La Riproducibilità nella ricerca – Proposta di attività didattica

Docente di riferimento: Dott. Stefano Biffani - Istituto di Biologia e Biotecnologia Agraria, CNR.

Docenti: Dott.ssa Vittoria Asti, Dott. Arnaud Molle - Dipartimento di Scienze Medico-Veterinarie, UNIPR

Background

I progressi raggiunti in questi ultimi anni nelle scienze ed applicazioni (bio)informatiche hanno contribuito a stimolare nel mondo scientifico ed applicato due fenomeni straordinari: una raccolta di dati su larga scala e ad alto rendimento (*high-throughput*) e lo sviluppo e l'implementazione di algoritmi statistici sempre più complessi per le analisi dei dati stessi. Questi due fenomeni hanno portato enormi progressi nella scoperta scientifica, ma hanno sollevato due serie preoccupazioni. La complessità delle moderne analisi dei dati solleva interrogativi sulla loro **riproducibilità**, intendendo con questo la capacità, da parte di ricercatori indipendenti, di ricreare i risultati ottenuti dagli autori originali utilizzando i dati originali e le stesse tecniche di analisi. Una delle principali cause è legata alla mancanza di disponibilità di dati originali e/o del codice informatico. Una preoccupazione più generale è la replicabilità delle scoperte scientifiche, che riguarda la frequenza con cui i risultati sono confermati da ricerche completamente indipendenti. Sebbene riproducibilità e replicabilità siano argomenti correlati tra loro, gli stessi si concentrano su aspetti diversi del progresso scientifico. Un altro aspetto della **riproducibilità,** che tra l'altro è legato all'origine del suo utilizzo, riguarda molto più semplicemente la capacità di documentare e quindi trasferire in primo a luogo a se stessi, ma anche ai propri colleghi e/o collaboratori, le metodologie utilizzate nell'ambito di una certa analisi. Molto spesso viene eseguita un'analisi ma, nell'ambito dello stesso gruppo di ricerca, non si è più in grado di replicarla o ripeterla.

Oggi giorno il progresso informatico contribuisce a generare i due fenomeni sopra descritti ma anche a fornire gli strumenti per poter rendere maggiormente *riproducibile* l'attività di ogni ricercatore.

Proposta

L'obiettivo del corso, strutturato in 2 settimane (18 ore/settimana), include:

1. *Introduzione ai concetti di base della statistica.*
2. *Introduzione ad approcci di statistica descrittiva e inferenziale.*
3. *Introduzione all'utilizzo del software R e Rstudio.*
4. *Presentazione delle origini della riproducibilità della ricerca.*
5. *Differenze tra riproducibilità e replicabilità.*
6. *Gestione dei metodi e degli strumenti per rendere la ricerca riproducibile.*

Per raggiungere questi obiettivi il corso è strutturato in una parte teorica ed una pratica.

La parte teorica prevede: l'introduzione alla statistica e all'uso di R/Rstudio (settimana 1); lo studio dei concetti di riproducibilità e replicabilità (settimana 2). La parte teorica comprende anche esempi concreti, problematiche e possibili soluzioni.

La parte pratica prevede una parte preliminare (settimana 1) ed una successiva a quella teorica sulla riproducibilità (settimana 2). La parte pratica prevede l'utilizzo di **uno** degli strumenti oggi disponibile per rendere riproducibile la ricerca: il software R attraverso l'utilizzo della piattaforma RStudio ed in particolare del pacchetto Rmarkdown.

Nato oltre 20 anni fa, il linguaggio di programmazione R si è profondamente evoluto nel corso del tempo sino a diventare oggi uno strumento estremamente versatile che permette non solo la

gestione e l'analisi (più o meno avanzata) di dati di origine diversa (SQL, MSDB, csv, txt, xls, pdf, html, json, etc) ma la possibilità di creare reports (doc, pdf, html), pagine e applicazioni web, presentazioni con un unico strumento. Nel tempo è anche venuta meno la limitazione relativa alla dimensione dei dati da gestire ed oggi esistono pacchetti sviluppati ad hoc per gestire i cosiddetti *big data*. Procedure e librerie sono disponibili per le più svariate analisi statistiche ma anche per la semplice gestione dei dati (e.g. trasformazione di variabili).

Nell'ampio panorama di pacchetti oggi disponibili, la libreria *tidyverse* spicca per la sua versatilità, fornendo strumenti per l'importazione dei file (di qualsiasi formato) e per la loro successiva manipolazione e visualizzazione. A questa libreria ed al suo utilizzo si affiancheranno altri strumenti per l'analisi statistica vera e propria e per la cosiddetta "riproducibilità" della ricerca e della routine (e.g. creazione di pipeline dedicate con reportistica di output automatizzata).