Final Technical Report: **UNIVERSITY OF UTAH**
"Automated Multiplex Sequencing of Microbial Genomes"
DE-FG03-94ER61950: 9/15/94 to 3/14/98

P.I.: Robert B. Weiss, Ph.D., Dept. of Human Genetics, University of Utah, Salt Lake City, UT.

The primary goal of this research was to test a unique DNA sequencing strategy for sequencing microbial genomes. During the testing of this strategy, this project generated the complete genome sequence of a hyperthermophilic archaeon: *Pyrococcus furiosus*.

Pyrococcus furiosus is the most well studied member of the unusual class of organisms known as the extreme hyperthermophiles. These organisms live at the extremes of temperature and pressure where life is still possible. *Pyrococcus furiosus* was isolated from geothermally heated marine sediment in the shallow waters off Vulcano Island, Italy. It grows optimally at 100°C, deriving its energy by a fermentative metabolism using protein, peptide and sugar mixtures. The organism is fast growing, capable of dividing every forty minutes. The genome of this organism is also of phylogenetic interest, since it is part of the Thermococcales branch within the Archaea.

Extreme hyperthermophiles play an important role in advancing the fundamental understanding of protein biochemistry, RNA and DNA metabolism, and protein interactions. How has the macromolecular machinery of the cell adapted to function at 100°C? Proteins from organisms living at moderate temperatures unfold, or denature, when heated. The proteins from hyperthermophiles maintain their three-dimensional shapes at these high temperatures. The genome sequence provides a resource for beginning to understand why these proteins can maintain their shapes and activities at extreme temperatures.

Extremely stable proteins have potential uses in biotechnology as rugged industrial catalysts. *Pyrococcus furiosus* possesses a diverse metabolism that provides a wide variety of potentially useful biocatalysts. These enzymes have potential uses as environmentally safe reagents for transforming biomass to derive energy and specialty chemicals, and in degrading organic compounds for environmental detoxification.

This project first generated a plasmid library of cloned *P. furiosus* genomic DNA. The genomic DNA was isolated from a fresh cell pellet grown in the laboratory of Dr. Frank Robb, at the Center of Marine Biotechnology, University of Maryland. Genomic DNA was isolated and cloned into an *E. coli* vector system designed to stably propagate medium insert (10 kb) DNA fragments. The library contained 2400 clones, with an average insert size of 5.2 kb. The average insert size was below the target of 10 kb average insert size, primarily due to the difficulty in isolating intact high molecular weight genomic DNA from *P. furiosus*. This library was supplemented with a cosmid library comprising 400 clones with an average insert size of 35 kb. A unique feature of the library was the distribution of the plasmid library into a series of related vectors carrying different sequence tags flanking the cloned inserts. These tags were used in the transposon mapping phase of the project.

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.

The second stage of the project generated the complete genomic sequence in three strategic phases. Initially, paired end reads were generated from all the clones. These sequence reads served as a sequence-based scaffold to guide the assembly of sub-clone library. At this phase, all plasmid inserts were sized on agarose gels; this provided length information to the paired-end read data set. A subset of 100 plasmid inserts containing an average insert size of 8.5 kb was chosen for an initial round of full-insert mapping and sequencing.

The mapping and sequencing phase involved the generation of a minimal set of transposon inserts evenly spaced across the insert. DNA sequencing using the transposon inserts as priming sites for dideoxy sequencing followed this process. Both the transposon mapping and sequencing phase used molecular procedures based on separation and transfer of DNA fragments onto microporous nylon membranes, and probing of these membranes with DNA probes. The DNA hybridization and detection was performed on automated instruments capable of imaging fluorescence from fluorogenic DNA probe/enzyme conjugate detection chemistry.

Typically, 288 independent transposon inserts were isolated in a fragment. Their locations were mapped via restriction digestion and multiplex probing. Subsets of clones were retrieved from frozen storage and processed for multiplex sequence. In the final analysis, 670 plasmid inserts were processed through transposon mapping. The majority of these also were processed through multiplex sequencing, yielding 85 % of the total sequence coverage of the genome. At this point, the sequence process was transitioned to fluorescent dye terminator chemistry using ABI 377 sequencing instruments. These dye terminator reads complemented the dye primer chemistry used in the multiplex reactions, leading to an increase in data quality.

