DOI: 10.26314/GARR-Conf17-proceedings-08

Microbial Resource Research Infrastructure: stato e prospettive sull'integrazione dei dati

Paolo Romano[1], Giovanna Cristina Varese[2]

[1]Ospedale Policlinico San Martino, Genoa,
[2]Mycotheca Universitatis Turinensis, Università di Torino

Abstract. In questo manoscritto, viene presentato lo stato dell'infrastruttura Microbial Resource Research Infrastructure (MIRRI), con particolare attenzione alla situazione nazionale e alla realizzazione di un nodo italiano. Sono inoltre presentati lo stato attuale e le prospettive sullo scambio di dati tra collezioni di servizio per i microorganismi (microbial domain Biological Resource Centers, mBRC) e sulla realizzazione di un sistema informativo condiviso, MIRRI-IS, che sia al contempo in grado di offrire un accesso integrato ai cataloghi degli mBRC e di interoperare con sistemi specializzati sull'analisi dei microorganismi e, in generale, con le più rilevanti banche dati di biologia molecolare, nell'ottica di inserire i dati relativi ai microrganismi nel contesto di un ambiente bioinformatico realizzato con un approccio FAIR (Findable, Accessible, Interoperable, Reusable). Infine, viene ipotizzato un coinvolgimento di GARR nella realizzazione di un prototipo a livello nazionale.

Keywords. Microorganismi, infrastruttura di ricerca, integrazione dati, interoperabilità di sistemi

Introduzione

Nell'ambito dell'iniziativa ESFRI è compresa la Microbial Resource Research Infrastructure (MIRRI) (http://www.mirri.org/) che, nella fase preparatoria, ha incluso 16 partner e 28 istituti collaboranti da 19 stati europei. La missione di MIRRI consiste nel superamento della frammentazione esistente nell'offerta di risorse e servizi in ambito microbiologico. La sua azione è focalizzata sulle esigenze, opportunità e sfide poste ai microbial domain Biological Resource Centres (mBRCs) e agli utilizzatori di microorganismi, sia industriali sia del mondo accademico. MIRRI intende offrire un punto di accesso unico ai servizi e alle risorse offerte dai mBRC.

Il raggiungimento degli obiettivi di MIRRI richiede anche una maggiore interoperabilità tra i sistemi informativi dei mBRC e un'accresciuta offerta di dati. Con l'avvento delle tecnologie high-throughput e il conseguente spostamento della ricerca dai dati a livello cellulare a quelli molecolari, è diventato necessario includere nei cataloghi dei mBRC informazioni di sequenze e d'interazione tra molecole. È necessario implementare un'architettura informatica in grado di gestire dati di sequenza, fenotipici e immagini, che sono intrinsecamente "big data".

1. I sistemi informativi degli mBRC

La maggior parte dei mBRC europei propone il proprio catalogo on-line. Esistono pochi esempi di accesso integrato a più cataloghi. Inoltre, i sistemi informativi dei mBRC sono

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9 © 2018 Associazione Consortium GARR

disomogenei per modalità di accesso ed eterogenei nei contenuti e nel formato dei dati, sostanzialmente non in grado di interoperare e lontani dal un approccio FAIR (Findable, Accessible, Interoperable, Reusable).

L'accesso integrato a più cataloghi è possibile tramite Common Access to Biological Information and Resources (CABRI, http://www.cabri.org/) (Romano P et al. 2005), StrainInfo (http://www.straininfo.net/) (Verslyppe B et al. 2014) e il Global Catalogue of Microorganisms (GCMs, http://gcm.wfcc.info/) (Wu L et al. 2013). CABRI consente l'accesso integrato a 25 cataloghi che includono più di 130.000 risorse microbiologiche. La sua implementazione si basa sull'adozione di dataset e formato dati condivisi (http://www.cabri.org/guidelines/catalogue/CPdata.html). L'indicizzazione dei cataloghi e la ricerca dei loro contenuti è effettuata tramite Sequence Retrieval System (SRS), un motore di ricerca per database di biologia molecolare. Gli utenti possono eseguire ricerche diversificate utilizzando l'"Extended Query Form" dell'interfaccia SRS o l'apposito modulo di ricerca "Simple Search".

