AN ARB TOOLKIT FOR CUSTOM DATABASES & MASSIVE PHYLOGENIES

Steven Essinger[1], Erin Reichenberger[1], Gail L. Rosen[1], Chris Blackwood[2]

[1]Electrical & Computer Engineering, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, US

[2]Department of Biological Sciences, Kent State University, 217 Cunningham Hall, Kent, OH 44242, US

Goals:

The overall goal of this project is to identify microorganisms that do not produce extracellular enzymes to degrade plant polymers, but still benefit from enzymes produced by other members of the microbial community ("cheaters"). To accomplish this goal, we are focusing on particularly relevant gene families of transporters and extracellular enzymes. Bioinformatics tools are required to look for evolutionary patterns and conserved sequence regions amenable to primer design within these gene families.
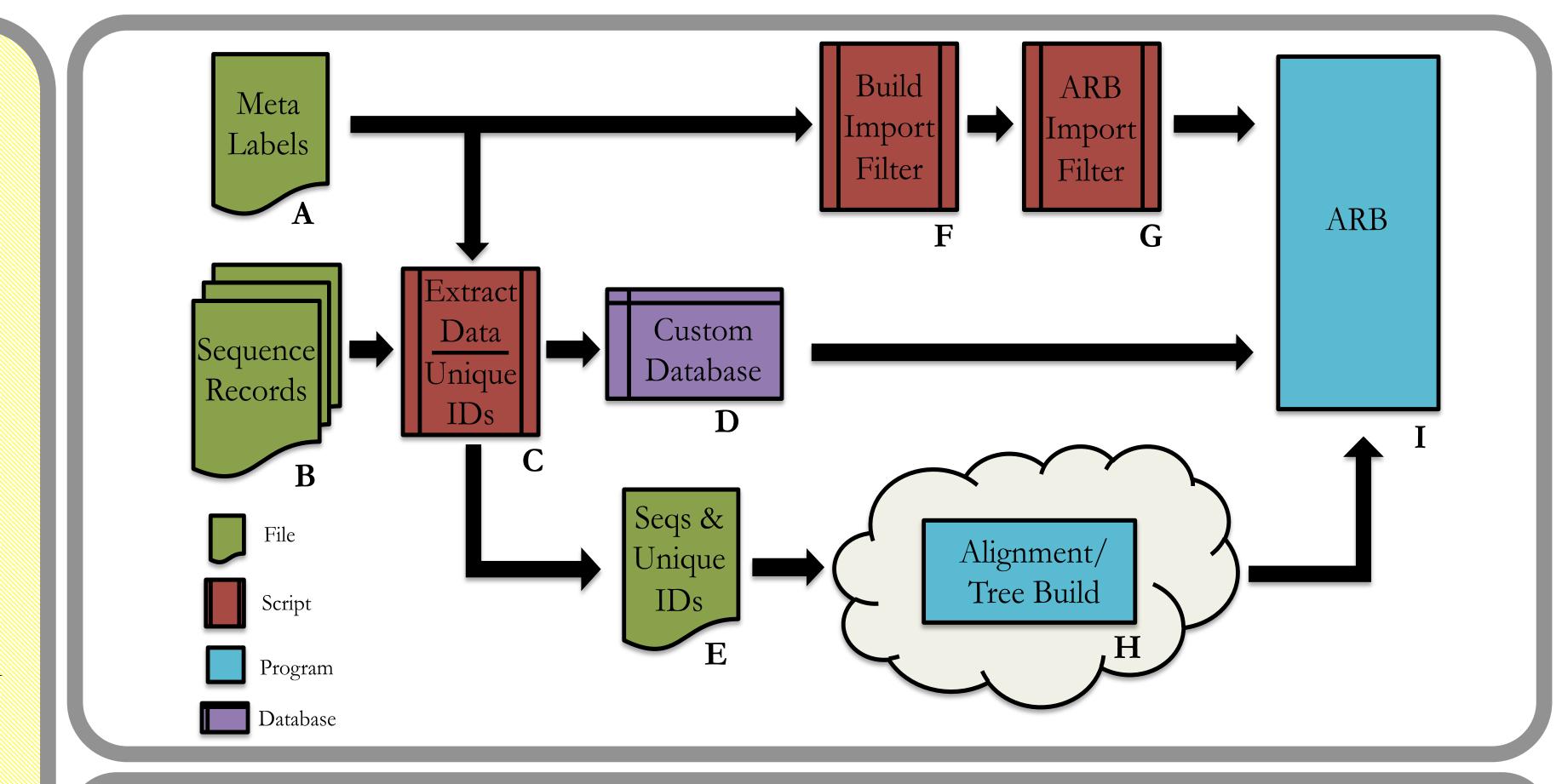
Challenges:

