CALIFORNIA INSTITUTE OF TECHNOLOGY,
Biology Division

May 5, 1999

Re: DE-FG03 98 ER 62550

"Complete sequencing of the 2.3 Mbp genome of the hyperthermophilic archaeon *Pyrobaculum aerophilum*".

P.I.: Melvin I. Simon

FINAL

PROGRESS REPORT

1/01/98 - 12/31/98

Pyrobaculum aerophilum is a hyperthermophilic archaeon discovered from a boiling marine water hole at Maronti Beach, Italy that is capable of growth at 110°C. This microorganism can grow aerobically, unlike most of its thermophilic relatives. Due to the tolerance to oxygen, it is possible to grow this microbe in the presence of air, i.e. on plates. Therefore, it is a good candidate a model organism for studying archaeal biology and thermophilism. Sequencing the entire genome of this organism will provide a wealth of information on the evolutionary and phylogenetic relationship between archaea and other organisms as well as the biology of thermophilism. We have constructed a physical map that covers estimated 2.3 Megabase pair genome using a 10X fosmid library. The map currently consists of 96 overlapping fosmid clones. We have completed sequencing the entire genome using in random shotgun approach with the supplement of oligonucleotide primer directed sequencing. Total 16,098 random sequences corresponding to approximately 3.5X genomic coverage were obtained by sequencing from both ends with vector-specific primers the 2-3 kbp genomic DNA fragments cloned into pUC18/19 vector after shearing have been assembled into a number of contigs using Phrap program developed by Dr. Phil Green at University of Washington, Seattle. Gaps and regions of low quality base calls have been er a total of 2,300 directed sequencing and reassembly. Our current full length genomic sequence still suffers from low data quality: only approximately 99% of the nucleotide sequences are accurate. This is mainly due to the low redundancy (3.5 fold) in random sequencing. We plan to perform 2-3,000 more directed sequencing to polish the sequence to 99.99% accuracy. Final polishing of the sequence data and annotation is currently being performed by UCLA team and Caltech sequencing core facility.

Melvin I. Simon

Professor and Chairman, Division of Biology

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.