statements to reflect your configuration options–see the comments in the file for information about the options. Once finished either:

1. **Paste the file into MySQL**
    (a) Copy the contents of the file
    (b) Type `USE nrseq_fasta_importer;` in MySQL
    (c) Paste contents of the file.
2. Save the file and `source /location/to/config_inserts.sql`.

1.2.4 4. Configure Apache Tomcat database connection

This section describes how to connect Tomcat to the YRC_NRSEQ and nrseq_fasta_importer databases.

On the MySQL side:

Log in to MySQL as root:

```
shell> mysql --user=root mysql
```

Create the MySQL user:

```
mysql> CREATE USER 'nrseq_user'@'localhost' IDENTIFIED BY 'password';
```

Replace `nrseq_user` with the username you would prefer, `localhost` with the relative hostname of the machine connecting to the MySQL database (usually localhost), and `password` with your preferred password.

Grant the necessary privileges in MySQL:

```
GRANT ALL ON YRC_NRSEQ.* TO 'nrseq_user'@'localhost'
GRANT ALL ON nrseq_fasta_importer.* TO 'nrseq_user'@'localhost'
```

Replace `nrseq_user` and `localhost` with the username and hostname you used when creating the user.

On the Tomcat side:

Add the following to `$CATALINA_HOME/conf/context.xml`, inside the `<Context></Context>` root element. Be sure to change `nrseq_user` and `password` to the username and password you set up above. If necessary, change `localhost` and `3306` to the hostname and port of your MySQL server.

```
<Resource      name="jdbc/nrseq_fasta_importer"
               auth="Container"
               type="javax.sql.DataSource"
               factory="org.apache.commons.dbcp.BasicDataSourceFactory"
               maxActive="50"
               maxIdle="1"
               maxWait="10000"

                                minIdle="0"
                                minEvictableIdleTimeMillis="21600000"
                                timeBetweenEvictionRunsMillis="30000"
                                validationQuery="select 1 from dual"
                                testOnBorrow="true"









               username="nrseq_user"
               password="password"
               driverClassName="com.mysql.jdbc.Driver"
               url="jdbc:mysql://localhost:3306/nrseq_fasta_importer?autoReconnect=true&amp;tcpKe

<Resource      name="jdbc/nrseq"
               auth="Container"
               type="javax.sql.DataSource"
               factory="org.apache.tomcat.dbcp.dbcp.BasicDataSourceFactory"
               maxActive="10"
               maxIdle="1"
               maxWait="10000"

                                minIdle="0"
                                minEvictableIdleTimeMillis="21600000"
                                timeBetweenEvictionRunsMillis="30000"
```

```
                                        validationQuery="select 1 from dual"
                                        testOnBorrow="true"



                username="nrseq_user"
                password="password"
                driverClassName="com.mysql.jdbc.Driver"
                url="jdbc:mysql://localhost:3306/YRC_NRSEQ?autoReconnect=true"/>
```

1.2.5 5. Install WAR file into Apache Tomcat

Download the latest release of nrseq-fasta-importer from github at https://github.com/yeastrc/nrseq-fasta-importer/releases

Unzip the downloaded file and copy nrseq-fasta-importer.war into `$CATALINA_HOME/webapps/`. The WAR file should automatically deploy. If not, restart Tomcat to force the file to deploy.

Your web application should now be available at http://your.host:8080/nrseq-fasta-importer/ (Depending on how you have configured your web server, the `:8080` may not be different or not required.) If you have a firewall running, will need to allow access through this port.

1.2.6 6. Configure security for Apache Tomcat (optional but recommended)

To prevent unauthorized access to your nrseq-fasta-importer web application, it is recommended that you set up user authentication. These instructions describe how to set up basic authentication for Tomcat. For more detailed instructions, see https://tomcat.apache.org/tomcat-7.0-doc/realm-howto.html

First, add the following lines within the `<tomcat-users></tomcat-users>` element in `$CATALINA_HOME/conf/tomcat-users.xml`. Substitute `USERNAME` and `PASSWORD` with the username and password you wish to use to secure access to your web application.

```
<role rolename="nrseq-fasta-upload-group"/>
<user username="USERNAME" password="PASSWORD" roles="nrseq-fasta-upload-group"/>
```