✧ Tens of thousands of DNA, RNA and protein sequences must be analyzed:
- 'Non-standard' meta-data annotations from multiple sources
- Alignment and Phylogenetic Tree Construction
- Primer Design

✧ Lack of suitable in-house computational resources

Solution:

A pipeline (right) has been developed to couple the analytical power of the ARB environment with powerful external computational resources. The solution includes a set of python scripts for:

1. Extracting sequence/meta-data information from a variety of resources
2. Generating an ARB import filter for this custom database of sequence information
3. Method for constructing phylogenetic trees externally, then importing back to ARB for analysis

All python scripts, ARB import filters and a tutorial are available at:
http://www.ece.drexel.edu/gailr/EESI/tutorial.php

Phylogenetic Tree: The tree (right) consists of 1200 sequences from the sugar transport gene family. It was constructed using MAFFT and RAxML via the Cipres Science Gateway. Analysis was performed in ARB enabled by the pipeline shown above. Over 50 different meta-labels were imported for each AA sequence. The leaf descriptors (shown) were parsed out of respective GenBank files.

A: Meta Labels

This is a user supplied text file containing the names of the meta-data fields. For each sequence in the database, all corresponding meta-data will be assigned to its respective name from this file.

B: Sequence Records

Sequence records (e.g. GenBank) from which to extract desired sequences and meta-data information.

C: Extract Data/Assign Unique IDs

Python script for extracting information from GenBank files and NCBI Taxonomy. This script also generates a unique identification number for each database entry. This serves two purposes: 1) phylogenetic trees constructed outside of ARB may be imported into ARB and associated with meta-data and 2) future meta-data may be added to existing sequence records within the ARB environment.

D: Custom Database

This is the fasta formatted custom database. Each entry consists of a tab-delimited header followed by a sequence on the next line. The header information starts with the unique ID followed by all corresponding meta-data, ordered according to the list of meta-data names provided in the meta-labels text file described in section A.
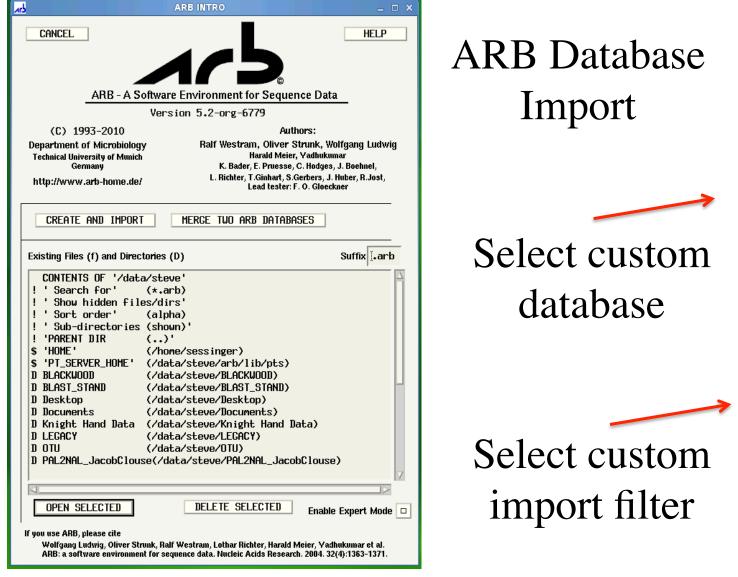
E: Sequences & Unique IDs

