Lucas B. Chaney
June 11, 2003
BBSI Research Project Proposal (Outline)

1. Familiarization and analysis (June 11 – 22)
   a. Locate currently available programs
      i) Many programs are available that perform a wide variety of useful operations for the completion of gaps in a genome sequence. Locating these programs will require searching the internet and literature for any publicly available programs.
      ii) Estimated time: 1 – 3 days
      iii) Expected product: web-page with easy access to all available programs
   b. Analyze currently available programs
      i) The features offered by a program and the strength of a program's operation of a specific function is not always readily apparent. By reading available documentation (published papers and online information) in addition to testing programs and attempting to find what algorithms are being utilized a clear picture of what can and cannot be done by current programs should emerge.
      ii) Estimated time: 3 – 10 days
      iii) Expected product: database or table allowing users to select and then access the best program for their current task. Ideally, easily updatable as new programs become available
   c. Prepare a flowchart
      i) Organizing the information gathered above into a simple flowchart would provide a useful and concise representation of the data gathered through the analysis above.
      ii) Estimated time: 1 – 2 days
      iii) Expected product: flowchart with clear representation of how to transform a shotgun-sequence into a most complete/most accurate set of contigs and what programs to then utilize for best gap-closing primer design (GCPD)

The difficulty and scale of steps 2 & 3(below) will depend on the quality of currently available programs, and the ease of interfacing scripts with the programs found and analyzed (above).

2. Assembly (June 23 – July 2)
   a. Processing (largely complete)
      i) Initial assembly with all available assembly programs to generate contigs. Will not take any time unless previously unknown/unused programs are found.
      ii) Estimated time: <1 day
      iii) Expected product: 3 or more different contig assemblies generated by different programs
   b. Discrepancy identification
      i) Finding differences in contig assemblies of different programs. Will be greatly aided by a script or program.
      ii) Estimated time: 2 – 4 days
      iii) Expected product: location and nature of discrepancies in outputs of all programs
   c. Analysis
      i) Through the analysis of the programs used in contig assembly, it will be possible to explain and understand discrepancies found above. Selection of the best contigs will occur.
      ii) Estimated time: 1 – 3 days

iii) Expected product: knowledge of which contigs should be included, why, and which programs generated the best contigs

d. Re-assembly
   i) Incorporating the knowledge from above into a complete file with the best possible set of contigs through either by-hand selection or a written script.
   ii) Estimated time: 3 – 5 days
   iii) Expected product: best, most complete set of contigs

3. GCPD [Gap-closing primer design] (July 2 – August 7)
   a. Program (or script) construction
      i) If any one of the programs currently out is found to be adequate, GCPD could be completed with it and this step would be reduced to one of familiarization. More likely, a number of programs will be found which will be somewhat useful, but inadequate in isolation. A script could be written that will: convert a single query into all needed formats; query all programs; rank, score, and sort results; and return the best results to the user. Finally, if a simple script is incapable of completing these tasks a more involved program could be written which would do so.
      ii) Estimated time:
         (1) Use of program: 1 – 2 days (familiarization)
         (2) Script: 5 – 10 days
         (3) Program: 20 – 40 days
      iii) Expected product: described above
   b. Program usage (likely concurrent with above)
      i) This step will require the use of the product selected/created above to design primers for gap closure
      ii) Estimated time: Unknown
      iii) Expected product: primers needed for gap closure
   c. Wet-lab closure[1] (concurrent with above)
      i) Preformed by lab workers
      ii) Estimated time: Unknown
      iii) Expected product: New sequence information
   d. Incorporation of new information (concurrent with above)
      i) Inclusion of new sequence information with existing knowledge requiring contig reanalysis and perhaps resulting in new primer designs. Along with steps b. and c. (above) repeated as needed.
      ii) Estimated time: Unknown
      iii) Expected product: COMPLETED GENOME!!!

4. Annotation (Academic year[2])
5. Mining and vaccinology (Peter's job / Next summer)

[1] Not my realm! (LBC)
[2] Or this summer, if a program for GCPD is found and the sequence is completed