Second, add the following lines within the `<web-app></web-app>` root element in `$CATALINA_HOME/webapps/nrseq-fasta-importer/WEB-INF/web.xml` (the web.xml for your deployed nrseq-fasta-importer web app).

```
<security-constraint>
        <web-resource-collection>
                <web-resource-name>NRSEQ FASTA Upload Server</web-resource-name>
                <url-pattern>/*</url-pattern>
        </web-resource-collection>
        <auth-constraint>
                <description>Authorized NRSEQ FASTA Upload User</description>
                <role-name>nrseq-fasta-upload-group</role-name>
        </auth-constraint>
</security-constraint>
<security-role>
        <role-name>nrseq-fasta-upload-group</role-name>
</security-role>
<login-config>
        <auth-method>BASIC</auth-method>
        <realm-name>nrseq-fasta-upload-server</realm-name>
</login-config>
```

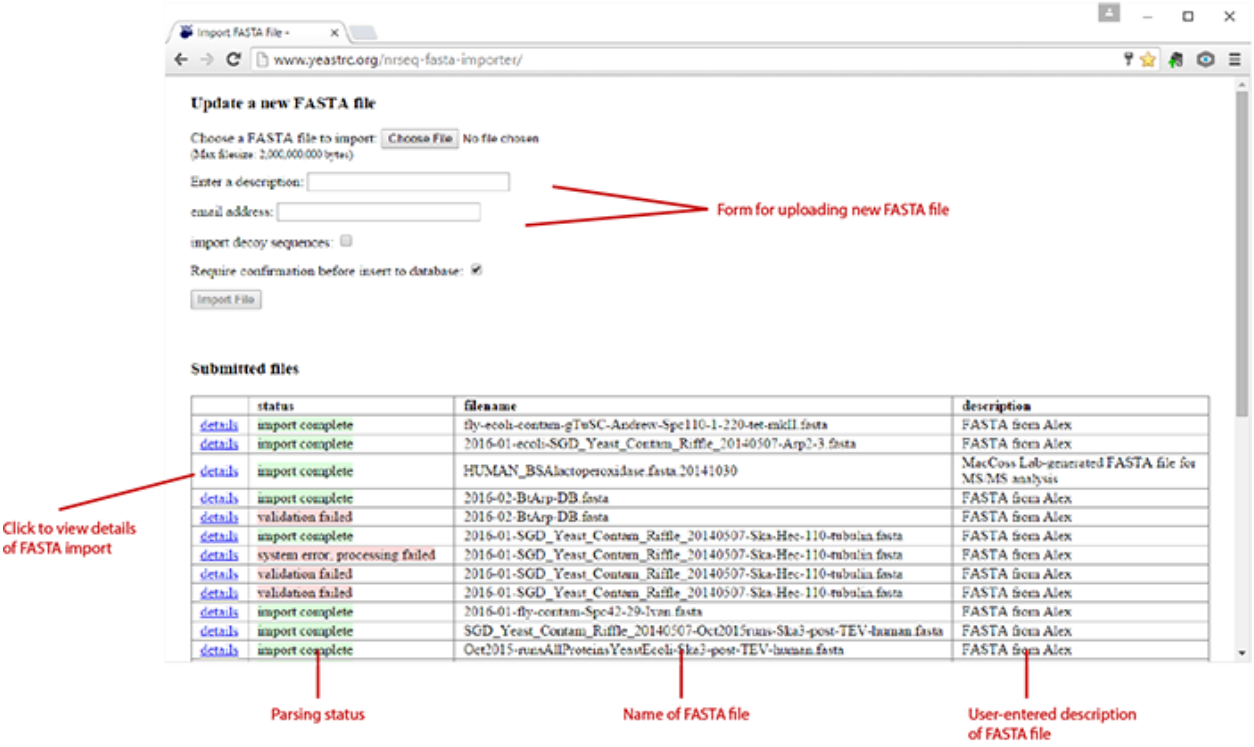
You will need to restart Tomcat for these changes to take effect.

1.3 Using nrseq-fasta-importer

1.3.1 Overview

The nrseq-fasta-importer web application provides a dynamic GUI to parsing FASTA files and importing them into the YRC_NRSEQ database. For more information about what this means, see About nrseq-fasta-importer.

The following screenshot shows the interface:



1.3.2 Upload the FASTA file

To upload a FASTA file, fill out the form at the top of the page. First click the `Choose File` button and select the FASTA file. Enter a description for the FASTA file, your email address (for status notifications), whether or not to import decoy sequences, and whether or not to require user confirmation before data is inserted into the YRC_NRSEQ database (recommended).

