# *Implementation of a Multifaceted Approach for Completion of a Microbial Genome*

*Lucas B. Chaney*

*Bioinformatics and Bioengineering Summer Institute, Virginia Commonwealth University, Richmond, Virginia 23220 USA*

Submitted June 14, 2004 to Jeff Elhai, Director BBSI

## INTRODUCTION

Recent technological innovations have made it possible for the powerful tool of genomics to reach many new research establishments. Institutions at the forefront of this expansion face a number of challenges as they gain the technical knowledge, procedural reliability, and "finishing instincts" of current field leaders. Until the personnel conducting the research have gained these key skills, through hard won experience, their progress is not likely to proceed at the impressive pace of such paragon facilities as TIGR.

Continued effort is required in the completion of the *Streptococcus sanguis* genome. Despite substantial[1] coverage by whole genome shotgun sequences, a large number of unsequenced regions located between contigs, gaps[2], remain. Thus far a number of methods, collectively belonging to a set of processes known as gap closure, have been attempted in an effort to sequence the portion of DNA that spans adjacent contigs. Each of these attempts have produced varying degrees of success. Clone pair information has been exploited to suggest around 40 gaps that were successfully closed (Xu *et al.*, unpublished data). An alignment based on protein data from *S. pneumoniae* R6 (Hoskins *et al.*, 2001) suggested 120 contig pairings, but only 3 gaps were closed using the PCR products generated (Xu *et al.*, unpublished data). Additionally, a relatively expensive method of multi-step PCR has been used in a genome walking strategy that has been 50-60% effective in extending outward from the known ends of contigs (Xu *et al.*, unpublished data).

## METHODS

I have developed a more comprehensive, BLASTN (Altschul *et al.*, 1990) comparison of the current set of contigs to the NCBI "nr" nucleotide database to suggest a further group of possible contig pairs. The program "BOX-C" (Chaney, manuscript) generated 26 suggested alignments, which was then compared to the previously generated list of possible contig pairs. All BOX-C suggested pairs proved to be independent of those previously suggested, and a sample of 5 have been selected for analysis with PCR. These PCR tests will be conducted using a mix of previous and redesigned primers. Previous primers are being used to minimize delay and reduce financial cost, while new primers have been designed and will be synthesized in an effort to produce a higher yield than the previous trials for which some suspicions—of primer specificity, labeling, and purity—are present. Due to the involvement of at least one previously designed primer in 4 of the 5 reactions being undertaken, it may be necessary to redesign and freshly synthesize the reused primers if the tests with mixed primers are unsuccessful.

## POSSIBLE RESULTS AND THEIR IMPLICATIONS

Subsequent work will be dependent upon the yield of the BOX-C suggested alignments. If the PCR results confirm these as largely successful, more of the putative pairings will be tested and hopefully verified. Otherwise, a variety of genome walking approaches might be utilized. Two such approaches are currently at use in various branches of the laboratory. One utilizes short non-specific primers in conjunction with end specific primers (Kitten *et al.*, manuscript) while the other uses a "joiner" attached to a sequence near a contig's end and two long, highly specific primers in one stage of PCR followed by shorter primers in a second PCR stage (Xu *et al.*, unpublished data). Either method is expected to produce a product that, when sequenced, adds to the length of known sequence, extending a current contig. In the event the contig numbers decrease sufficiently[3], through BOX-C gap closure or genome walking PCR, it is likely multiplex PCR (Tettelin *et al.*, 1999) will then be attempted.

This combination of approaches is expected to be useful for closing many of the smaller gaps, but may not work as well for some of the larger[4] unsequenced regions. For these problem areas it may be necessary

[1] Approximately 13x
[2] Approximately 250
[3] Fewer than 100 (estimated)
[4] Suggested to be as large as approximately 22kb

to generate a mid-sized clone library in plasmids or cosmids, and/or utilize a BAC alignment.

## BUDGET

- PCR Primers ($20 per reaction) $600.00
- PCR Kits ($4.25 per reaction) $127.50
- Electrophoresis gels (for verification of PCR) $32.50
- Sequencing of PCR products ($8 per sequencing) $240.00

## REFERENCES


Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) A basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.

Hoskins,J., Alborn,W., Arnold,J., Blaszczak,L., Burgett,S., DeHoff,B., Estrem,S., Fritz,L., Fu,D., Fuller,W., Geringer,C., Gilmour,R., Glass,J., Khoja,H., Kraft,A., Lagace,R., LeBlanc,D., Lee,L., Lefkowitz,E., Lu,J., Matsushima,P., McAhren,S., McHenney,M., McLeaster,K., Mundy,C., Nicas,T., Norris,F., O'Gara,M., Peery,R., Robertson,G., Rockey,P., Sun,P., Winkler,M., Yang,Y., Young-Bellido,M., Zhao,G., Zook,C., Baltz,R., Jaskunas,S., Rosteck,P., Skatrud,P. and Glass,J. (2001) Genome of the bacterium Streptococcus pneumoniae strain R6. *J Bacteriol.*, **19**, 5709-5717.

Tettelin,H., Radune,D., Kasif,S., Khouri,H. and Salzberg,S. (1999) Optimized multiplex PCR: efficiently closing a whole-genome shotgun sequencing project. Genomics, 62, 500-507.