USMI Galaxy Demonstrator(UGD): a collection of tools to integrate microorganism information







D.P. Colobraro and P. Romano IRCCS AOU San Martino IST, Genoa, Italy {danielepierpaolo.colobraro,paolo.romano}@hsanmartino.it





Background

Information on microbial strains is mainly collected in heterogeneous repositories, although some web platforms exist that archive it in limited ways. A comprehensive approach to the merging of information on microbial strains is one of the aims of MIRRI. The USMI Galaxy Demonstrator (UGD) was developed to demonstrate, at least partially, the feasibility of this approach.

Scope

Galaxy

Excel to mcl

computer

catalogues

BRENDA

<u>Uniprot</u>

Get microbial data

<u>Upload File</u> from your

Taxonomy from NCBI

TaxonID from NCBI

Get Catalogues Get microbial

Microbial INSDC rRNA from

Retrieve external information

ECNumber ECNumber from

Protein FASTA from Uniprot

FASTA from INSDC retrieves

PMID and DOI from PubMed

From alignment of proteins to

INSDC rRNA related to

FASTA format from ENA

Strains in Catalogue

The goal of UGD is to extend and integrate the contents of catalogues provided by microbial domains Biological Research Centers (mBRCs) by leveraging on the well know data integration tool Galaxy and by exploiting the explicit links existing with the most important biological databases. UGD is a friendly tool, supporting microbe data curation. It avoids manual, potentially long searches on the web and support users in the identification and selection of microorganisms of theirs interest.

USMI Galaxy Demonstrator (UGD)

The UGD is publicly available on-line at http://bioinformatics.hsanmartino.it:8080/. Developed tools are available in two sections (Get microbial data and Retrieval external information) under the label

Figure 1. The UGD menu. Below, is a short description of tools. --- BASIC TOOLS FOR MIRRI

> **Upload file** is a Galaxy' generic tool Get Catalogues is a 'data source' tool to import catalogues from external sources **Taxonomy** retrieves all taxonomy information Microbial INSDC rRNA retrieves related information given a catalogue acronym

'BASIC TOOLS FOR MIRRI'. See fig. 1.

TaxonID retrieves taxonomy ID for all strains **ECNumber** gathers information for enzyme names included in a given catalogue **Uniprot** retrieves protein accession number for all strains in a catalogue **Protein FASTA** retrieves protein sequences by using the relative accession number **INSDC rRNA** retrieves rRNA accession number related to all strains in a catalogue PMID and DOI retrieves Pubmed IDs and Digital Object Identifiers (DOIs) of given

The tools in the second section are enrich basic information provided by mBRC catalogues. Data are retrieved from NCBI, BRENDA and UNIPROT. The tool 'From alignment of proteins to microbial strain' (fig. 2) integrates microbial information with strain number, taxon ID and link to CABRI and related DBs after a blastp alignment. The tool 'Compound and (fig. 3) retrieves strain enzyme' number, link to CABRI, compound / ligand accession number, EC number, name and synonyms for a given

4. Link(s) to other DBs biological term. From alignment of proteins to microbial strain (Galaxy Tool Version 1.1.2) Options Select alignment □ ② □ No tabular dataset available No data? See tip below Identity >= ✓ Execute This tool provides the strain numbers and links, which are retrieved from CABRI [1], to microbial catalogues. Figure 2. From alignment of proteins to microbial strain tool input form

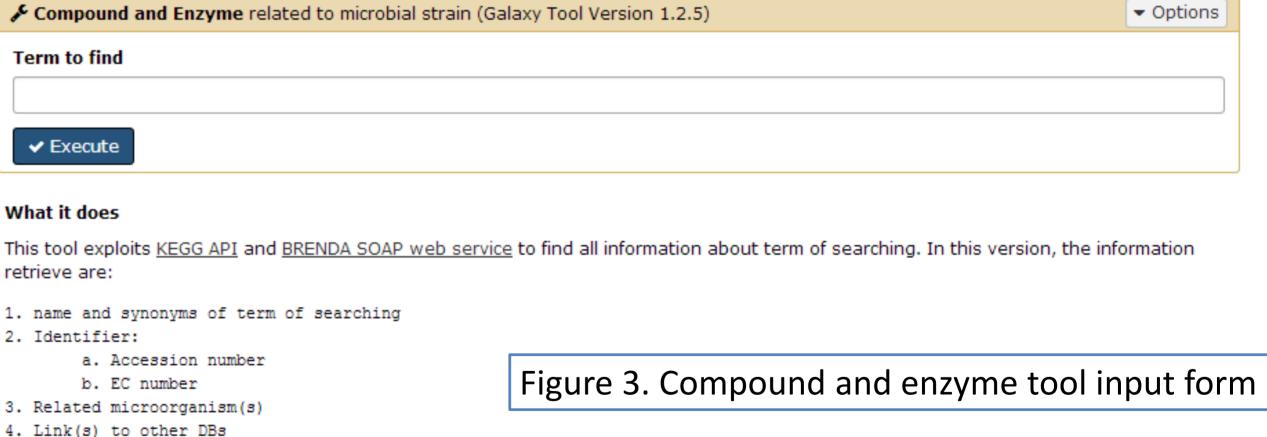
CABRI and MIRRI

Common access to biological resource and information (CABRI) network services (http://www.cabri.org/) offer access to 28 European mBRCs catalogues.

Microbial Resource Research Infrastructure (MIRRI) is a European Research Infrastructure which recently finished its preparatory phase. It aims at connecting all European mBRCs with the objective of providing improved and extended services to the research and industry communities.

Results

The fundamental information about microorganisms are integrated with related taxonomy, literature, sequences and chemical compound on the basis of the strain number, the unique identifier for a given culture, the species name or the qualified species name, in proper format with related databases. The developed tools are able to identify which microorganisms are related to a molecule or a protein of interest to the end user by integrating information from both mBRCs catalogues and external data sources.



Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 312251.

References

The MIRRI Project: www.mirri.org

The Galaxy project: http://galaxyproject.org

bibliographic references **FASTA from INSDC** retrieves rRNA sequences

Compound and Enzyme by using accession number related to microbial strain

<u>microbial strain</u>