Uploading sequences from a database

Uploading sequences from a database by sequence names

To perform uploading of sequences from the database by sequence names, we need to use the `upload_seq_names_db.pl` component. This component will only run if EGene is in database mode. Trying to use this component in XML mode will cause the pipeline to abort.

Configuration parameters in the `.cnf` file:

```
#==============================================================
PHASE=Upload sequences from a database
program = upload_seq_names_db.pl
#--------------------------------------------------------------
names_file = names.txt
#==============================================================
```

There is only one parameter, `names_file`, which specifies the path (absolute or relative) to a file containing a description of the names of the sequences to be uploaded. This file should contain one specification per line, where the specification can be an arbitrary Perl regular expression. For example, if the file contains:

```
TnMz01.*
TnMy0120
[AB].*
```

the component will upload all sequences with names starting with the characters `TnMz01` (specified by the line `TnMz01.*`), the sequence named `TnMy0120` (specified by the line `TnMy0120`), and all sequences whose names start with a capital A or capital B (specified by the line `[AB].*`). For more information on Perl regular expressions please consult a Perl book.

Uploading sequences from a database using an SQL query

To perform uploading of sequences from the database using an SQL query, we need to use the `upload_sql.pl` component. This component will only run if EGene is in database mode. Trying to use this component in XML mode will cause the pipeline to abort.

Configuration parameters in the `.cnf` file:

```
#==============================================================
PHASE=Upload
program = upload_sql.pl
#--------------------------------------------------------------
sql_query=SELECT SEQUENCE_ID FROM SEQUENCE WHERE SEQUENCE_NAME
~ TnMz.*
#==============================================================
```

There is only one parameter, `sql_query`, which specifies an SQL query to be issued to the database. The query should return a set of sequence identifiers, with any other result causing the crash of the pipeline. In the example above, we have a simple SQL query that returns all sequence identifiers of sequences with names starting with the characters `TnMz`. For more information on possible SQL queries, please consult the PostgreSQL documentation and the EGene database entity-relationship diagram.