Il database StrainInfo comprende alcuni metadati estratti dai cataloghi dei mBRC. L'analisi e integrazione di questi dati consente l'identificazione dei ceppi presenti nelle diverse collezioni, ma originate dallo stesso ceppo iniziale, e permette così di creare una sintesi del contenuto dei cataloghi centrata su ceppi identici. Da questa sintesi ("strain passport") sono resi disponibili link alle collezioni e a database esterni in grado di fornire informazioni estese su tassonomia, sequenze, riferimenti bibliografici e altro. StrainInfo comprende dati su ca. 300.000 ceppi presenti in più di 60 cataloghi con ca. 700.000 numeri di collezione diversi (http://www.straininfo.net/stats).

GCM è il database che include il maggior numero di cataloghi di mBRC. Alcune delle sue caratteristiche, però, non vanno incontro alle esigenze degli utilizzatori. Ad esempio, i dati dei cataloghi devono essere trasferiti manualmente in GCM e questo comporta che molti cataloghi non siano aggiornati. Inoltre, le possibilità di ricerca nel database sono limitate. È necessario quindi realizzare un sistema informativo che superi le limitazioni di CABRI, StrainInfo e GCM. Il nuovo sistema informativo deve allo stesso tempo includere dataset più estesi, ad esempio alle sequenze, e nuove funzionalità che siano in grado di eseguire alcune analisi di tipo bioinformatico.

2. L'architettura proposta e alcuni risultati preliminari

Nel corso della fase preparatoria di MIRRI, è stata definita un'architettura informatica per il MIRRI Information System (Smith D et al. 2017) (figura 1). L'architettura prevede le seguenti componenti:

- un formato standard per lo scambio di dati tra mBRC, sviluppato a partire dal Microbiological Common Language (MCL) di StrainInfo (Verslyppe B et al. 2010),
- un dataset minimo per i dati essenziali di ogni ceppo disponibile, destinato ad evolvere nel tempo sino a comprendere ogni informazione potenzialmente interessante per valutare le applicazioni dei singoli ceppi, da definire come Minimum Information about Biological Resources (MIaBRe),
- un'interfaccia "user-friendly" per le ricerche da inglobare all'interno di un Collaborative Working Environment (CWE),

- opportune Application Programming Interfaces (APIs) e Web Services / automatic workflow per i più diffusi e adottati software di integrazione (come Galaxy (Goecks J et al. 2010) e Taverna (Wolstencroft K et al. 2013).

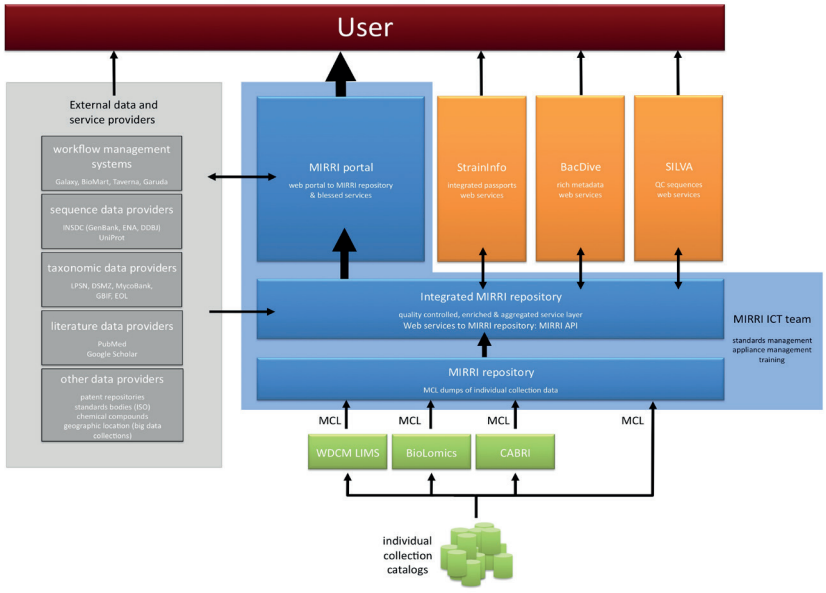

Figura 1
Possibile architettura per il MIRRI Information System

Per valutare la fattibilità delle scelte architetturali, tre diversi prototipi, chiamati "MIRRI demonstrators", sono stati sviluppati durante la fase preparatoria di MIRRI. I tre prototipi affrontano aspetti diversi, ma ugualmente rilevanti, del futuro sistema informativo.