Genome closure and completion was attained with multiple rounds of finishing procedures. The paired end plasmid scaffold allowed extension of the sequence islands into regions of sequence contiguity. The plasmid library topped out at 34 sequence islands covering 99% of the genome. The cosmid library spanned numerous gaps, reducing the sequence islands to 20. Isolating PCR fragments from genomic DNA spanned the final gaps. The sequence was completed with both sub-clone libraries and primer walking using the cosmids and PCR fragments as the physical substrate.

The project determined the complete sequence of this organism's genome. It is 1,908,253 base pairs in length, with a GC-content of 40.8%. The unannotated version can be retrieved via ftp from <http://www.genome.utah.edu>. Upon acceptance of the manuscript in preparation that describes the complete sequence, the annotations will be available in GenBank at the National Center for Biotechnology Information. Recently, the complete sequence of a distantly-related species, *Pyrococcus horikoshii* has also been determined. This species was isolated from a hydrothermal vent at a depth of 1395 meters in the Sea of Japan. Comparative analysis is revealing complex gene rearrangements and changes in gene content between these two *Pyrococcus* species.

The genome content and organization reveals many potential operons, one rRNA operon, 46 tRNAs, 22 insertion elements, 7 SR elements, and 14 inteins. 50 % of the ORFs are of unknown function, and of this class 19% are in common between the two *Pyrococcus* species. The 22 putative insertion elements in the genome of *P. furiosus* are not found in the *P. horikoshii* genome.

The genome of *P. furiosus* is 170 kb larger than *P. horikoshii*, with about 70% DNA identity conserved within open reading frames. Genome to genome dot plot alignment reveals the remnant of a conserved diagonal. The longest co-linear segment between the two genomes is 70 kb, and much of the remnant diagonal is interspersed with inversions, deletions and insertions. The list of major gene clusters present in *P. furiosus* but not in *P. horikoshii* include: maltose/trehalose transport, phosphate uptake system, major parts of the urea and TCA cycle, and amino acid metabolism of tryptophan, aromatics, arginine, and isoleucine/valine. The maltose/trehalose transport operon is within a 17 kb segment flanked by putative insertion elements. This segment is also found in *Thermococcus litoralis*, another isolate from a Mediterranean marine geothermal location. The high identity (>99%) between these two segments suggests a recent lateral transfer event between *T. litoralis* and *P. furiosus*.

In conclusion, this project generated the complete genome sequence of *P. furiosus* while testing several elements of a new DNA sequencing strategy. The most successful element was the ability to stably propagate medium inserts plasmid libraries of foreign DNA in standard *E. coli* strains. The combination of paired end sequence reads with these medium insert plasmids provides a new, tested method for whole genome shotgun strategies applicable to other genomes.

Publications from this project (as of 11/99):

- Riera, J., Robb, R.T., Weiss, R., Fontecave. (1997) Ribonucleotide reductase in the archaeon *Pyrococcus furiosus*: a critical enzyme in the evolution of DNA genomes? Proc. Natl. Acad. Sci. U.S.A. 94:475-478.
- Brown JR, Robb FT, Weiss R, Doolittle WF (1997). Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. J Mol Evol 1997 Jul;45(1):9-16.
- Ghosh M, Grunden AM, Dunn DM, Weiss R, Adams MW (1998) Characterization of native and recombinant forms of an unusual cobalt-dependent proline dipeptidase (prolidase) from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol 1998 Sep;180(18):4781-9.
- Roy R, Mukund S, Schut GJ, Dunn DM, Weiss R, Adams MW (1999) Purification and Molecular Characterization of the Tungsten-Containing Formaldehyde Ferredoxin Oxidoreductase from the Hyperthermophilic Archaeon *Pyrococcus furiosus*: the Third of a Putative Five-Member Tungstoenzyme Family. J Bacteriol 1999 Feb 15;181(4):1171-1180.
- Maeder DL, Weiss RB, Dunn DM, Cherry JL, Gonzalez JM, DiRuggiero J, Robb FT (1999) Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences. Genetics 1999 Aug;152(4):1299-305.