Once the form is submitted, an overlay will appear on the page:

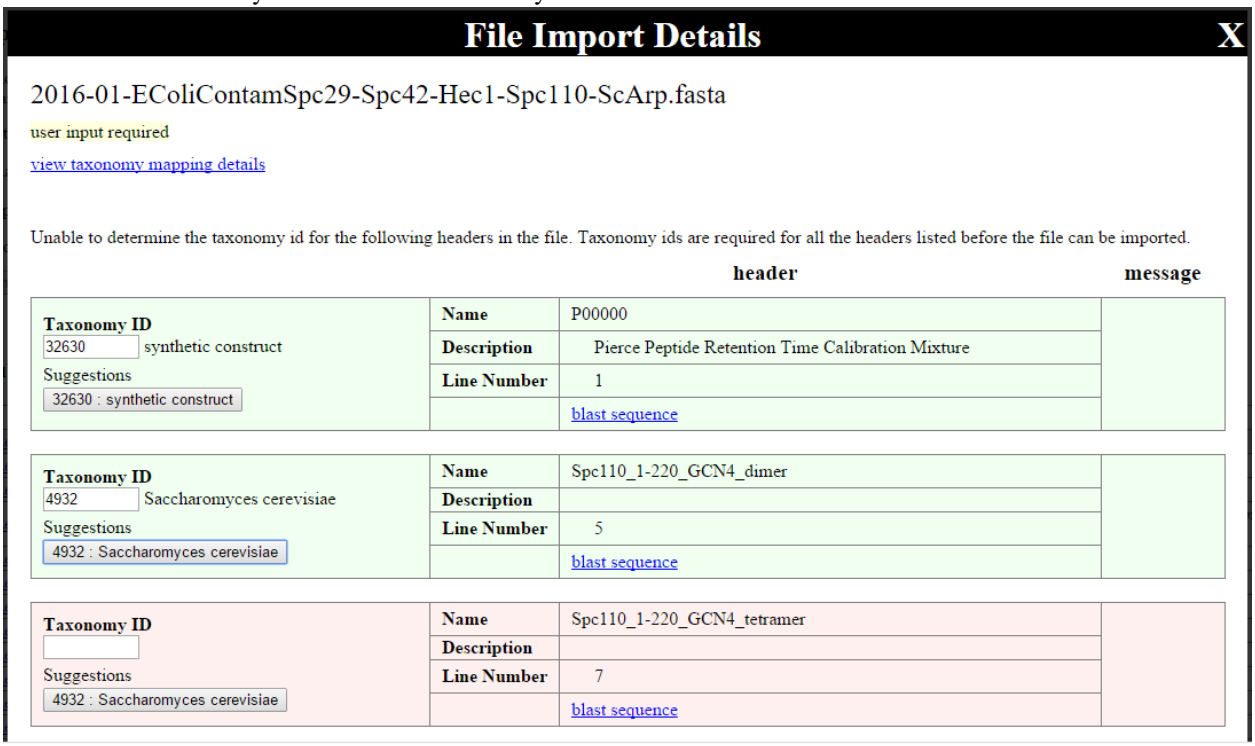
This overlay will display what is happening (FASTA validation, taxonomy determination, etc), and given a progress indication by showing how many sequences are left to process.

If validation fails, the overlay will indicate the failure. The FASTA file will need to be corrected and re-uploaded.

If taxonomy determine fails for too many entries (currently set to 200), the upload process will fail. Please see About nrseq-fasta-importer for more information about what to do in this situation.

1.3.3 User input of taxonomy IDs

If the taxonomy ID could not be determined for some of the FASTA entries (up to 200), an overlay will appear that allows users to manually enter the NCBI taxonomy ID for the FASTA headers:

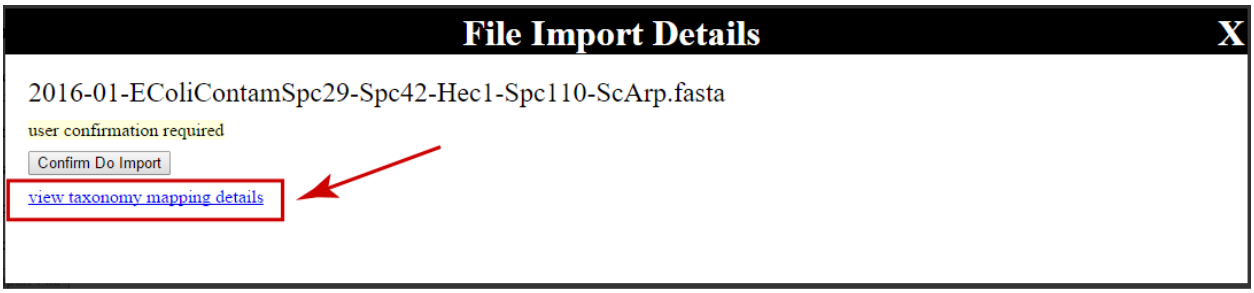


Each entry will be shaded red until a taxonomy ID is added, when it will turn green. When a taxonomy ID is entered, a web services lookup is made to NCBI to retrieve the name for the taxonomy, and this is displayed so that the user may verify the correctness of the ID.Suggestions for taxonomy ID are provided, based on taxonomy IDs associated with that sequence and name in the past. To choose the suggested taxonomy ID, click the button showing the suggestion.

