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

Due to the fragmentation of microbial information and the several branch of human activities encompassed by microorganism applications, a comprehensive approach for merging information on microbes is needed. Although on line service providers collect several data on microorganisms and provide services for microbial Biological Resource Centres (mBRCs), such services are still limited both in contents and aims. The USMI Galaxy Demonstrator (UGD), an implementation of the Galaxy framework exploiting the XML-based Microbiological Common Language (MCL), is meant to support researchers to make an integrated access to enriched information from microbial catalogues, as well as to help mBRC curators in validating and enriching the contents of their catalogues. Researchers and mBRC curators may exploit the UGD to avoid manual, potentially long, searches on the web and to identify and select microorganisms of interest.

UGD tools are written in Python, version 2.7. They allow to enrich the basic information provided by catalogues with related taxonomy, literature, sequence and chemical compound data retrieved from some of the main databases on the basis of the strain number, i.e. the unique identifier for a given culture, and the species names. The data is retrieved by querying database Web Services using either the Simple Object Access Protocol (SOAP) or the Representational State Transfer (REST) access protocols. The MCL format provides a versatile way to archive and exchange data among mBRCs.

Galaxy is a well-known, open, web-based platform which offers many tools to retrieve, manage and analyze different kind of information arising from any life science domain. By exploiting Galaxy flexibility,UGD implements some tools and workflows that can be used to find and integrate several information on microorganisms. UGD tools integrate basic information which may support mBRC staff in the insertion of all fundamental strain information in a proper format allowing integration and interoperability with external databases. They also extend the output by adding information on source materials, including species and strain numbers, and retrieve associated microorganisms which use a compound or an enzyme in whatever metabolic pathway by returning the accession number, synonyms, links to external databases, taxon name, and strain number of the requested molecule.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2766v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

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

Daniele Pierpaolo Colobraro[1], Paolo Romano[1]

[1]IRCCS AOU San Martino IST, Genoa, Italy

Introduction

The information on microorganisms is fragmented and the vision of biological potentials is compromised because of the lack of efforts to merge the data sources on microbes. Microorganism application domains encompass several branch of human activities, including health, food, energy, waste management. For this reason, a comprehensive approach merging the information on microbes is needed.
The Microbial Resource Research Infrastructure (MIRRI) research infrastructure, that recently finished its preparatory phase, aims to connect the European microbial Biological Resource Centres (mBRCs) with the goal of providing improved and extended services to the research and industry [1]. One of the main objectives of MIRRI is the design of improved integration methods and tools for mBRCs data with the aim of: i) assessing available information, ii) pointing out discrepancies, errors and gaps, iii) carrying out in-silico analyses, and iv) curating mBRC catalogue data. Although StrainInfo [2] and the Global Catalogue of Microorganisms (GCM) [3] collect several data on microorganisms and provide services for mBRCs, they are limited both in contents and aims. In order to offer an improved access to microorganisms data through a well-known and widely available tool, an implementation of the web-based Galaxy framework [4] has been set up. It exploits an XML-based language, the Microbiological Common Language (MCL) [5], that represents a data exchange format for sharing mBRCs catalogue information.
The USMI Galaxy Demonstrator (UGD) is meant to support both researchers and mBRC staffs to perform bioinformatics pipelines, importing available microbial catalogues, enriching them with enzyme data, ribosomal RNA sequences and taxon IDs. In this context, our purposes are both to improve annotations on microorganisms collected by mBRCs and to reach the specific strain number via a protein sequence or a biological compound used by microorganisms. Researchers may exploit the UGD to avoid manual, potentially long, searches on the web and to identify and select microorganisms of interest.
In this abstract, we present the developed tools available in UGD, which are available at the address http://bioinformatics.hsanmartino.it:8080/.