Il prototipo legato al Bacterial Diversity Metadatabase (BacDive, https://bacdive.dsmz.de/) (Söhngen C et al. 2015) mira a estendere i contenuti dei cataloghi e a migliorarne la qualità dei dati. Si tratta di un lavoro impegnativo che ha portato a descrivere in maniera qualitativamente elevata un numero limitato di ceppi. Lo sforzo necessario a descrivere con precisione i ceppi è però superiore a quello che un collezione può normalmente permettersi. Il prototipo ha comunque consentito di identificare e caratterizzare molte informazioni utili e dimostra come l'estensione dei contenuti dei cataloghi sia possibile e si possa raggiungere progressivamente, selezionando ambiti d'interesse specifico e concentrandosi sui ceppi di maggior interesse, portando a collezioni più specializzate, con un numero minore di ceppi, ma appropriatamente caratterizzati e descritti.

Il prototipo di StrainInfo mira a ottenere una migliore integrazione dei contenuti delle collezioni tramite una precisa identificazione dei ceppi comuni, in possesso di più collezioni. Rende così possibile sia la riorganizzazione e focalizzazione delle collezioni, sia lo scambio di dati e l'arricchimento della descrizione dei ceppi comuni.

L'USMI Galaxy Demonstrator (UGD, http://bioinformatics.hsanmartino.it:8080/), è un server Galaxy per i curatori e utenti dei cataloghi dei mBRC ed è mirato a facilitare la gestione dei dati di catalogo ("data curation") e l'integrazione nei cataloghi di dati estratti da database esterni (Colobraro DP and Romano P. 2015). Questo server consente anche di accedere e

utilizzare terminologie e formati standard, migliorando così la qualità dei dati contenuti nel catalogo. UGD consente quindi l'integrazione dei dati dei cataloghi con informazioni di altri database sfruttando un software molto diffuso e senza la necessità di sviluppi informatici.

3. Esigenze e prospettive a livello nazionale

L'effettiva realizzazione di MIRRI dipende, come per ogni infrastruttura ESFRI, dall'attivo coinvolgimento degli stati interessati a sostenerla, che sono chiamati a sostenere sia il nodo centrale, con funzioni di coordinamento, sia le esigenze della comunità scientifica e industriale nazionale. Mentre ogni accordo per il nodo centrale viene stabilito nel contesto europeo, le forme e i modi del sostegno nazionale possono variare a seconda della specifica situazione nazionale.

In Italia esistono molte collezioni di ceppi microbici, ma poche che abbiano una chiara e definita missione di servizio. Inoltre, il coordinamento tra collezioni è molto limitato. La creazione di una rete tra le collezioni, di servizio e non, ha comunque una grossa ricaduta potenziale in vari settori, quali le diverse declinazioni delle biotecnologie, sia nell'ambito della ricerca sia in quello industriale, la salute e l'ambiente. Siamo quindi convinti che il governo italiano dovrebbe sostenere la creazione di una rete nazionale di mBRC.

Recentemente, è stata costituita una Joint Research Unit per la creazione di un nodo italiano di MIRRI (MIRRI-IT). Ad essa hanno già aderito gli enti che hanno partecipato alla fase preparatoria di MIRRI: le Università di Torino, Perugia e Modena e Reggio Emilia, l'Ospedale Policlinico San Martino e il CNR. Lo sviluppo di una stretta connessione tra i mBRC italiani è uno degli obiettivi importanti della JRU, insieme all'implementazione di un sistema informativo per l'accesso integrato ai dati relativi ai microorganismi conservati presso le collezioni, da sviluppare secondo l'architettura di MIRRI-IS. Per questo obiettivo sono richieste notevoli risorse IT. Data la natura pubblica della JRU e dei suoi aderenti, tutti membri del Consortium GARR, il partner ideale è la rete GARR.

