BIOINFORMATIC AUTOMATION APPROACH TO QUALITY ASSESSMENT OF HIGH THROUGHPUT DNA SEQUENCING DATA

Rosanna Alcoverde, Andrea Zuccolo, Aswathy Sebastian, Michele Braidotti, Dave Kudrna, Rod A. Wing
Arizona Genomics Institute, Dept. Plant Sciences/BIO5,
University of Arizona, Tucson AZ

National Science Foundation
Plant Genome Program

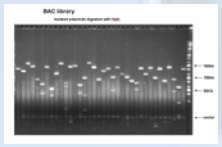
Abstract

Systematic management and assessment of the exponential increase of DNA sequence data is a major burden experienced by high throughput genomics research laboratories. The Arizona Genomics Institute produces ~6 Mbp of sequence data per day from a multitude of species and sources. It was proposed to create an automated check system for quality and contamination of the data produced; a pipeline that automatically retrieves sequence code on a daily basis, performs some assessment and reports a summary. This pipeline is a collection of computer codes which automatically: (1) extracts a random portion (10%) of each DNA sequence project produced daily, (2) sends the data to a super computer to perform similarity searches, (3) retrieves outputs from the super computer and extracts and sorts the desired information, and (4) displays summary reports in an easily readable form. Since this pipeline will be used daily, a set of efficient and straightforward tools were needed.

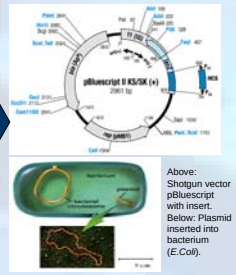
We utilized the programming language Perl and BioPerl modules to write the pipeline. These are object-oriented modules designed for the most common bioinformatics needs. The University of Arizona super computer "Aura" was used to perform BLAST analyses on the sequences. The "Aura" computer is used because of its speed and because it maintains updated genbank deposited data sets for which the AGI DNA sequences are compared. We adopted the outputs to confirm species, check for contaminations, and report success.

Overview of Genomics

First isolate DNA from the desired organism.



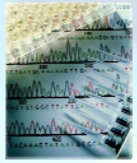
Isolated high molecular weight DNA is then "cut up" at specific sequences with restriction enzymes (a process known as digestion) resulting in fragments of the chromosome ranging from 120-150kb on average. These DNA fragments are ligated into BAC vectors (resulting in what is called a plasmid) and then transformed into *E.Coli* cells for **amplification**. All the resulting cells with the duplicated DNA insert are stored (usually in 384 well plates) making up a BAC (bacterial artificial chromosome) library



Above: Shotgun vector pBluescript with insert. Below: Plasmid inserted into bacterium (*E.Coli*).

BAC clones are broken down into shotgun libraries. These libraries have smaller DNA inserts (ranging from 2-3 kb) in their plasmids. This process is done first by lysing the *E.Coli* cells containing the BAC plasmids. The following steps consist of precipitation, shearing, end repair, size selection, elution from gel, ligation, then finally transformation which yields the final result: *E.Coli* cells containing plasmids with small DNA inserts that have been subcloned from a BAC clone

Sample Trace File and instruments utilized in sequencing process.

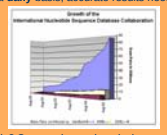


Sequencing the BAC clones (in essence parts of the organism's genome) is now possible with the smaller inserts. The sequencing process uses fluorescently tagged ddNTPs complementary to the target DNA. Thus, by analyzing the tagged nucleotides, one finds the base pair sequence. The output from the sequencing machine (a trace file) is then passed to an automated system: PHRED and/or LUCY. This collection script of various programs removes the vector sequences (leaving high quality output) and converts the code to a FASTA file.

My Summer Project (Concentration Area on Bioinformatics)

Analyzing Sequence Data

Problem: An Exponential increase of Genomics data has created a critical need for managing the output. On a **daily** basis, accurate results need to be **partially** examined.



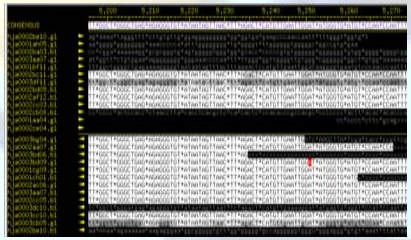
Is our data what we think it is? Sequencing and analyzing unwanted code (contamination), is not only detrimental to the genome information collected, but it wastes time and money. A **checkpoint** before proceeding is critical.

