

Shotgun Sequence Alignment and Haplotype Discovery

A Java Program for Discovering Haplotypes from Shotgun
Data

Luke Blanchard John P. Daigle
Na'el Mohammed Abu-halaweh

Department of Computer Science
Georgia State University

12.07.06

Outline

- 1 Introduction
 - The Biological Problem
 - The Computation Problem
 - Alignment
 - Haplotype Discovery
- 2 Solution
 - Software Overview
 - Current Status

Outline

- 1 Introduction
 - The Biological Problem
 - The Computation Problem
 - Alignment
 - Haplotype Discovery
- 2 Solution
 - Software Overview
 - Current Status

Our Goal

The goal of this project is to create a tool that biologists can use to discover haplotypes within a population using shotgun sequencing.

Shotgun Sequencing

- Means of sequencing DNA
- Uses small fragments produced by Chain Termination
- Small fragments must be built into full genome.

What is a Haplotype?

In this context, a Haplotype is a set of *Single Nucleotide Polymorphisms*.

Single Nucleotide Polymorphism

When two sequences of DNA from two members of the same species differ by a single nucleotide.

Why do we care?

- Humans
 - Evolutionary Research
 - Genetic Testing for Disease
- Virus/Bacteria
 - Evolutionary Research
 - Genetic Testing for Drug Susceptibility/Resistance

Biological Problem Statement

- To find the SNP positions in a shotgun sequences of DNA
- To use that data to generate a “best guess” of the haplotypes.

Several Problems

There are several algorithmic challenges in this project.

- Three NP Complete Problems
- One String Matching
- Two Graph Theory

Clearly, approximation algorithms will be needed.

Sequence Alignment

Application

Input: a set of DNA fragments.

Output: a realistic sequence that overlaps all fragments

Analysis

This is an instance of the shortest superstring problem.

- 1 An algorithm that provides a good approximation for the shortest superstring will also provide a good guess for the original DNA sequence.
- 2 An algorithm that solves this problem will produce an approximately correct alignment for each input string.

Finding Conflicts

Application

Input: A set of shotgun reads, correctly aligned.

Output: A SNP conflict graph

Analysis

The SNP conflict graph is a representation of which DNA fragments cannot belong to the same species.[1]

- This is a simple problem to solve
- producing a conflict graph is crucial for the next step

Minimum Number of Species

Application

Input: A SNP conflict graph

Output: an approximation of the minimum number of species represented by the graph.

Analysis

- This is an example of the *minimum coloring problem*
- Cannot be solved perfectly and quickly
- Good approximations are well established[2]
- Only solves half the problem

Haplotype Discovery

Definition

Input: The SNP conflict graph

Output: The maximum number of species that the dataset can contain

Analysis

- The *Maximum Clique Problem* is equivalent to this
- More difficult than the previous problem
- Good heuristics do not exist to solve this problem

Outline

- 1 Introduction
 - The Biological Problem
 - The Computation Problem
 - Alignment
 - Haplotype Discovery
- 2 Solution
 - Software Overview
 - Current Status

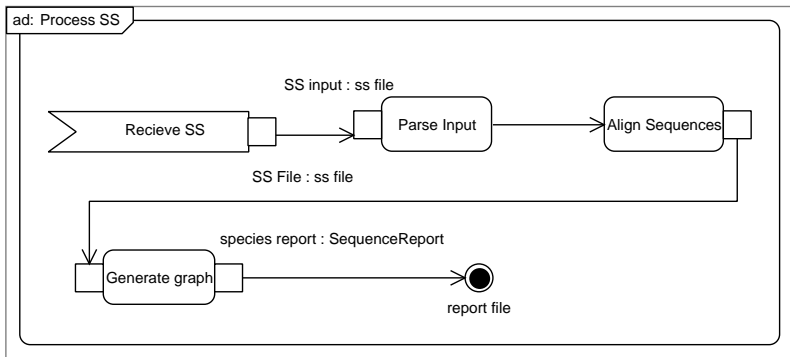
The Haplotype Discovery System I

Implementation

The Haplotype Discovery System is a Java program that produces information about the probable number of haplotypes and accompanying phenotypes from an input of shotgun sequences.

