Bioinformatics: From Algorithms to Applications
Saint Petersburg, Russia, July 27-28, 2020

Installing and Searching BLAST Databases in a Data Science Framework

Graham Alvare[1], Abiel Roche-Lima[2], Brian Fristensky[3]*

Access Norwest Co-op Community Health, Winnipeg, Canada[1] RCMI Program, Medical Science Campus, University of Puerto Rico[2] and Department of Plant Science, University of Manitoba, Winnipeg, Canada[3] *Corresponding author brian.fristensky@umanitoba.ca

ABSTRACT

Data science embodies a pipeline of processes: acquisition, cleaning and organization of data, quality control and assurance, validation, and downstream visualization and analytics. Because of the overwhelming number of tools for each of these steps, the greatest challenge is often making those tools work in concert to facilitate a thorough and insightful analysis.

The BIRCH [1] system is a framework consisting of hundreds of bioinformatics tools, unified through the BioLegato family of programmable graphical applications [2]. Each BioLegato application represents a specific class of biological objects, packaging together the data and the methods for each class of objects. We describe BioLegato applications for BLAST searches, implementing data science principles. For example, in **blncbi** the user retrieves sequences from NCBI using a graphical Entrez query builder. Amino acid sequences matching the query pop up in **blprotein**, a BioLegato application that lets the user run protein-specific tasks. A protein can be selected for a BLAST search, and output will appear in **bpfetch,** a BioLegato spreadsheet object for protein hits. **blpfetch** makes it easy to scan hundreds of hits, refining the list into one or more subsets for retrieval. Sequences are retrieved to a new **blprotein** object for downstream analysis. For example, proteins aligned using mafft would pop up in a **blpalign** object. Because each object is a separate window with a small screen footprint, the user has more of a sense of working directly with the data compared to web pages that fill the screen.

BioLegato makes it easy to experiment with the data at all steps in a pipeline. Because output of each step appears in a new BioLegato object, there are no dead ends. Output from one step can be used directly as input for subsequent steps because BioLegato takes care of things like file format conversion, which is a tedious and sometimes error-prone part of using tools at the command line. We call this process ad hoc pipelining. Ad hoc pipelining enables the user to learn from each step before going to the next. We also describe **blastdbkit**, a Python script run from BioLegato, for downloading and managing BLAST databases on the users's computer.

Together, BioLegato applications provide a seamless point and click pipeline for sequence database searches, within the context of the larger BIRCH system. New programs can be added to any BioLegato application by creating a file using BioLegato's PCD language [2], which specifies parameters to be set and a shell command to run the program. In this way, the core BIRCH functions can be integrated with locally-installed bioinformatics software. BIRCH Web site: http://home.cc.umanitoba.ca/~psgendb

