A GALAXY APPROACH TO INTEGRATE MICROBIAL DATA: THE USMI GALAXY DEMONSTRATOR

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

GCC 2015

Background

Applications of microorganisms in several fields like health, food and waste management, just to cite few, are the final outcomes of research that make available information via heterogeneous repositories. The microBiological Resource Centres (mBRCs) collect microbial material in catalogues and provide information via own websites. Microbial catalogues rarely gather metabolite, genomics, proteomics and taxonomy data, although these information characterize and validate the microbial species that are collected into mBRC. Some web platforms provide the automatic services to integrate several information, but don’t let researchers set an analysis pipeline to results integration. For this reason, we have set up a first implementation of the local web-based framework, Galaxy, to support the researcher or mBRCs staff to perform bioinformatics pipelines.

Scope

Our purposes are: i) find a method to merge all microbiological sources, ii) offer a clear vision of microorganism data, iii) support curators of catalogues to improve annotations, iv) make this data available to analysis pipelines and v) integrate information.

USMI Galaxy Demonstrator (UGD)

The USMI Galaxy Demonstrator, Galaxy version 15.05, is under active developement and publicly available on-line at http://galaxy.nettab.org:8088. As shown in fig. 1, the developed tools are available in two sections, *Get microbial data* and *Retrieval external information*, under the general label ‘BASIC TOOLS FOR MIRRI’. Galaxy allows to set up workflows to rerun, store and share both specific analyses and data. As shown in fig. 2, tools may be set in various ways in order to define own pipelines. Indeed, implementing basic tools as modular elements allows to make up several pipelines.

CABRI, Common access to biological resource and information, Network Services (http://cabri.org) offer access to 28 catalogues from European Biological Resources Centers (BRCs), since 2000.

MIRRI, Microbial Resource Research Infrastructure, is a pan-European distributed research infrastructure in its preparatory phase which aims to connect all European **mBRCs,** microBiological Resource Centres, with the aim of providing improved and extended services to the research and industry communities. **MIRRI** wants to reach the integration of information on microorganisms with further data that can be found and retrieved from a wide range of biological databases like NCBI, EMBL, BRENDA and UNIPROT.

Conclusion

UGD may be the central point for up- and down-stream analyses oriented to identification and enrichment of microorganism annotations.

