Background:
The International Barcode of Life project (iBOL) continuously records and catalogs species information and stores data on the Barcode of Life Database system (BOLD). Advances in DNA analysis have lead to a rapid increase in the volume of data available for researchers involved with the iBOL project to study. As a consequence, modern statistical techniques are playing an increasingly important role in the analysis of such large volumes of data. A popular software package for statistical analysis is R, however existing methods to retrieve data from BOLD into R are inconvenient, time-consuming or return limited information. Therefore we have introduced a more accessible system to provide convenient and direct access to the data stored on BOLD into R.

Advantages of Data Frames
Data retrieved using BOLD.R is stored as data frames. Data frames offer the following advantages:

- **Flexible**: Each column can store a different type of data (e.g. strings, numeric, logical etc.), therefore we are not restricted to just one type of data in a data frame.
- **Labelling**: Each column can be given a meaningful column name.
- **Modifying**: The user can make changes to the data in a data frame without affecting the original data on BOLD.
- **Filtering**: The user can filter data by values in a column.
- **Merging**: Straightforward to merge data and remove duplicates.

Significance
We developed an R library called BOLD.R which allows users to access data directly from BOLD into R via current APIs maintained by BOLD. Users can access their own private data by logging in to BOLD using BOLD.R, or they can access public data without the need to login. All data accessed using BOLD.R is stored in R with a consistent internal structure and this allows the user to employ the suite of functions provided by BOLD.R along with existing packages to perform statistical analysis on DNA sequences.

Data collection and validation

![Diagram of data collection and validation process]

Key Field | Sequence Alignments | Custom fields
---|---|---
Record | Presence | Accession | BIN.
COI barcode | - | - | -
CYTB barcode | - | - | -

Figure 1: An example of a data frame with real data accessed from BOLD.

Long Term Plan
- **Employ the analytical engine of BOLD**: Develop the BOLD API and BOLD.R to harness the power of the analytical engine of BOLD through cloud computing.
- **Establish a script repository**: Users can store their own scripts privately or make scripts public for others to use.
- **Integrate popular scripts as Plugins on BOLD**.

Summary
BOLD.R is a powerful, flexible and convenient package which provides the user with the capacity to access and analyze large amounts of private and public data on BOLD directly through R. It can therefore become an invaluable tool to assist researchers make informed decisions through the automation of data processing and enable the analysis of extremely large datasets.

References

This project is funded and supported by

1. Biodiversity Institute of Ontario, University of Guelph, Guelph, ON, N1G 2W1, Canada.
2. Department of Mathematics and Statistics, University of Guelph, Guelph, ON, N1G 2W1, Canada.
3. Department of Computer Science, University of Guelph, Guelph, ON, N1G 2W1, Canada.

Note: All of the plots, maps and figures above were obtained using BOLD.R and existing R libraries.