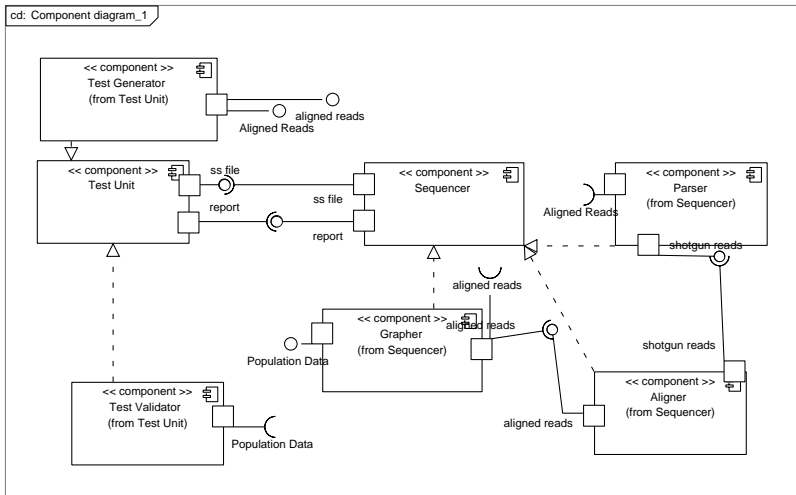
Major components are an alignment module for aligning sequences, a graphing module for haplotype discovery, and several report generation options.

The Haplotype Discovery System II

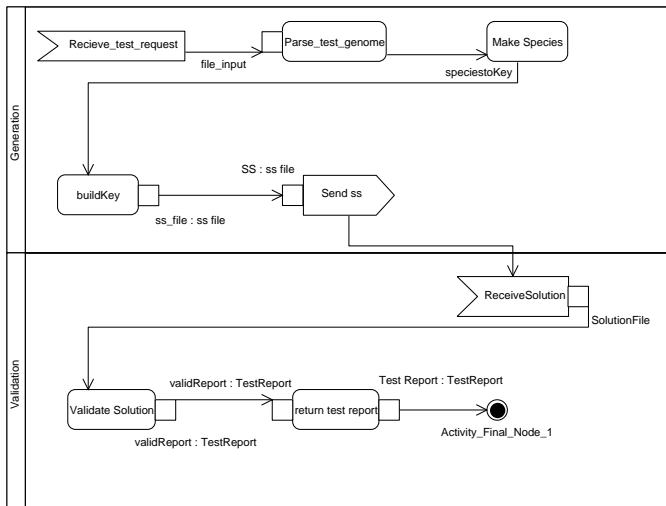


Created with Poseidon for UML Community Edition. Not for Commercial Use.

Components



Test Unit



Simulation

- Relevant Input Parameters
 - 1 Alignment overlap
 - 2 Number of Species
 - 3 Length of original supersequence
 - 4 Number of SNPS
 - 5 Read (fragment) length mean & standard deviation
- Parameters contained in XML file (with others)
- Software is in test stage
 - XML file is read by a generator
 - generator simulates the creation of shotgun reads
 - reads are fed to aligner, aligned reads to grapher
 - reports are produced

Problems and Promise at r.405

alignment component The aligner is known to be inaccurate.

Typically, it does not appear to have a significant or indeed any overlap with the true alignment.

However, the innaccuracy is not significant for overall results

graphing component The detection of SNP positions is excellent with the correctly aligned data and with the data from the aligner. The graphing component consistently finds the correct number of species.

validation and reporting components Reporting is primitive, but effective and expandable.

Evaluation

- qualified success
- much future development needed
- significant challenges remain in all components

References I



G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz, “Snps problems, complexity, and algorithms,” in *ESA '01: Proceedings of the 9th Annual European Symposium on Algorithms*, (London, UK), pp. 182–193, Springer-Verlag, 2001.

2



A. Blum, “New approximation algorithms for graph coloring,” *J. ACM*, vol. 41, no. 3, pp. 470–516, 1994.