Methods

UGD tools are written in Python, version 2.7. Due to the MCL, some tools are implemented with the aim of integrating information provided by mBRC catalogues. These tools are able to enrich the basic information provided by catalogues 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. The data are retrieved from the National Center for Biotechnology Information (NCBI) [6], the European Nucleotide Archive (ENA) [7], the BRaunschweig ENzyme DAtabase (BRENDA) [8] and Uniprot [9].
The tool ‘From alignment of proteins to microbial strain’ exploits the source of sequences returned by a BLAST [10] search in order to verify if the microorganism is available and possibly establish a link with data from the Common Access to Biological Resources and Information (CABRI) [11]. In this tool, a lower similarity threshold for blastp analysis is also requested.
The tool ‘Compound and Enzyme’ is meant to retrieve information on microorganisms involved to some extent with compounds, ligands and enzymes of interest, again with the goal of requesting these strains from collections. Information on a molecule whose name is given by the user are first retrieved from Kyoto Encyclopedia of Genes and Genomes (KEGG) [12] RESTful Web Service [13] or BRENDA SOAP web service [14]. Information on strains related to the given molecule found in these findings are then compared with CABRI and finally returned to the user.

Results

Galaxy is a well-known, open web-based platform which offers many tools to retrieve, manage and analyze different kind of information arising from any life science domain [4]. By exploiting its flexibility, the UGD implements some tools and workflows that can be used to find and integrate several information on microorganisms.
The first group of tools integrate basic information which will facilitate the staff of mBRCs in the inclusion of all fundamental information, in the proper format, thus allowing integration with related databases.
The 'From alignment of proteins to microbial strain' tool extends the usual output of blastp [10] by adding information on source materials, including species and strain numbers. Links to information on strains present in CABRI are provided when a microorganism is available in one of the partner mBRCs. This extended output, which includes protein accession number and definition, percentage of identity to the query sequence, taxon ID and name, strain number, microorganism catalogue and link to CABRI, is presented in a tabular form.
The tool 'Compound and Enzyme' is designed to retrieve associated microorganisms which use a compound or an enzyme in whatever metabolic pathway and it returns the accession number, synonyms, links to available databases, taxon name, and strain number of the requested molecule.
All these tools, specially the last two, are meant to offer fast ways both to integrate microorganism data and to set up, due to the modularity of Galaxy tools, the workflows automatizing a part of steps needed in a bioinformatics pipeline.
Also, the MCL format provides a versatile way to share and store all data retrieved. For these reasons, the UGD may be a first approach in the 'omics' world that increasingly requires the computational skills in order to speed up bioinformatics analysis.
This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 312251.

References

1. Microbial Resource Research Infrastructure - MIRRI project – http://www.mirri.org/
2. Verlyppe et al. Semantic integration of isolation habitat and location in StrainInfo. *BMC Genomics* 2013, 14:933 - http://www.straininfo.net
3. Linhuan Wu et al. Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources. *BMC Genomics* 2013, 14:933 - http://gcm.wfcc.info/
4. Goecks et al. Galaxy: a comprehensive approach for supporting accessible, reproducible and transparent computational research in life sciences. *Genome Biology* 2010, 11:R86
5. Verslyppe B. et al. "Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons." *Research in microbiology* 161.6 (2010): 439-445
6. National Center for Biotechnology Information - www.ncbi.nlm.nih.gov
7. European Nucleotide Archive - ENA - http://www.ebi.ac.uk/ena
8. Schomburg I. et al. BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem Sci*. 2002 Jan;27(1):54-6
9. Uniprot - http://www.uniprot.org/
10. Protein BLAST - http://blast.ncbi.nlm.nih.gov/Blast.cgi
11. CABRI: Common Access to Biological Resources and Infromation - http://www.cabri.org
12. KEGG: Kyoto Encyclopedia of Genes and Genomes - http://www.genome.jp
13. KEGG API - http://www.kegg.jp/kegg/docs/keggapi.html
14. BRENDA Soap Access - http://www.brenda-enzymes.org/soap.php