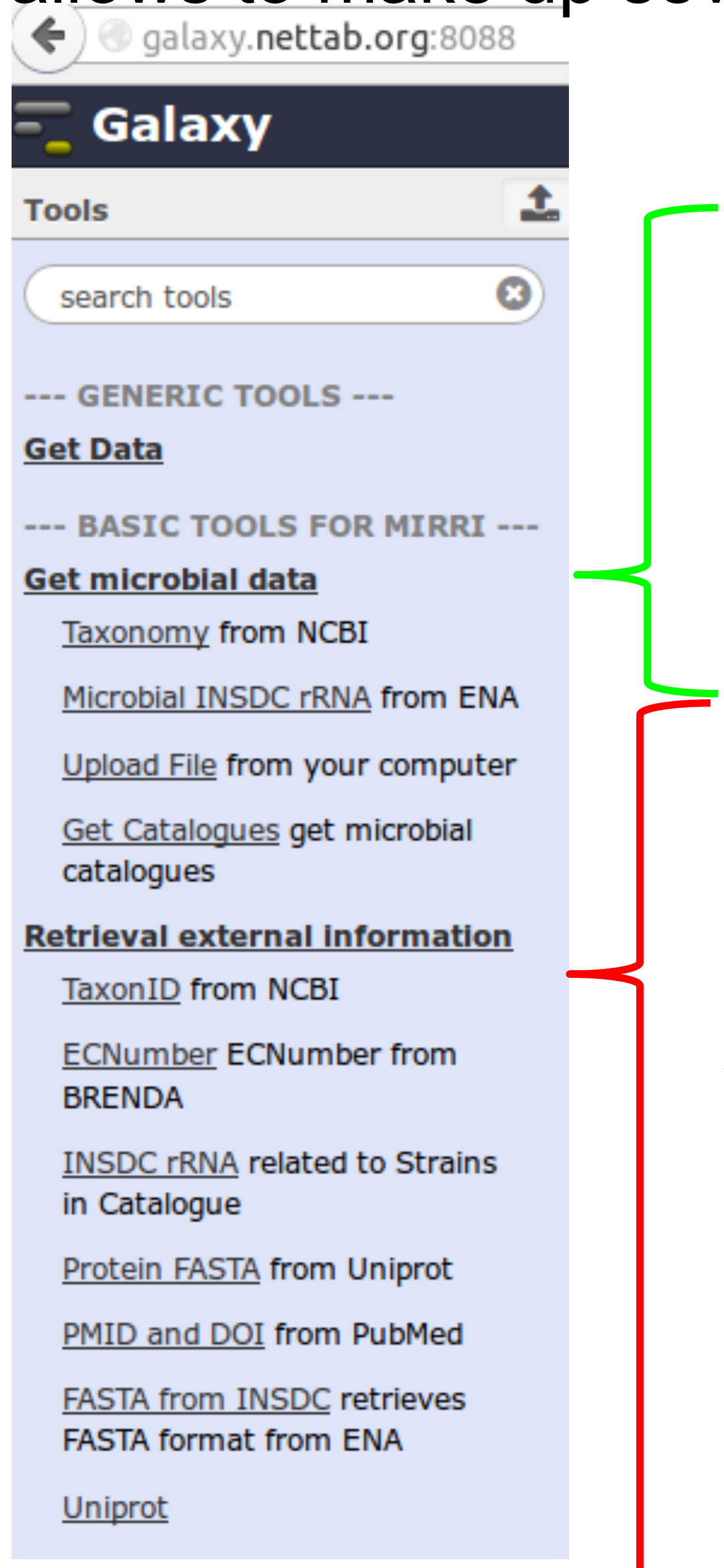
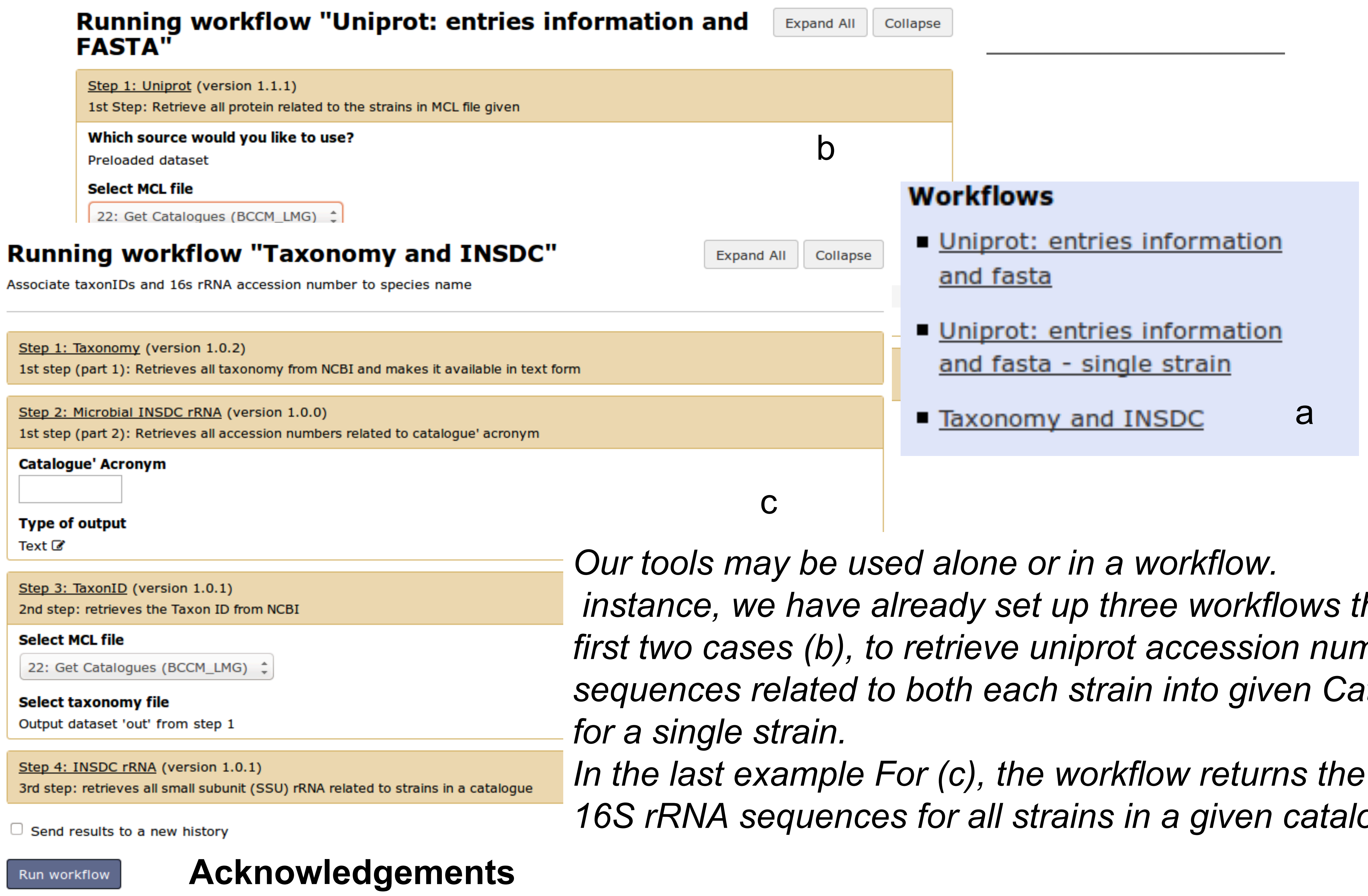