Once all rows are green, click the button at the bottom of the list to re-initiate taxonomy determinations. Once this succeeds, the FASTA may be imported.

1.3.4 Inspect taxonomy ID assignments

To view the taxonomy ID assignments for the proteins in your FASTA file, click the `view taxonomy mapping details` link present in the details overlay for the FASTA file:



This will prompt the user to save an XML file, which may be viewed in a web browser or text editor. The file has the following syntax:

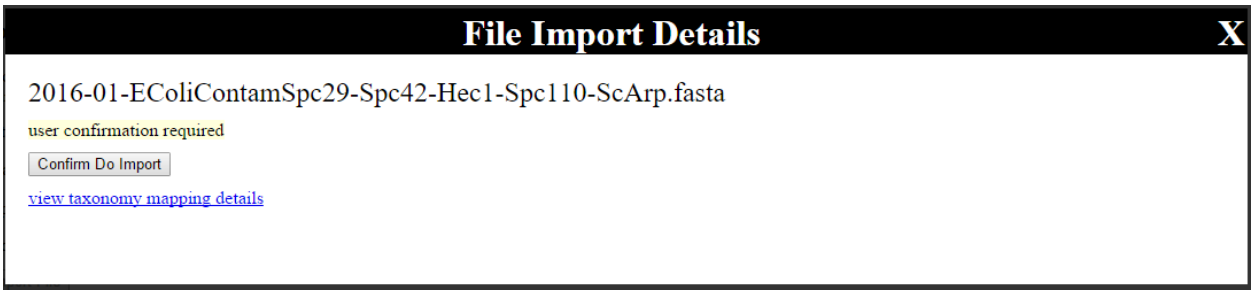
```
<intermediate-import-file>
        <import-file-entry>
                <headerLineNumber>1</headerLineNumber>
                <importFileHeaderEntryList>
                        <headerDescription>Pierce Peptide Retention Time Calibration Mixture</headerD
                        <headerName>P00000</headerName>
                        <taxonomyId>32630</taxonomyId>
                </importFileHeaderEntryList>
                <sequence>
                        SSAAPPPPPRGISNEGQNASIKHVLTSIGEKDIPVPKPKIGDYAGIKTASEFDSAIAQDKSAAGAFGPELSRELGQS
                </sequence>
        </import-file-entry>
                <import-file-entry>
                <headerLineNumber>5</headerLineNumber>
                <importFileHeaderEntryList>
                        <headerName>Spc110_1-220_GCN4_dimer</headerName>
                        <taxonomyId>4932</taxonomyId>
                </importFileHeaderEntryList>
                <sequence>
                        GSMDEASHLPNGSLKNMEFTPVGFIKSKRNTTQTQVVSPTKVPNANNGDENEGPVKKRQRRSIDDTIDSTRLFSEAS
                </sequence>
        </import-file-entry>
        <import-file-entry>
                <headerLineNumber>7</headerLineNumber>
                <importFileHeaderEntryList>
                        <headerName>Spc110_1-220_GCN4_tetramer</headerName>
                        <taxonomyId>4932</taxonomyId>
                </importFileHeaderEntryList>
                <sequence>
                        GSMDEASHLPNGSLKNMEFTPVGFIKSKRNTTQTQVVSPTKVPNANNGDENEGPVKKRQRRSIDDTIDSTRLFSEAS
                </sequence>
        </import-file-entry>
</intermediate-import-file>
```

There is a `<import-file-entry>` for each FASTA entry in the file. Each one contains the associated sequence, header(s) and associated taxonomy ID(s) found for that FASTA entry.

1.3.5 Import the FASTA file

If the `Require confirmation before insert to database:` option was not selected on the upload form, import will automatically begin after successful validation and determination of taxonomy IDs.

Otherwise, user confirmation is required. Confirmation may be given in the overlay showing the import status:



The status overlay for an import may be accessed by clicking the `details` link for the row for that FASTA file in the interface or from the link in the status email received from the web application.

Click `Confirm Do Import` to import the FASTA file to the database.

Upon successful completion, that status message will change to `import complete` and another confirmation email will be sent.