Revised
January 30, 2004

Lucas B. Chaney
January 23, 2004
CPSC 401

IRC Proposal Draft (and Weekly Update 1)

0. Background

   Tremendous effort and interest has recently been applied to the process of genomics, the determination of the entire DNA sequence of an organism of interest to researchers. A popular and highly successful method which is currently employed is known as whole genome shotgun (WGS) sequencing. While tremendous computing power and programming effort has been applied to the problem of WGS genome assembly, researchers often find that while large "contigs" are not difficult to obtain, there exist too few tools for use in gap closure.

1. Project Overview

   I seek to offer a simple tool which is easy to use by those with limited computing skills that will suggest contig alignment and thus facilitate gap closure. The program will utilize: the contig files produced by available assembly programs, NCBI BLAST, and Primer3.

2. Program Details

   The program will process a contig file, searching the ends of the contigs with BLAST to suggest local alignments of the ends of the contigs on related (sequenced) organisms, parse the BLAST output to remove "impossible" suggestions (those which are directionally incompatible, or suggest an overlap which an assembly program should have recognized), and, finally, utilize Primer3 to suggest primer pairs for PCR to determine if the suggested gap was accurate.

3. Necessary Steps
   a. Create a program which processes a file of contigs and prepares a file or table which consists of the sequences at the ends of those contigs
   b. Utilize BLAST to compare the file/table of ends to completed sequences
      i. Ideally this will be completed within the application I write. To this end, it will be necessary to learn the NCBI API (Advanced Programmer Interface)
   c. Parse the BLAST output for suggested alignments
      i. This will take into account the organism on which the match was found, the direction in which the match was oriented, the ends (left or right) of the contigs in the suggested match, and the gap size suggested by the match
      ii. Taking into account the number of organisms on which the match was found would eventually be useful, as the set of completed genomes grows
   d. Present the suggested alignments in an easily understandable visual style
   e. Utilize Primer3 to design the primer pairs for closing suggested gaps
      i. Some documentation has been found for the Primer3 program, suggesting this could be accomplished in the parent application if "Boulder" interface format is learned

4. Possible Extensions

   Due to some previous work in this area, it is possible that the above, "minimal", specifications could be exceeded. The initial priority will fall towards creating an application which is inclusive, and does not require the user to call BLAST or Primer3 individually. The next extension would come by incorporating a simple graphical user interface (GUI) which would encourage use of the program by biologists.

5. Challenges

   Foreseeable challenges primarily lie in interfacing with the NCBI API, successfully implementing Boulder I/O with Primer3, creating clear and attractive graphical output of results, and writing a simple but powerful (and preferably UNIX and Windows compatible) GUI.

6. Resources

   NCBI API (or C++ Toolkit) Information
   http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=toolkit.TOC&depth=2
   Primer3 v0.9 Readme (including information on Boulder I/O)
   http://www.broad.mit.edu/ftp/distribution/software/README.primer3_0_9_test
   GNN Assembling the genome
   http://www.genomenewsnetwork.org/articles/03_00/assemble_genome_3_24.shtml
   GNN Genome Sequencing
   http://www.genomenewsnetwork.org/whats_a_genome/Chp2_1.shtml
   Further Boulder-io information
   http://stein.cshl.org/software/boulder/
   http://www.broad.mit.edu/genome_software/other/boulder.html
   http://www.cfcl.com/cfcl/rdm/carny/1999.08.htmlc
   http://search.cpan.org/~lds/Boulder-1.30/Boulder.pod