galaxy.nettab.org:8088
Galaxy
Tools
search tools
--- GENERIC TOOLS ---
Get Data
--- BASIC TOOLS FOR MIRRI ---
Get microbial data
Taxonomy from NCBI
Microbial INSDC rRNA from ENA
Upload File from your computer
Get Catalogues get microbial catalogues
Retrieval external information
TaxonID from NCBI
ECNumber ECNumber from BRENDA
INSDC rRNA related to Strains in Catalogue
Protein FASTA from Uniprot
PMID and DOI from PubMed
FASTA from INSDC retrieves FASTA format from ENA
Uniprot

Figure 1

***Taxonomy** retrieves all taxonomy information*
***Microbial INSDC rRNA** retrieves information by using a Catalogue acronym*
***Upload file** is a Galaxy’ generic tool*
***Get Catalogues** is a ‘data_source’ tool to import catalogues from external-web storage*

***TaxonID** retrieves taxonomy ID for all strains*
***ECNumber** gathers information when enzyme names are collected in Catalogue*
***Protein FASTA** retrieves protein sequences by using protein accession number*
***INSDC rRNA** retrieves rRNA accession number related to Strains in Catalogue*
***PMID and DOI** retrieves Pubmed IDs and Digital Object Identifiers (DOIs) of given bibliographic references*
***FASTA from INSDC** retrieves rRNA sequences by using accession number*
***Uniprot** retrieves protein accession number by using all strains*


Expand All
Collapse
b
Workflows
Uniprot: entries information and fasta
Uniprot: entries information and fasta - single strain
Taxonomy and INSDC
a
Expand All
Collapse
c
Send results to a new history
Run workflow

Running workflow "Uniprot: entries information and FASTA"

Step 1: Uniprot (version 1.1.1)
1st Step: Retrieve all protein related to the strains in MCL file given

Which source would you like to use?
Preloaded dataset

Select MCL file
22: Get Catalogues (BCCM_LMG)

Running workflow "Taxonomy and INSDC"
Associate taxonIDs and 16s rRNA accession number to species name

Step 1: Taxonomy (version 1.0.2)
1st step (part 1): Retrieves all taxonomy from NCBI and makes it available in text form

Step 2: Microbial INSDC rRNA (version 1.0.0)
1st step (part 2): Retrieves all accession numbers related to catalogue' acronym

Catalogue' Acronym

Type of output
Text

Step 3: TaxonID (version 1.0.1)
2nd step: retrieves the Taxon ID from NCBI

Select MCL file
22: Get Catalogues (BCCM_LMG)

Select taxonomy file
Output dataset 'out' from step 1

Step 4: INSDC rRNA (version 1.0.1)
3rd step: retrieves all small subunit (SSU) rRNA related to strains in a catalogue

Figure 2

Our tools may be used alone or in a workflow. instance, we have already set up three workflows that allow, in the first two cases (b), to retrieve uniprot accession number and protein sequences related to both each strain into given Catalogue file and for a single strain.
In the last example For (c), the workflow returns the taxonomy and 16S rRNA sequences for all strains in a given catalogue file.

References

- The MIRRI Project: www.mirri.org
- The Galaxy project: http://galaxyproject.org

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.