Pipeline Plan:

Take Genome Data	Analyze Sequences	Present Output
Aura (supercomputer) takes random 10% of institute's **daily** sequence output (may be from more than 1 project).	Supercomputer uses BLAST on the sequences against its updated databases. The desired information is extracted from the BLAST output file via Perl and BioPerl code.	Information on the quality of the sequence reads is displayed in a readable form and analyzed: contamination is checked to ensure high quality genomic data is used in reconstructing the BAC clone.

The **hypothesis** is that the pipeline result will provide necessary input for the analysis of quality of the genomic data the institute is producing. Thus, the pipeline **is** the checkpoint needed.

Start of pipeline: Supercomputer takes FASTA files to BLAST.



Reconstruction of BAC: Individual shotgun sequence reads are assembled in Consed. The output is the consensus that reconstructs the BAC sequence.

Databases

In a general sense, a database is an organized body of related information. In bioinformatics, it can be defined as a collection of sequences made so that BLAST could be readily run against it.

GenBank is a major genetic sequence database. More specifically, it is an annotated collection of all publicly available nucleotide sequences and amino acid sequences. Continuing to grow at an exponential rate, it doubles every ten months and currently contains over 65,369,091,950 bases in 61,132,599 sequence records. It is built from submissions from individual laboratories, as well as "bulk submissions from large scale sequencing centers." Thus, the input is controlled by the submitters; consequently, redundancy is a characteristic attributable to the GenBank database.

Produced at the National Center for Biotechnology Information (NCBI), GenBank is part of the International Nucleotide Sequence Database Collaboration which also includes the DNA Data Bank of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL) data library from the European Bioinformatics Institute (EBI). Genomic data is added to this collaboration from around the world from over 100,000 distinct organisms.

NCBI uses the collection of data from GenBank to create derivative databases. For example, nr (non-redundant) databases where redundant sequences have been removed (such as in the RefSeq derivative database) are seen.

Aura (the super computer)

•Format of Input: Input files must be in FASTA format (nucleotide or protein). There may be many sequences in one input file.

•Aura contains several blast databases including swissprot, est, nt, and nr. The nr database was the one used in this pipeline.

•Updating BLAST databases: The first day of each month, each database is downloaded from NCBI.

•Format of output: by default (for each query sequence), the BLAST output contains a list of one line descriptions of the hit along with pair-wise alignments.

With an updated nr database system, the supercomputer located right on the University of Arizona campus is utilized in the pipeline system for its efficiency.

Pipeline Start

FASTA Files

Sequence Data to be blasted against databases

BLAST output file (1)

BLAST

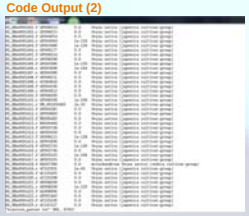
Basic **L**ocal **A**lignment **S**earch **T**ool

This is an algorithm that compares primary sequences of proteins or (in this case) nucleotides of DNA. It's a widely used bioinformatics program. It is ran on the supercomputer.

Checking for Contamination

Code written to extract and display species information

Code Output (2)

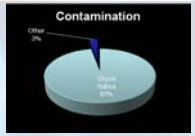


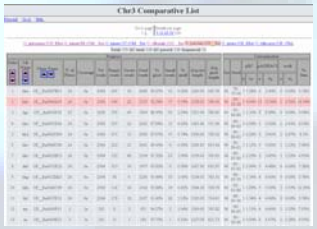
Displays query, hit name, e-value, and species information (critical component).

Pipeline Flow of files Summary:

BLAST is run on the FASTA file and produces a BLAST output file (1). This output file is the input for our Perl and BioPerl codes which extract the desired data (parses it) and displays it as another file (2). This resulting file is stored in a database where a webpage reads from it and finally presents it in a readable form that is meaningful (see screenshot below). This information is then analyzed to ensure that the sequence data tested is of high quality to be utilized. Generally, anything above 15% contamination is investigated.

What output represents in a basic sense.





Tool Box

- •Unix Commands
- •PERL
- •BioPerl

Linux is the specific kind of Unix utilized. These commands were employed in **Secure Shell** or **SSH** (a network protocol that allows data to be exchanged over a secure channel between two computers). Interaction with the supercomputer was done via SSH.

Perl and BioPerl were the programming languages used in parsing the BLAST output files. Perl is known as the first choice in bioinformatics because of its ability to deal with strings and pattern matches. However, the majority of the code in the pipeline was written in BioPerl. This object-oriented language is a collection of perl modules employed in bioinformatics applications.