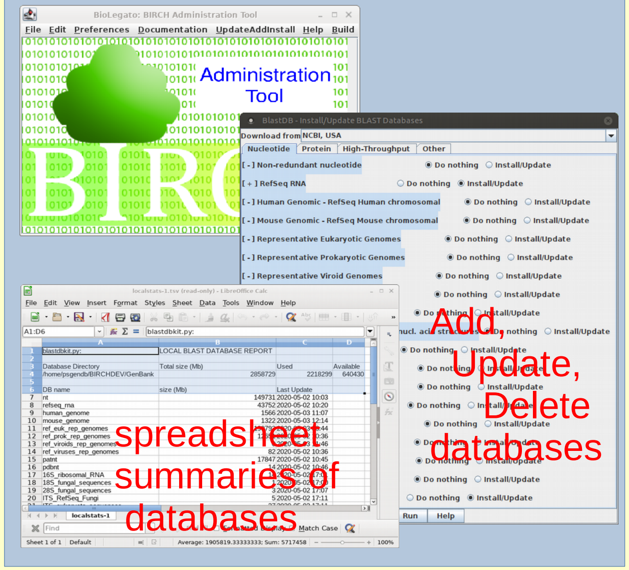
REFERENCES

1. Fristensky B (2007) BIRCH: A user-oriented, locally-customizable, bioinformatics system. BMC Bioinformatics, 8:5

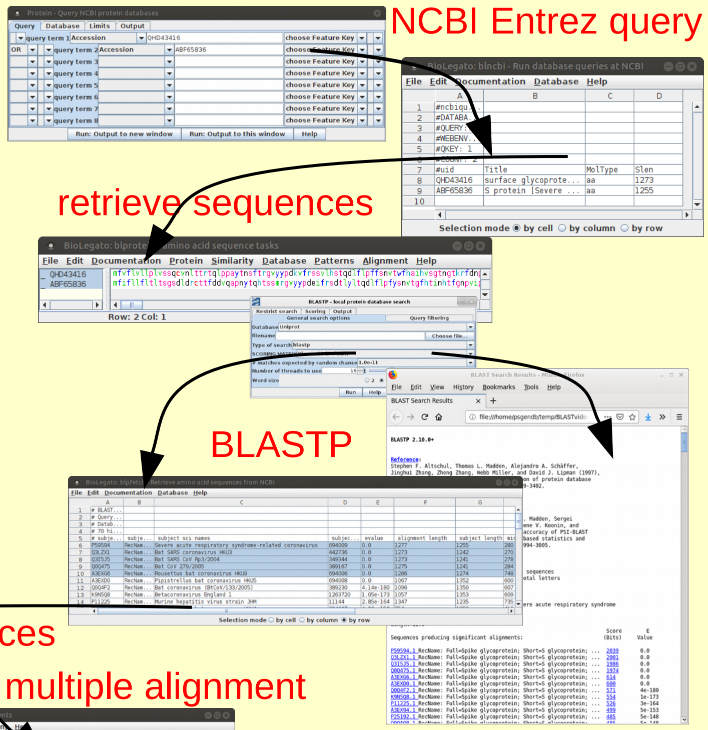
2. Alvare GGM, Roche-Lima A, Fristensky B (2012) BioPCD - A Language for GUI Development Requiring a Minimal Skill Set Int. J. Computer Appl. 57:9-16.

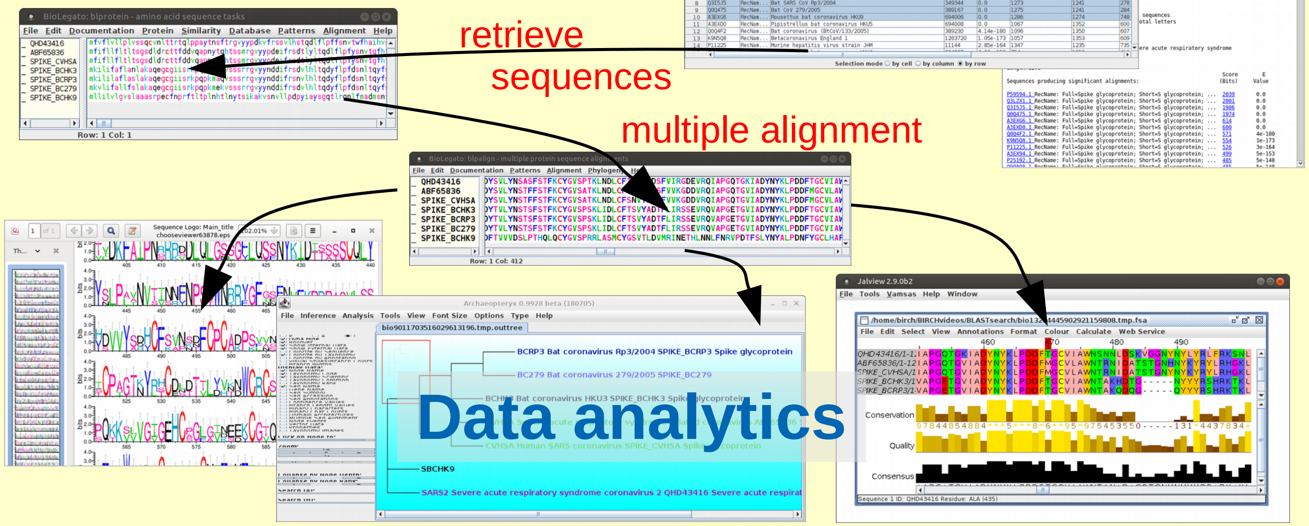
Watch other BIRCH videos on the BIRCH YouTube channel.

blastdbkit - Point and click installation of BLAST databases


BioLegato: BIRCH Administration Tool
File Edit Preferences Documentation UpdateAddInstall Help Build
Administration Tool
BIRCH
BlastDB - Install/Update BLAST Databases
Nucleotide Protein High-Throughput Other
[-] Non-redundant nucleotide
[+] RefSeq RNA
[-] Human Genomic - RefSeq Human chromosomal
[-] Mouse Genomic - RefSeq Mouse chromosomal
[-] Representative Eukaryotic Genomes
[-] Representative Prokaryotic Genomes
[-] Representative Viroid Genomes
○ Do nothing ○ Install/Update
Run Help
LOCAL BLAST DATABASE REPORT
Add, Update, Delete databases
spreadsheet summaries of databases

A Data Science approach to BLAST: *ad hoc* pipelining


Protein - Query NCBI protein databases
Query Database Limits Output
Run: Output to new window
Run: Output to this window
Help
NCBI Entrez query
BioLegato: blncbi - Run database queries at NCBI
File Edit Documentation Database Help
Selection mode ● by cell ○ by column ○ by row
retrieve sequences
BioLegato: blprotein - amino acid sequence tasks
File Edit Documentation Protein Similarity Database Patterns Alignment Help
BLASTP - local protein database search
BLASTP
BLAST Search Results
BioLegato: blpfetch - retrieve amino acid sequences from NCBI


BioLegato: blprotein - amino acid sequence tasks
File Edit Documentation Protein Similarity Database Patterns Alignment Help
retrieve sequences
multiple alignment
BioLegato: blpalign - multiple protein sequence alignment
Archaeopteryx
Jalview 2.9.0b2
Conservation
Quality
Consensus
Data analytics