This is the sequence file for alignment and phylogenetic tree building using any external algorithms and computing resources of choice. It is a fasta-formatted file containing only the unique ID in the header.
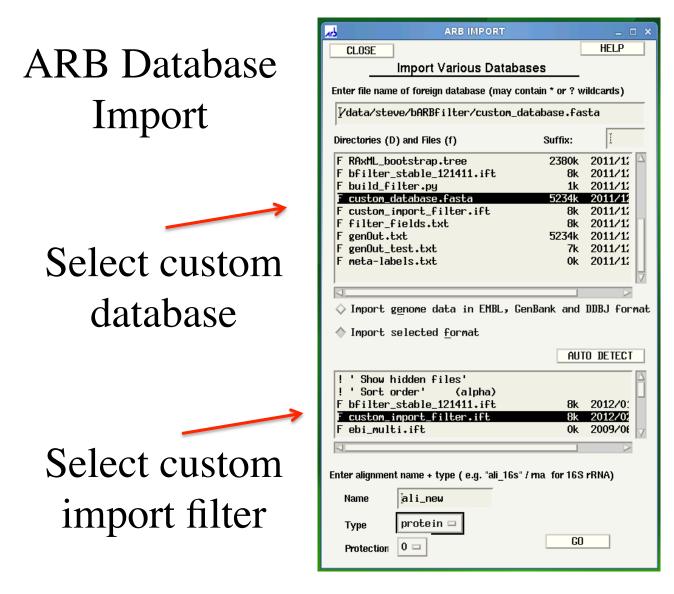
F: Build Import Filter

This is a python script for creating the import filter for your custom database. The meta-labels text file described in section A is all that is required by the script to build the filter. The script will output a file called custom_import_filter.ift.
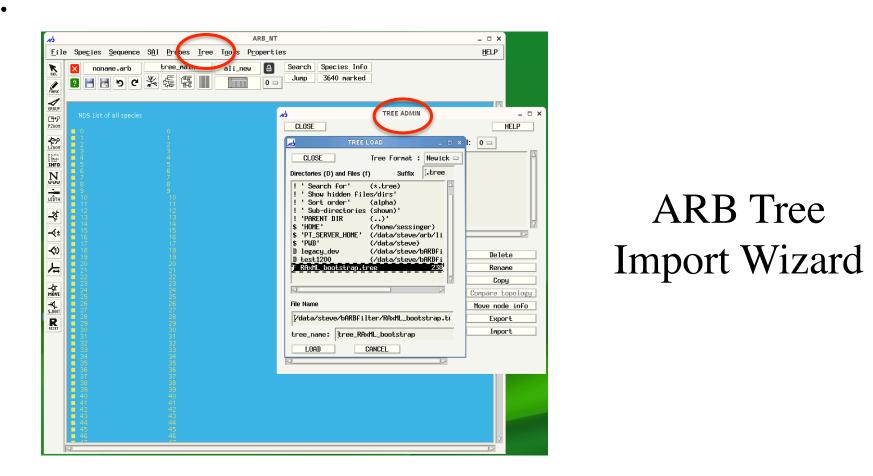
G: ARB Import Filter

This is the custom .ift ARB import filter for your database, which is automatically generated by the python script described in section F. The filter must be placed into the ARB import filter directory '/arb/lib/import/'. To import your custom database, follow the typical ARB procedure for creating a new database.

ARB Database Import

Select custom database

Select custom import filter

H: External Alignment & Tree Build

The choice of tool for alignment and phylogenetic tree building is at the discretion of the user. For example, you may be interested in building your tree using the maximum likelihood algorithm offered by RAxML, but desire to use external resources due to inadequate internal computational capabilities. The Cipres Science Gateway offers a potential solution for both sequence alignment (e.g. MAFFT) and tree building (e.g. RAxML) using the TeraGrid computing cluster. The user would submit the sequences.fasta file from section E to Cipres and then build a pipeline within Cipres to construct the tree. The tree output from this process would then by imported into ARB using the tree import function found under 'Tree/Tree Admin/Import' within the ARB environment.

ARB Tree Import Wizard

I: ARB

This is the ARB program. Information on downloading, installation and use may be found at the ARB website: http://www.arb-home.de