4. Conclusioni

Abbiamo presentato in sintesi gli obiettivi e lo stato attuale di MIRRI, un'infrastruttura di ricerca sulle risorse microbiologiche in via di sviluppo nel contesto di ESFRI. Uno degli obiettivi fondamentali di MIRRI è la realizzazione di un sistema informativo flessibile in grado di integrare i dati delle collezioni partecipanti e di numerosi altri database e software bioinformatici per offrire agli utilizzatori di ceppi microbici un ambiente informativo ricco di dati e applicazioni e di facile utilizzo. Abbiamo quindi presentato l'architettura informatica ipotizzata per la realizzazione di MIRRI-IS e alcuni prototipi software che hanno dimostrato la fattibilità delle sue principali componenti. Infine, abbiamo presentato la situazione nazionale, evidenziando come sia ipotizzabile la realizzazione di una piattaforma simile a MIRRI-IS per le esigenze nazionali. Tale piattaforma potrebbe logicamente e utilmente essere implementata nel contesto della rete GARR.

Riferimenti bibliografici

Colobraro DP, Romano P. (2015) A Galaxy approach to microbial data integration: the

USMI Galaxy Demonstrator. Conference Proceedings 12th BITS Annual Meeting, June 3-5, 2015, Milano, Italy. Milanesi L, Mauri G, Masseroli M (Eds). BUP - Bononia University Press SpA, Bologna Italy, 2016, pp. 43-45.

Goecks J, Nekrutenko A, Taylor J, Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible and transparent computational research in life sciences. Genome Biology, 11:R86

Romano P, Kracht M, Manniello MA, Stegehuis G, Fritze D. (2005) The role of informatics in the coordinated management of biological resources collections. Applied bioinformatics 4 (3), 175-186.

Smith D, Stackebrandt E, Casaregola S, Romano P, Glöckner FO. (2017) MIRRI Recommendations for Exploiting the Full Potential of Micro-Organism Data. Ann Biom Biostat 4(1):1027.

Söhngen C, Bunk B, Podstawka A, Gleim D, Vetcininova A, Reimer LC, Overmann J. (2015) BacDive - the Bacterial Diversity Metadatabase. Nucleic Acids Res. 2015.

Verslyppe B, Kottmann R, De Smet W, De Baets B, De Vos P, Dawyndt P. (2010) Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. Research in microbiology 161.6:439-445

Verslyppe B, De Smet W, De Baets B, De Vos P, Dawyndt P. (2014) StrainInfo introduces electronic passports for microorganisms. Syst Appl Microbiol. Feb;37(1):42-50.

Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, Bhagat J, Belhajjame K, Bacall F, Hardisty A, Nieva de la Hidalga A, Balcazar Vargas MP, Sufi S, Goble C. (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res. 41(Web Server issue):W557-561

Wu L, Sun Q, Sugawara H, Yang S, Zhou Y, McCluskey K, Vasilenko A, Suzuki K, Ohkuma M, Lee Y, Robert V, Ingsriswang S, Guissart F, Philippe D, Ma J. (2013) Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources. BMC Genomics, 14:933.

Autori

Paolo Romano - paolo.romano@hsanmartino.it

Bioingegnere, lavora come bioinformatico dal 1993. I suoi interessi di ricerca sono legati all'integrazione di dati biomedici e allo sviluppo di procedure automatiche per l'analisi dei dati. Si è occupato a lungo di database per le risorse biologiche. Dal 2012 lavora nel laboratorio di proteomica. Organizza i workshop NETTAB sulle tecnologie ICT emergenti per la ricerca biomedica.

Giovanna Cristina Varese - cristina.varese@unito.it

Professore Associato in Botanica Sistematica presso l'Università di Torino. Responsabile Scientifico della Mycotheca Universitatis Taurinensis (MUT), una delle più grandi collezioni ex situ di microrganismi in Italia e in Europa. Partecipa attivamente alla creazione dell'Infrastruttura Europea MIRRI e sta coordinando la creazione del network italiano delle collezioni di microorganismi MIRRI-IT.